Pandemic Genomic Surveillance: SARS-CoV-2 Real-time Genomic Epidemiology to Identify and Track Variants in South Africa and Africa

Houriiyah Tegally
Doctor of Philosophy (Ph.D.)
(Medicine)

UNIVERSITY OF KWAZULU-NATAL
INYUVESI YAKWAZULU-NATALI

School of Laboratory Medicine and Medical Sciences, College of Health Sciences
University of KwaZulu-Natal
Durban, South Africa
2022

Supervisor:
Prof. Tulio de Oliveira

Table of Contents

Declaration 1: Authorship

The research described in this thesis was carried out in the KwaZulu-Natal Research Innovation and Sequencing Platform (KRISP) and in the Department of Virology, School of Laboratory Medicine and Medical Sciences, College of Health Sciences, University of KwaZulu-Natal, in Durban, South Africa. It was conducted between October 2019 and October 2022 under the supervision of Prof Tulio de Oliveira. This research is covered by ethics certification by the UKZN Research Office (BREC/00003933/2022).

The work contained in this thesis has not been previously submitted for a degree or diploma in any other higher education institution. This thesis contains no material previously published or submitted for publication by another person except where due reference has been made.

Houriiyah Tegally
Date: 28th October 2022

Prof. Tulio de Oliveira
Date: 28th October 2022

Declaration 2: Plagiarism

I **Houriiyah Tegally** declare that:

(i) The research reported in this thesis is my original work, except where otherwise indicated.
(ii) This thesis has not been submitted for any degree or examination at any other university.
(iii) This thesis does not contain other persons' data, pictures, graphs, or other information unless specifically acknowledged as being sourced from other persons.
(iv) This thesis does not contain other persons' writing unless specifically acknowledged as being sourced from other researchers. Where other written sources have been quoted, then:

a) Their words have been re-written, but the general information attributed to them has been referenced.

b) Where their exact words have been used, their writing has been placed inside quotation marks and referenced.

(v) Where I have reproduced a publication of which I am an author, co-author, or editor, I have indicated in detail which part of the publication was written by myself alone and have fully referenced such publications.
(vi) This thesis does not contain texts, graphics or tables copied and pasted from the internet unless specifically acknowledged, and the source being detailed in the thesis and the references.

Date: 28th October 2022

Abstract

The SARS-CoV-2 pandemic has both been one of the largest public health emergencies of modern times and an unprecedented opportunity to track the epidemic progression of an evolving virus. Globally, more than 13 million genomic sequences have been generated, over 140,000 of which was from surveillance in African countries. The work presented in this thesis employs methods of real-time genomic surveillance and epidemiology, genome assembly, phylogenetic analysis and phylodynamic modelling to characterise the evolution of SARS-CoV-2 in South Africa and Africa with immediate public health impact globally. Chapter 2 describes the setting up and results of genomic surveillance in all provinces during the first wave of infections in South Africa. Insights included the description of three local lineages that caused over half of infections in the first wave and the establishment of surveillance baselines that enabled rapid characterization of variants of concern in upcoming waves. Chapter 3 provides a description of, for the first time in the world, the emergence of a SARS-CoV-2 variant of concern. The study gives an overview of the detection of the Beta variant, its association with an accelerating epidemic in the Eastern Cape province and the inferred phylogeography of how the variant spread to coastal provinces during summer holidays in South Africa. Chapter 4 describes the characterisation and phylodynamics of the Omicron variant of concern in record time at the start of the 4th wave of infections in southern Africa. Chapter 5 provides insights into the continued evolution of Omicron into sublineages BA.4 and BA.5, which went on to dominate the epidemic in other parts of the world in mid-2022. Finally, Chapter 6 and 7 are comprehensive studies of continental genomic surveillance in Africa, giving insights into establishment of epidemics from introductions from external sources, cross-border viral movements and the expansion of genomic surveillance on the continent to cover blindspots. This thesis also contributed to a number of other studies where genomic sequencing of SARS-CoV-2 helped to answer critical questions during pandemic response, which is described in the last chapter. In conclusion, this thesis exemplifies how genomic epidemiology can be utilised in real-time to track the evolution of a pandemic pathogen as well as rapidly raise alarms of detected global health threats.

Publications

Publications from this Ph.D. Thesis:

1. **The evolving SARS-CoV-2 epidemic in Africa: Insights from rapidly expanding genomic surveillance. Tegally H,** San JE, Cotten M, …, **de Oliveira T***, Happi C, Lessells R, Nkengasong J, Wilkinson E*, ***Science*** (2022), eabq5358. doi: 10.1126/science.abq5358:.

2. **Emergence of SARS-CoV-2 Omicron lineages BA.4 and BA.5 in South Africa.** **Tegally H**, Moir M, Everatt J, Giovanetti M, Scheepers C, Wilkinson E, Subramoney K, Moyo S, Amoako D, Althaus C, Anyaneji U, Kekana D, Viana R, Giandhari J, Maponga T, Maruapula D, Choga W, Mayaphi S, Mbhele N, Gaseitsiwe S, Msomi N, Naidoo Y, Pillay S, Sanko T, San J, Scott L, Singh L, Magini N, Smith-Lawrence P, Stevens W, Dor G, Tshiabuila D, Wolter N, Preiser W, Treurnicht F, Venter M, Davids M, Chiloane G, Mendes A, McIntyre C, O'Toole A, Ruis C, Peacock T, Roemer C, Williamson C, Pybus O, Bhiman J, Glass A, Martin D, Rambaut A, Gaseitsiwe S, von Gottberg A, Baxter C, Lessells R, **de Oliveira T**, ***Nature Medicine*** (2022), https://doi.org/10.1038/s41591-022-01911-2

3. **A year of genomic surveillance reveals how the SARS-CoV-2 pandemic unfolded in Africa .** Wilkinson E*, Giovanetti M*, **Tegally H*,** San JE, … , Nkengasong J, **de Oliveira T**, ***Science*** (2021), DOI: 10.1126/science.abj4336:.

4. **Rapid epidemic expansion of the SARS-CoV-2 Omicron variant in southern Africa.** Viana R*, Moyo S*, Amoako DG*, **Tegally H*,** Scheepers C, Althaus CL, Anyaneji UJ, Bester PA, Boni MF, Chand M, Choga WT, Colquhoun R, Davids M, Deforche K, Doolabh D, du Plessis L, Engelbrecht S, Everatt J, Giandhari J, Giovanetti M, Hardie D, Hill V, Hsiao NY, Iranzadeh A, Ismail A, Joseph C, Joseph R, Koopile L, Kosakovsky Pond SL, Kraemer MUG, Kuate-Lere L, Laguda-Akingba O, Lesetedi-Mafoko O, Lessells RJ, Lockman S, Lucaci AG, Maharaj A, Mahlangu B, Maponga T, Mahlakwane K, Makatini Z, Marais G, Maruapula D, Masupu K, Matshaba M, Mayaphi S, Mbhele N, Mbulawa MB, Mendes A, Mlisana K, Mnguni A, Mohale T, Moir M, Moruisi K, Mosepele M, Motsatsi G, Motswaledi MS, Mphoyakgosi T, Msomi N, Mwangi PN, Naidoo Y, Ntuli N, Nyaga M, Olubayo L, Pillay S, Radibe B, Ramphal Y, Ramphal U, San JE, Scott L, Shapiro R, Singh L, Smith-Lawrence P, Stevens W, Strydom A, Subramoney K, Tebeila N, Tshiabuila D, Tsui J, van Wyk S, Weaver S, Wibmer CK, Wilkinson E, Wolter N, Zarebski AE, Zuze B, Goedhals D, Preiser W, Treurnicht F, Venter M, Williamson C, Pybus OG,

Bhiman J, Glass A, Martin DP, Rambaut A, Gaseitsiwe S, von Gottberg A, **de Oliveira T**, ***Nature*** (2022), doi: 10.1038/s41586-022-04411-y:.

5. **Detection of a SARS-CoV-2 variant of concern in South Africa. Tegally H**, Wilkinson E, Giovanetti M, Iranzadeh A, Fonseca V, Giandhari J, Doolabh D, Pillay S, San E, Msomi N, Mlisana K, Gottberg A, Walaza S, Allam M, Ismail A, Mohale T, Glass A, Engelbrecht S, Zyl G, Preiser W, Petruccione F, Sigal A, Hardie D, Marais G, Hsiao M, Korsman S, Davies M, Tyers L, Mudau I, York D, Maslo C, Goedhals D, Abrahams S, Laguda-Akingba O, Alisoltani-Dehkordi A, Godzik A, Wibmer Cos, Sewell B, Lourenco J, Alcantara Ls, Kosakovsky Pond S, Weaver S, Martin D, Lessells R, Bhiman J, Williamson C, **de Oliveira T**, ***Nature*** (2021), https://doi.org/10.1038/s41586-021-03402-9:.

6. **Sixteen novel lineages of SARS-CoV-2 in South Africa. Tegally H**, Wilkinson E, Lessells R, Giandhari J, Pillay S, Msomi N, Mlisana K, Bhiman J, Gottberg A, Walaza S, Fonseca V, Allam M, Ismail A, Engelbrecht S, Van Zyl G, Preiser W, Williamson C, Pettruccione F, Sigal A, Gazy I, Hardie D, Hsiao M, Martin D, York D, Goedhals D, San EJ, Giovanetti M, Lourenco J, Alcantara LCJ, **de Oliveira T**, ***Nature Medicine*** (2021), https://doi.org/10.1038/s41591-021-01255-3:.

7. **Unlocking the efficiency of genomics laboratories with robotic liquid-handling. Tegally H**, San JE, Giandhari J, **de Oliveira T**, ***BMC Genomics*** (2020), 729:https://doi.org/10.1186/s12864-020-07137-1.

Collaborative publications during the Ph.D.

1. **Urgent need for a non-discriminatory and non-stigmatizing nomenclature for monkeypox virus.** Happi C, Adetifa I, Mbala P, Njouom R, Nakoune E, Happi A, Ndodo N, Ayansola O, Mboowa G, Bedford T, Neher RA, Roemer C, Hodcroft E, **Tegally H**, O'Toole Á, Rambaut A, Pybus O, Kraemer MUG, Wilkinson E, Isidro J, Borges V, Pinto M, Gomes JP, Freitas L, Resende PC, Lee RTC, Maurer-Stroh S, Baxter C, Lessells R, Ogwell AE, Kebede Y, Tessema SK, **de Oliveira T**, ***PLoS Biology*** (2022), doi: 10.1371/journal.pbio.3001769.:.

2. **Genomic epidemiology of the SARS-CoV-2 epidemic in Brazil.** Giovanetti M, Slavov SN, Fonseca V, Wilkinson E, **Tegally H**, … **de Oliveira T**, Holmes EC, Haddad R, Sampaio SC, Elias MC, Kashima S, Junior de Alcantara LC, Covas DT, ***Nature Microbiology*** (2022), doi: 10.1038/s41564-022-01191-z:.

3. **SARS-CoV-2 Genetic Diversity and Lineage Dynamics in Egypt during the First 18 Months of the Pandemic.** Roshdy WH, Khalifa MK, San JE, **Tegally H**, Wilkinson E, Showky SM, Darren P, Moir M, Naguib AEN, Gomaa MR, Fahim MAEH, Mohsen AGR, Hassany M, Lessells R, Al-Karmalawy AA, EL-Shesheny R, Kandeil AM, Ali MA, **de Oliveira T**, *Viruses* (2022), https://www.mdpi.com/1999-4915/14/9/1878:.

4. **Building genomic sequencing capacity in Africa to respond to the SARS-CoV-2 pandemic.** **de Oliveira T,** Wilkinson E, Baxter C, **Tegally H**, Giandhari J, Naidoo Y, Pillay S, ***Science*** (2022), https://www.science.org/content/resource/pandemic-preparedness-changing-world-fostering-global-collaboration, 2022:.

5. **Omicron BA.4/BA.5 escape neutralizing immunity elicited by BA.1 infection..** Khan K, Karim F, Ganga Y, Bernstein M, Jule Z, Reedoy K, Cele S, Lustig G, Amoako D, Wolter N, Samsunder N, Sivro A, San JE, Giandhari J, **Tegally H**, Pillay S, Naidoo Y, Mazibuko M, Miya Y, Ngcobo N, Manickchund N, Magula N, Karim QA, von Gottberg A, Abdool Karim SS, Hanekom W, Gosnell BI; COMMIT-KZN Team, Lessells RJ, **de Oliveira T,** Moosa MS, Sigal A, ***Nature*** Communications (2022), doi: 10.1038/s41467-022-32396-9:.

6. **Identification of SARS-CoV-2 Omicron variant using spike gene target failure and genotyping assays, Gauteng, South Africa, 2021.** Subramoney K, Mtileni N, Bharuthram A, Davis A, Kalenga B, Rikhotso M, Maphahlele M, Giandhari J, Naidoo Y, Pillay S, Ramphal U, Ramphal Y, **Tegally H**, Wilkinson E, Mohale T, Ismail A, Mashishi B, Mbenenge N, **de Oliveira T**, Makatini Z, Fielding BC, Treurnicht FK, ***J Med Virol.*** (2022), 4(8):3676-3684. doi: 10.1002/jmv.27797:.

7. **Tracking the 2022 monkeypox outbreak with epidemiological data in real-time.** Kraemer MUG, **Tegally H,** Pigott DM, Dasgupta A, Sheldon J, Wilkinson E, Schultheiss M, Han A, Oglia M, Marks S, Kanner J, OBrien K, Dandamudi S, Rader B, Sewalk K, Bento AI, Scarpino SV, **de Oliveira T**, Bogoch II, Katz R, Brownstein JS, ***The Lancet Infectious Diseases*** (2022), DOI:https://doi.org/10.1016/S1473-3099(22)00359-0:.

8. **Omicron infection enhances Delta antibody immunity in vaccinated persons.** Khan K, Karim F, Cele S, Reedoy K, San EJ, Lustig G, **Tegally H**, Rosenberg Y, Bernstein M, Jule Z, Ganga Y, Ngcobo N, Mazibuko M, Mthabela N, Mhlane Z, Mbatha N, Miya Y, Giandhari J, Ramphal Y, Naidoo T, Sivro A, Samsunder N, Kharsany A, Amoako D, Bhiman J, Manickchund N, Karim Q, Magula N, Abdool

Karim SS, Gray G, Hanekom W, von Gottberg A, Milo R, Gosnell B, Lessells R, Moore P, **de Olveira T**, Moosa M-Y S, Sigal A, ***Nature*** (2022), https://doi.org/10.1038/s41586-022-04830-x:.

9. **Emergence and phenotypic characterization of the global SARS-CoV-2 C.1.2 lineage.** Scheepers C, Everatt J, Amoako DG, **Tegally H**, Wibmer CK, Mnguni A, Ismail A, Mahlangu B, Lambson BE, Martin DP, Wilkinson E, San JE, Giandhari J, Manamela N, Ntuli N, Kgagudi P, Cele S, Richardson SI, Pillay S, Mohale T, Ramphal U, Naidoo Y, Khumalo ZT, Kwatra G, Gray G, Bekker LG, Madhi SA, Baillie V, Van Voorhis WC, Treurnicht FK, Venter M, Mlisana K, Wolter N, Sigal A, Williamson C, Hsiao NY, Msomi N, Maponga T, Preiser W, Makatini Z, Lessells R, Moore PL, **de Oliveira T**, von Gottberg A, Bhiman JN, ***Nature Commununications*** (2022), 13(1):1976. doi: 10.1038/s41467-022-29579-9:.

10. **Replacement of the Gamma by the Delta variant in Brazil: Impact of lineage displacement on the ongoing pandemic.** Giovanetti M, Fonseca V, Wilkinson E, **Tegally H**, San EJ, Althaus CL, Xavier J, Nanev Slavov S, Viala VL, Ranieri Jerônimo Lima A, Ribeiro G, Souza-Neto JA, Fukumasu H, Lehmann Coutinho L, Venancio da Cunha R, Freitas C, Campelo de A E Melo CF, Navegantes de Araújo W, Do Carmo Said RF, Almiron M, **de Oliveira T**, Coccuzzo Sampaio S, Elias MC, Covas DT, Holmes EC, Lourenço J, Kashima S, de Alcantara LCJ, ***Virus Evolution*** (2022), doi: 10.1093/ve/veac024:.

11. **Selection analysis identifies clusters of unusual mutational changes in Omicron lineage BA.1 that likely impact Spike function.** Martin DP, Lytras S, Lucaci AG, Maier W, Grüning B, Shank SD, Weaver S, MacLean OA, Orton RJ, Lemey P, Boni MF, **Tegally H**, Harkins GW, Scheepers C, Bhiman JN, Everatt J, Amoako DG, San JE, Giandhari J, Sigal A; NGS-SA, Williamson C, Hsiao NY, von Gottberg A, De Klerk A, Shafer RW, Robertson DL, Wilkinson RJ, Sewell BT, Lessells R, Nekrutenko A, Greaney AJ, Starr TN, Bloom JD, Murrell B, Wilkinson E, Gupta RK, **de Oliveira T**, Kosakovsky Pond SL, ***Mol Biol Evol.*** (2022), doi: 10.1093/molbev/msac061:.

12. **Rapid Replacement of SARS-CoV-2 Variants by Delta and Subsequent Arrival of Omicron, Uganda, 2021.** Bbosa N, Ssemwanga D, Namagembe H, Kiiza R, Kiconco J, Kayiwa J, Lutalo T, Lutwama J, Ssekagiri A, Ssewanyana I, Nabadda S, Kyobe-Bbosa H, Giandhari J, Pillay S, Ramphal U, Ramphal Y, Naidoo Y, Tshiabuila D, **Tegally H**, San EJ, Wilkinson E, **de Oliveira T**, Kaleebu P, ***Emerg Infect Dis*** (2022), doi: 10.3201/eid2805.220121:.

13. **Targeted Sanger sequencing to recover key mutations in SARS-CoV-2 variant genome assemblies produced by next-generation sequencing.** Singh L, San JE, **Tegally H**, Brzoska PM, Anyaneji UJ, Wilkinson E, Clark L, Giandhari J, Pillay S, Lessells RJ, Martin DP, Furtado M, Kiran AM, **de Oliveira T**, ***Microb Genom.*** (2022), doi: 10.1099/mgen.0.000774:.

14. **Comparison of SARS-CoV-2 sequencing using the ONT GridION and the Illumina MiSeq.** Tshiabuila D, Giandhari J, Pillay S, Ramphal U, Ramphal Y, Maharaj A, Anyaneji UJ, Naidoo Y, **Tegally H**, San EJ, Wilkinson E, Lessells R, **de Oliveira T**, ***Research Square*** (2022), https://doi.org/10.21203/rs.3.rs-1249711/v1:.

15. **Persistent SARS-CoV-2 infection with accumulation of mutations in a patient with poorly controlled HIV infection.** Maponga TG, Jeffries M, **Tegally H**, Sutherland A, Wilkinson E, Lessells R, Msomi N, van Zyl G, **de Oliveira T**, Preiser W, ***Clin Infect Dis.*** (2022), ciac548. doi: 10.1093/cid/ciac548:.

16. **T cell responses to SARS-CoV-2 spike cross-recognize Omicron.** Keeton R, Tincho MB, Ngomti A, Baguma R, Benede N, Suzuki A, Khan K, Cele S, Bernstein M, Karim F, Madzorera SV, Moyo-Gwete T, Mennen M, Skelem S, Adriaanse M, Mutithu D, Aremu O, Stek C, du Bruyn E, Van Der Mescht MA, de Beer Z, de Villiers TR, Bodenstein A, van den Berg G, Mendes A, Strydom A, Venter M, Giandhari J, Naidoo Y, Pillay S, **Tegally H**, Grifoni A, Weiskopf D, Sette A, Wilkinson RJ, **de Oliveira T**, Bekker LG, Gray G, Ueckermann V, Rossouw T, Boswell MT, Bihman J, Moore PL, Sigal A, Ntusi NAB, Burgers WA, Riou C, ***Nature*** (2022), doi: 10.1038/s41586-022-04460-3:.

17. **Escape from recognition of SARS-CoV-2 Beta variant spike epitopes but overall preservation of T cell immunity.** Riou C, Keeton R, Moyo-Gwete T, Hermanus T, Kgagudi P, Baguma R, Valley-Omar Z, Smith M, **Tegally H**, Doolabh D, Iranzadeh A, Tyers L, Mutavhatsindi H, Tincho MB, Benede N, Marais G, Chinhoyi LR, Mennen M, Skelem S, du Bruyn E, Stek C; SA-CIN, **de Oliveira T**, Williamson C, Moore PL, Wilkinson RJ, Ntusi NAB, Burgers WA, ***Science Translational Medicine*** (2021), DOI: 10.1126/scitranslmed.abj6824:.

18. **Reduced amplification efficiency of the RNA-dependent-RNA-polymerase target enables tracking of the Delta SARS-CoV-2 variant using routine diagnostic tests.** Valley-Omar Z, Marais G, Iranzadeh A, Naidoo M, Korsman S, Maponga T, Hussey H, Davies MA, Boulle A, Doolabh D, Laubscher M, Wojno J, Deetlefs JD, Maritz J, Scott L, Msomi N, Naicker C, **Tegally H, de Oliveira T**, Bhiman J, Williamson C,

Preiser W, Hardie D, Hsiao NY, ***J Virol Methods*** (2022), 302:114471. doi: 10.1016/j.jviromet.2022.114471:.

19. **Track Omicrons spread with molecular data.** Scott L, Hsiao NY, Moyo S, Singh L, **Tegally H**, Dor G, Maes P, Pybus OG, Kraemer MUG, Semenova E, Bhatt S, Flaxman S, Faria NR, **de Oliveira T**, ***Science*** (2021), DOI: 10.1126/science.abn4543 :.

20. **SARS-CoV-2 Omicron has extensive but incomplete escape of Pfizer BNT162b2 elicited neutralization and requires ACE2 for infection.** Cele S, Jackson L, Khoury DS, Khan K, Moyo-Gwete T, **Tegally H**, San JE, Cromer D, Scheepers C, Amoako DG, Karim F, Bernstein M, Lustig G, Archary D, Smith M, Ganga Y, Jule Z, Reedoy K, Hwa SH, Giandhari J, Blackburn JM, Gosnell BI, Abdool Karim SS, Hanekom W; NGS-SA; COMMIT-KZN Team, von Gottberg A, Bhiman JN, Lessells RJ, Moosa MS, Davenport MP, **de Oliveira T**, Moore PL, Sigal A, ***Nature*** (2021), doi: 10.1038/s41586-021-04387-1.:.

21. **Rapid replacement of the Beta variant by the Delta variant in South Africa.** **Tegally H**, Wilkinson E, Althaus C, Giovanetti M, San J, Giandhari J, Pillay S, Naidoo Y, Ramphal U, Msomi N, Mlisana K, Amoako D, Everatt J, Mohale T, Nguni A, Mahlangu B, Ntuli N, Khumalo Z, Makatini Z, Wolter N, Scheepers C, Ismail A, Doolabh D, Joseph R, Strydom A, Mendes A, Davids M, Mayaphi S, Ramphal Y, Maharaj A, Karim W, Tshiabuila D, Anyaneji U, Singh L, Engelbrecht S, Fonseca V, Marais K, Korsman S, Hardie D, Hsiao N, Maponga T, van Zyl G, Marais G, Iranzadeh A, Martin D, Alcantara L, Bester P, Nyaga M, Subramoney K, Treurnicht F, Venter M, Goedhals D, Preiser W, Bhiman J, vonGottberg A, Williamson C, Lessells R, ***de Oliveira T***, ***medRxiv*** (2021), MEDRXIV-2021-264018v1:.

22. **The emergence and ongoing convergent evolution of the SARS-CoV-2 N501Y lineages .** Martin DP, Weaver S, **Tegally H**, San EJ, Shank SD, Wilkinson E, Lucaci AG, Giandhari J, Naidoo S, Pillay Y, Singh L, Lessells RJ; NGS-SA; COVID-19 Genomics UK (COG-UK), Gupta RK, Wertheim JO, Nekturenko A, Murrell B, Harkins GW, Lemey P, MacLean OA, Robertson DL, **de Oliveira T**, Kosakovsky Pond SL, ***Cell*** (2021), https://doi.org/10.1016/j.cell.2021.09.003:.

23. **Implementation of an efficient SARS-CoV-2 specimen pooling strategy for high throughput diagnostic testing.** Singh L, Anyaneji UJ, Ndifon W, Turok N, Mattison SA, Lessells R, Sinayskiy I, San EJ, **Tegally H**, Barnett S, Lorimer T, Petruccione F, **de Oliveira T.** ***Scientific Reports*** (2021), 11(1):17793. doi: 10.1038/s41598-021-96934-z:.

24. **HIV-1 and SARS-CoV-2: Patterns in the evolution of two pandemic pathogens.** Fischer W, Giorgi EE, Chakraborty S, Nguyen K, Bhattacharya T, Theiler J, Goloboff PA, Yoon H, Abfalterer W, Foley BT, **Tegally H**, San EJ, **de Oliveira T**, Network for Genomic Surveillance in South Africa (NGS-SA), Gnanakaran S, Korber B, ***Cell Host Microbe*** (2022), 29(7):1093-1110. doi: 10.1016/j.chom.2021.05.012:.

25. **Multiplex qPCR discriminates variants of concern to enhance global surveillance of SARS-CoV-2.** Vogels CBF, Breban MI, Ott IM, Alpert T, Petrone ME, Watkins AE, Kalinich CC, Earnest R, Rothman JE, Goes de Jesus J, Morales Claro I, Magalhaes Ferreira G, Crispim MAE, Brazil-UK CADDE Genomic Network, Singh L, **Tegally H**, Anyaneji UJ; Network for Genomic Surveillance in South Africa, Hodcroft EB, Mason CE, Khullar G, Metti J, Dudley JT, MacKay MJ, Nash M, Wang J, Liu C, Hui P, Murphy S, Neal C, Laszlo E, Landry ML, Muyombwe A, Downing R, Razeq J, **de Oliveira T**, Faria NR, Sabino EC, Neher RA, Fauver JR, Grubaugh ND, ***PLoS Biology*** (2021), DOI: 10.1371/journal.pbio.3001236:.

26. **Transmission dynamics of SARS-CoV-2 within-host diversity in two major hospital outbreaks in South Africa.** San JE, Ngcapu S, Kanzi AM, **Tegally H**, Fonseca V, Giandhari J, Wilkinson E, Nelson CW, Smidt W, Kiran AM, Chimukangara B, Pillay S, Singh L, Fish M, Gazy I, Martin DP, Khanyile K, Lessells R, **de Oliveira T.** ***Virus Evolution*** (2021), 7(1):veab041. doi: 10.1093/ve/veab041. eCollection 2021 Jan:.

27. **Tracking the international spread of SARS-CoV-2 lineages B.1.1.7 and B.1.351/501Y-V2.** O'Toole A, Hill V, Pybus OG, Watts A, Bogoch II, Khan K, Messina JP, COVID-19 Genomics UK (COG-UK) consortium, Network for Genomic Surveillance in South Africa (NGS-SA), Brazil-UK CADDE Genomic Network, **Tegally H**, Lessells RR, Giandhari J, Pillay S, Tumedi KA, Nyepetsi G, Kebabonye M, Matsheka M, Mine M, Tokajian S, Hassan H, Salloum T, Merhi G, Koweyes J, Geoghegan JL, de Ligt J, Ren X, Storey M, Freed NE, Pattabiraman C, Prasad P, Desai AS, Vasanthapuram R, Schulz TF, Steinbruck L, Stadler T, Swiss Viollier Sequencing Consortium, Parisi A, Bianco A, Garcia de Viedma D, Buenestado-Serrano S, Borges V, Isidro J, Duarte S, Gomes JP, Zuckerman NS, Mandelboim M, Mor O, Seemann T, Arnott A, Draper J, Gall M, Rawlinson W, Deveson I, Schlebusch S, McMahon J, Leong L, Lim CK, Chironna M, Loconsole D, Bal A, Josset L, Holmes E, St George K, Lasek-Nesselquist E, Sikkema RS, Oude Munnink B, Koopmans M, Brytting M, Sudha Rani V, Pavani S, Smura T, Heim A, Kurkela S, Umair M, Salman M, Bartolini B, Rueca M, Drosten C, Wolff T, Silander O, Eggink

D, Reusken C, Vennema H, Park A, Carrington C, Sahadeo N, Carr M, Gonzalez G, SEARCH Alliance San Diego, National Virus Reference Laboratory, SeqCOVID-Spain, Danish Covid-19 Genome Consortium (DCGC), Communicable Diseases Genomic Network (CDGN), Dutch National SARS-CoV-2 surveillance program, Division of Emerging Infectious Diseases (KDCA), **de Oliveira T**, Faria N, Rambaut A, Kraemer MUG., ***Wellcome Open Res*** (2021), 6:121. doi: 10.12688/wellcomeopenres.16661.1. eCollection 2021.:.

28. **Cross-Reactive Neutralizing Antibody Responses Elicited by SARS-CoV-2 501Y.V2 (B.1.351).** Moyo-Gwete T, Madzivhandila M, Makhado Z, Ayres F, Mhlanga D, Oosthuysen B, Lambson EB, Kgagudi P, **Tegally H**, Iranzadeh A, Doolabh D, Tyers L, Chinhoyi RL, Mennen M, Skelm S, Wibmer K C, Bhiman N J, Ueckermann V, Rossouw T, Boswell M, **de Oliveira T**, Williamson T, Burgers W, Ntusi N, Morris L, Moore P, ***NEJM*** (2021), DOI: 10.1056/NEJMc2104192:.

29. **Multiple Early Introductions of SARS-CoV-2 to Cape Town, South Africa.** Engelbrecht S, Delaney K, Kleinhans B, Wilkinson E, **Tegally H**, Stander T, van Zyl G, Preiser W, **de Oliveira T**, ***Viruses*** (2021), 22;13(3):526. doi: 10.3390/v13030526:.

30. **Escape of SARS-CoV-2 501Y.V2 from neutralization by convalescent plasma..** Cele S, Gazy I, Jackson L, Hwa SH, **Tegally H**, Lustig G, Giandhari J, Pillay S, Wilkinson E, Naidoo Y, Karim F, Ganga Y, Khan K, Bernstein M, Balazs AB, Gosnell BI, Hanekom W, Moosa MS; NGS-SA; COMMIT-KZN Team, Lessells R, **de Oliveira T**, Sigal A., ***Nature*** (2021), DOI: 10.1038/s41586-021-03471-w:.

31. **A novel variant of interest of SARS-CoV-2 with multiple spike mutations detected through travel surveillance in Africa. de Oliveira T**, Lutucuta S, Nkengasong J, Morais J, Paixao JP, Neto Z, Afonso P, Miranda J, David K, Ingles L, Amilton P A P R R C, Freitas H R, Mufinda F, Tessema K S , **Tegally H**, San E J, Wilkinson E, Giandhari J, Pillay S, Giovanetti M, Naidoo Y, Katzourakis A, Ghafari M, Singh L, Tshiabuila D, Martin D, Lessells R, ***medRxiv*** (2021).

32. **Safety and efficacy of the ChAdOx1 nCoV-19 (AZD1222) Covid-19 vaccine against the B.1.351 variant in South Africa.** Madhi SA, Baillie VL, Cutland CL, Voysey M, KoenAL, Fairlie L, Padayachee SD, Dheda K, Barnabas SL, Bhorat QE, Briner C, Kwatra G, Ahmed K, Aley P, Bhikha S, Bhiman JN, Bhorat AE, du plessis J, Esmail A, Groenewald M, Horne E, Hwa S-H, Jose A, Lambe T, Laubscher M, Malahleha M, Masenya M, Masilela M, McKenzie S, Molapo K, Moultrie A, Oelofse S, Pate Fl, Pillay S, Rhead S, Rodel H, Rossouw L, Taoushanis C, **Tegally H**, Thombrayil A, van Eck S, Wibmer C, Durham NM, Kelly EJ, Villafana T, Gilbert S,

Pollard AJ, **de Oliveira T**, Moore PL, Sigal A, Izu A, NGS-SA, Wits VIDA COVID vaccine trial group, ***NEJM*** (2021), https://doi.org/10.1056/NEJMoa2102214:.

33. **A genomics network established to respond rapidly to public health threats in South Africa.** Msomi N, Mlisana K, Willianson C, Bhiman JN, Goedhals D, Engelbrecht S, Van Zyl G, Preiser W, Hardie D, Hsiao M, Mulder N, Martin D, Christoffels A, York D, Giandhari J, Wilkinson E, Pillay S, **Tegally H**, James SE, Kanzi A, Lessells RJ, **de Oliveira T**, ***Lancet Microbe*** (2020), https://doi.org/10.1016/S2666-5247(20)30116-6:.

34. **Whole Genome Sequencing of SARS-CoV-2: Adapting Illumina Protocols for Quick and Accurate Outbreak Investigation During a Pandemic.** Pillay S, Giandhari J, **Tegally H**, Wilkinson E, Chimukangara B, Lessells R, Mattison S, Moosa Y, Gazy I, Fish M, Singh L, Khanyile KS, Fonseca V, Giovanetti M, Alcantara LCJ, **de Oliveira T**, ***Genes*** (2020), doi: https://doi.org/10.3390/genes11080949:11(8), 949.

35. **Early transmission of SARS-CoV-2 in South Africa: An epidemiological and phylogenetic report.** Giandhari J, Pillay S, Wilkinson E, **Tegally H**, Sinayskiy I, Schuld M, Lourenço J, Chimukangara B, Lessells R, Moosa Y, Gazy I, Fish M, Singh L, Khanyile KS, Fonseca V, Giovanetti M, Alcantara LCJ, Petruccione F, **de Oliveira T**, ***IJID*** (2020), doi: https://doi.org/10.1016/j.ijid.2020.11.128:.

Acknowledgements

The work described in this thesis happened largely during the response to the SARS-CoV-2 pandemic. Consequently, this research would not have been possible without the support of many, which I would like to acknowledge. First, I am grateful for the relentless support and generous guidance from my supervisor Tulio de Oliveira. His dedication to upholding equity towards science coming out of Africa during the pandemic was a driving force during times where this research had to be performed under pressure for pandemic response. His belief in my abilities was motivation to strive for excellence at every stage. Finally, Tulio taught me scientific rigour, encouraging me to always question remarkable results lest they are remarkable errors.

I owe thanks to Richard Lessells, a mentor and colleague who has challenged me scientifically to always bring results back to their public health significance. Richard has also shown considerate care and level-headed guidance on multiple occasions, for which I am grateful. I am immensely grateful for my colleagues within the KwaZulu-Natal Research Innovation and Sequencing Platform (KRISP) at UKZN and the Centre for Epidemic Response and Innovation (CERI) at Stellenbosch University alongside whom I have worked everyday during this PhD and who have consistently provided a collegial and collaborative environment. In alphabetical order they are Cheryl Baxter, Jennifer Giandhari, Monika Moir, Yeshnee Naidoo, Sureshnee Pillay, Jenicca Poongavanan, James Emmanuel San, Lavanya Singh, Eduan Wilkinson, Joicymara Xavier

I also thank the Network for Genomic Surveillance in South Africa (NGS-SA) and colleagues across the Africa PGI for their trust and close collaborations during the pandemic. Equally, I am grateful to have engaged closely with scientists in the field of viral genomic epidemiology globally on research work conducted towards pandemic response and presented in this thesis. It has been a humbling experience to be able to contribute to scientific knowledge alongside respected researchers during one of the biggest challenges of this century. In alphabetical order, these scientists are: Matthew Cotten, Marta Giovanetti, Moritz Kraemer, Darren Martin, Gerald Mboowa, Aine O'Toole, Oliver Pybus, Andrew Rambaut, Alex Sigal, Sofonias Tesema.

Finally, none of this would have been possible without the support of family and friends at home in Mauritius. First, I owe special thanks to my husband, Nadeem Mungloo who has been nothing but patient and caring during the many months of long research hours. My father's (Mehaad Tegally) and brother's (Baahir Tegally) encouragement have been invaluable. My mother's (Husna Tegally) lifelong example of hard work and tenacity has been my North Star throughout my scientific journey, even though she is not with us anymore to witness this milestone. Last but certainly not least, I owe thanks to Soufia Bham and Shabneez Badal for a lifetime of friendship, joy and enthusiastic support.

Chapter 1: Introduction

Background

Viruses

Viruses are obligate intracellular parasites that employ a range of different mechanisms to enter and replicate within host cells. They are the largest known group of organisms on the planet comprising close to 100 different families. This diversity is in turn classified into 7 recognized groups of the Baltimore system based on the organisation of their genetic material. The encoding messenger RNA can be produced from RNA or DNA, it can be single-stranded (ss) or double-stranded (ds) and, if it is single-stranded, its polarity can either be positive (5'-->3') or negative (3'-->5'). The 7 groups of viruses identified by the Baltimore classification system are known as groups I-VII and correspond to dsDNA, ssDNA, dsRNA, (+)ssRNA, (-)ssRNA, ssRNA-RT and dsDNA-RT viruses respectively.

The association between viruses and disease is described by two of their properties: pathogenicity and virulence. Pathogenicity simply defines the capability of an infectious agent to cause illness while virulence is a quantitative measure of the severity of the infection. Viruses become pathogenic for a certain host when they can enter, multiply inside and damage host cells [1]. This requires complex biological requirements, such as viral antigens matching host entry or replication receptors, resulting in only a minority of viruses possessing pathogenic abilities against particular hosts. Virulence is then mediated not only by viral success at replication and damage inside host cells but also by an array of host factors such as immune response [1].

Viruses have a compelling scale of genetic diversity, even within the same families. A core feature of RNA viruses particularly is their high evolutionary rate due to lack of a proofreading mechanism in RNA polymerases, causing high error rates or mutational frequency, which leads to mutation rates in the order of 10^{-4} per base per round of replication [2–4]. This, along with other mechanisms like recombination, allows RNA viruses to rapidly adapt to different environments with changes in virulence, epidemiology, or competence of vectors. It is this property that, for instance, allows for zoonotic spillover events where a virus not previously able to infect human cells gains that ability through random mutations. High error rates are directly linked to the smaller nature of RNA virus genomes. The average size of RNA viral genomes can be as small as 9 kb for and goes up to 29 kb for Coronavidae viruses. The link between the small size of RNA viruses and their high mutation rates has been explained by the knowledge that high error rates in larger genomes would inevitably cause lethal mutations at replication [5]. The small genome size is somewhat compensated by genome compression which results in gene overlaps [6].

SARS-CoV-2 genomic characteristics & evolution

The severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is a RNA virus of genome length of 29903 nucleotide bases belonging to the *Sarbecovirus* subgenus of *Coronaviridae* family and newly identified in December 2019 during an outbreak of SARS-like pneumonia cases in the city of Wuhan (Hubei Province) in China [7,8]. Other members of the coronavirus family include endemic viruses which cause mild symptoms in humans, often lumped into the grouping of the "common cold", while others have caused more serious outbreaks with important case fatality rates in humans, for example SARS and MERS in 2002-2004 [9] and 2012-2019 [10] respectively. At the beginning of 2020, the WHO declared SARS-CoV-2 as a public health emergency and later a pandemic [11]. Two years later, at the beginning of 2022, this virus had caused at least 400M recorded infections and 5M recorded deaths with cyclical epidemic waves still progressing in many parts of the world at the time of writing.

Genomic sequencing in Wuhan, China showed that the new coronavirus was typical of betacoronaviruses with the following genes: replicase ORF1ab, spike (S), envelope (E), membrane (M) and nucleocapsid (N), with 265 nt at the 5′ terminal end and 229 nt at the 3′ terminal end [12]. Analyses of its ancestry revealed that SARS-CoV-2 appeared to have emerged through recombination of bat sarbecoviruses, meaning that different segments of the genome originate from different viruses. The novel genomic structure allowed the virus to express new residues in the receptor binding domain (RBD) that enabled binding to human ACE2 receptors and therefore a zoonotic jump from bats or an intermediate host to cause infections and disease in humans, a new host. This is similar to the emergence of SARS-CoV and MERS-CoV. In fact, while coronaviruses are known to have slightly lower mutation rates due to the presence of an exonuclease, they are also well described for their ability to recombine with other members of the family during co-infection of the same hosts [13].

Throughout its global dissemination over an ongoing pandemic of already two years, the SARS-CoV-2 virus has had the opportunity to evolve extensively and this evolution has been meticulously monitored through viral genomic sequencing. For the first year or so, SARS-CoV-2 disseminated as lineages with mutations accumulating slowly, particularly on the spike gene of the virus, and thus without much effect on its phenotypic properties. Chapter 2 of this thesis describes the initial lineages of SARS-CoV-2 that circulated during the first wave of the epidemic in South Africa. The only exception was when the aspartic acid amino acid at position 614 on the spike was replaced by glycine (D614G) around February 2020 at the start of a massive expansion of the outbreak in Europe [14]. This mutation was retrospectively characterised as being associated with higher transmissibility of the virus [15]. Towards the end of 2020, more considerable mutational changes appeared within the spike protein, presumably from selective pressure [16], giving rise to what became known as variants of concern [17]. Chapters 3 and 4 describe research work that lead to the discovery and epidemic expansion of two of those variants of concern in South Africa, the

Beta and Omicron variants respectively. These variants of SARS-CoV-2 each possessed varying degrees of increased transmissibility and/or immune escape, thus making them better adapted for pathogenicity in humans.

Emerging viral pathogens

Over the past century, the understanding, prevention and treatment of communicable diseases have seen considerable progress. In a systematic analysis of global burden of disease between 1990-2019, the ten most important contributors to declining burden globally and combined for all ages included six infectious diseases (lower respiratory infections, diarrhoeal diseases, measles, tetanus, malaria and tuberculosis) [18]. While this contributes to the general shift of disease burden from communicable to non-communicable diseases globally (epidemiological transition), the situation for specific countries and regions vary widely. While the United States, western Europe and Australia, for example, had 40-50% of all DALYs (disability adjusted life year) made up of non-communicable diseases and injury before the pandemic, African countries largely had only around 10-20% of the same, showing a larger burden of infectious diseases in the latter. Yet, even globally, communicable illnesses have significant human and economic costs. For the year 2016 alone, the estimated combined burden of eight major infectious diseases (HIV/AIDS, malaria, measles, hepatitis, dengue fever, rabies, tuberculosis and yellow fever) was more than 156 million life years lost, which corresponded to a staggering cost of US$8 trillion [19].

A crucial consideration here is that this was evaluated before the massive overhaul of an infectious diseases pandemic (SARS-CoV-2) of over two years in 2020.The current coronavirus pandemic painfully illustrates how rapidly an infectious agent can cause chaos to global health and intensely impact both human lives and the economy. This phenomenon was neither unpredictable nor completely novel, although its scale has been unprecedented for modern society. Rather, the SARS-CoV-2 pandemic highlights the ongoing threat of emerging and re-emerging pathogens.

Emerging or re-emerging infectious diseases are those that have recently been identified in a population, those with rapidly rising incidence or geographic range or those of an increasing threat for the near future. They can be caused by novel infectious agents, often as a result of zoonoses, known pathogens that have spread to new locations or populations or that have caused previously unrecognised disease outcomes, or by the re-emergence of known agents whose incidence had largely subsided but started appearing again. Given the unpredictability of emergence, this type of infectious disease is known to have the potential for severe global impact such as pandemics, epidemic outbreaks, intense pressure on health systems, large mortality and morbidity burdens and general global instability.

Although technically, any microbial species could be the causative pathogen for emerging infectious diseases, the rapid evolution of viruses and their pathogenic nature tend to make them of particular threat to public health response. Unlike bacteria and other pathogenic microbes that can be treated with generic drugs, drugs and vaccines targeting viral pathogens can only be developed after identifying the virus (not immediately useful for spontaneous outbreaks – e.g. coronavirus, ebola). The current coronavirus pandemic highlights how rapidly viruses can spread as a result of increased globalisation and travel. Like SARS-CoV-2, the majority of viruses are transmitted between susceptible hosts through the respiratory tract. However, emerging infectious diseases can also transmit and spread differently. They can be foodborne, vector borne, sexually-transmitted or airborne and their global reach potentially made more serious by evolving conditions such as modern human behavior, global warming (diversification of vector or pathogen host environments) and greater connectedness between corners of the world (more difficult geographical containment). Furthermore, the threat of emerging novel pathogens becomes significantly more important in a setting of accelerating climate change, predicted to increase the risk of viral spillover events across species and ultimately into human transmission [20].

Before SARS-CoV-2, some of the most concerning emerging infectious diseases to humans in the last decade were caused by viral agents, including SARS in 2002-2004 [9], Ebola in 2014 [21] and Zika in 2016 [22]. Emerging and re-emerging pathogens pose several challenges to diagnosis, treatment, and public health surveillance. Identification of an emerging pathogen by conventional methods is difficult and time-consuming due to the "novel" nature of the agent, requiring a large array of techniques including cell cultures, inoculation of animals, cultivation using artificial media, histopathological evaluation of tissues (if available), and serological techniques using surrogate antigens. A strategy conceptualised to improve the characterization and response to emerging infectious diseases is the One Health approach aimed at considering human, animal and environmental health together to monitor and respond to public health threats and to learn how diseases spread in these three settings [23,24]. Within this approach, three stages of infectious diseases emergence are monitored: pre-emergence, localised emergence and pandemic emergence. At each of these stages, sequencing virus genomes can play a role in characterising and responding to outbreaks.

Virus genomes in Epidemiology

Early infectious disease outbreak investigations and response relied heavily on epidemiological data and routine laboratory testing of symptomatic patients to describe viral transmission events. These methods became insufficient in light of the rapid spread of viral transmitted infectious diseases. Their rapid molecular evolution and short generation time resulted in viral progeny of unique molecular architecture, sometimes capable of evading established control measures. This created a need for more developed approaches to reporting the emergency, spread, and evolution of viruses. There was also a need to understand the implication of evolutionary changes of these

viruses on phenotypic characteristics linked to complex host and pathogen traits. This is where sequencing of virus genomes and genomics came in.

Viral genomic sequencing is the process of reading the genetic code of a virus. In the past couple of decades, viral molecular sequences have played a growing role in understanding and mitigating emerging epidemics. Within disease outbreaks or epidemics, sampled genomes allow pathogen identification if novel, description of genetic diversity of viruses circulating, reconstruction of epidemic origins, rates of transmission, and vaccine development and drug design. Genomics is, in turn, the study of those sequenced genomes, and bioinformatics is the development of methods and tools to enable genomics, given that genomic data is often large and complex. The infusion of sequencing, and in the same thread genomics and bioinformatics, with epidemiology has resulted in an important public health tool to support epidemic responses and to elucidate the clinical and therapeutic relevance of genomic variants [25]. The various arms of this are described below.

Rapid-response diagnostics

The sequencing of virus genomes has played a key role in the rapid diagnosis of pathogenic agents in many infectious disease outbreaks. The first SARS-CoV-2 genome was published in January of 2020 [12], a month after the outbreak was reported in the Wuhan market in China. The early characterization of this novel virus' genome has been pivotal to control its spread through diagnostics development [26,27] and in informing vaccine and therapeutics development [28,29]. This is of particular importance to public health as traditional surveillance of emerging viral pathogens lacks the capacity to characterise highly divergent or novel pathogens [30]. In other cases, symptoms presented by patients with certain viral infections are often highly similar, for example, arboviruses such as Zika, Dengue, Chikungunya and Yellow fever all present with minor variations of the same symptoms that can easily result in wrong diagnosis and treatment if the infecting viral genome is not characterised [31].

Transmission Dynamics

Genetic sequence data of viruses can provide important insights into individual person-to-person transmission events [32]. Transmission-focused investigations such as the source and direction of transmission between infected hosts are key to informing activities such as contact tracing and isolation of hosts as a control measure. Genomic data can be used to support epidemiology in the hypothesis of linked infections [33]. Furthermore, with the use of next-generation sequencing (NGS) technologies, the presence of minority variants within patient viral populations can also be investigated. Several studies have relied on the utility of minority variants in informing the evolution and transmission of SARS-CoV-2. For example, Shen et. al (2020) [34] and Lythgoe et. al (2021) [35] investigated possible transmission events and diversity of SARS-CoV-2 using minority alleles while Popa et. al (2020) [36] used minor alleles to determine the dynamics of viral spread in

Austria. In parallel, minority analysis can also help to investigate co-infections of a single host with multiple virus genotypes and intra-host evolution of chronics infections [37].

Outbreak Investigation

In outbreak investigations, genomics adds a layer of support to epidemiological evidence in answering questions about the source of introduction of the pathogen and understanding transmission dynamics in the early stages, which is essential for informing effective outbreak control. This typically involves a series of methodological steps including sampling, sequencing, and genome assembly of outbreak and non-outbreak control isolates, mapping of genomic differences and inferring the relationship between the sequences, and evaluating the genomic linkages in the context of known epidemiological connections [24]. This process has been the foundation of numerous outbreak investigations during the SARS-CoV-2 pandemic - in airports, in universities, in farms and more [38,39]. This thesis contributed, for example, to solving the re-emergence of SARS-CoV-2 in the island of Mauritius after 9 months of no recorded COVID infections and ongoing government-managed 14-day quarantine of incoming passengers using genomic sequencing of outbreak samples [40].

Phenotypic studies/ Immune escape studies

Genetic alterations to pathogens during outbreaks, epidemics or pandemics can result in important phenotypic changes. Using genomic sequencing to track these mutations and inform phenotype-genotype studies is critical to mitigate the challenges that shifts in pathogen behaviour can pose to control measures. This includes pathogens developing complete or partial resistance to drugs or vaccines or evolving better pathogenic abilities to expand epidemics. Such genetic characterization has been critical in studying drug resistance to antiretrovirals in HIV treatment [41]. The genotype-phenotype association was also studied early in the SARS-CoV-2 pandemic [42] and more extensively when its impact was more seriously felt after variants of the virus caused new and bigger waves of infections worldwide [43–46]. Questions arose as to whether these variants were more transmissible, more virulent, or would cause more re-infections. Genomic sequencing has been key in answering these questions.

By studying mutations identified through sequenced viral genomes, scientists were able to quickly show that new SARS-CoV-2 variants were either better at binding to human ACE2 receptors *in-vitro*, hence supporting faster transmission, or better at evading immunity from neutralising antibodies, hence hinting towards easier re-infections. There were also more hopeful conclusions. In cohort studies, for example, sequencing demonstrated that T-cells against SARS-CoV-2 retained their activity against novel variants, suggesting that people with previous immunity would have less chances of suffering from severe disease even if re-infected with a new variant [47,48].

SARS-CoV-2 evolution also threatened to endanger plans to utilise fast-tracked and very effective vaccines as an emergency to curb an ongoing pandemic. Genomic sequencing played a key role in evaluating the effectiveness of various SARS-CoV-2 vaccines against these variants in human trials. Studies in South Africa, for example, reported no protection from mild to moderate infection from the Beta variant after two doses of ChAdOx1 nCoV-19 [49] and breakthrough infections in recipients of the Pfizer–BioNTech and Moderna vaccines [50]. Fortunately, other studies [51] involving other variants have registered more promising results. Together these studies highlight the usefulness of genomic sequencing in establishing the critical link between evolving genotype and phenotypes, even in the ongoing pandemic.

Molecular characterization of pathogen

Genomic sequencing of pathogens during an outbreak or epidemic can allow scientists to characterise key aspects of the evolution of a viral pathogen. For instance, by analysing the number and types of genetic changes happening in sampled viral genomes over time, scientists can estimate evolutionary rates of emerging viral pathogens and reconstruct their evolutionary path within transmission chains. This helps establish critical knowledge around an emerging pathogen, which will in turn permit robust inferences and estimates during outbreak investigation and inform more public health responses in a more accurate manner.

Establishing a virus' molecular clock is also tremendously important when characterising a pathogen. The molecular clock hypothesis describes the accumulation of mutations as being roughly constant with time. This was deduced after observation that the amount of genetic change between two organisms is proportional to the duration of time separating the two organisms from their last common ancestor [52,53]. Molecular clocks are at the foundation of methods used in outbreak investigations and to build time-scaled phylogenies. This time component is critical given that it provides estimations of the time of origin of viruses or the time when last common ancestors existed, which informs response to emerging pathogens, or preparedness against novel pathogens, in a plethora of ways [54].

Genomic Surveillance and Genomic Epidemiology in real-time

Well before the SARS-CoV-2 pandemic, scientists and health experts in the field called for a strengthening of disease surveillance systems [55,56] and, with the democratisation of sequencing, a shift towards a genomics-informed, real-time, global pathogen surveillance system, particularly following the costly Ebola and Zika outbreaks in 2014 and 2015 respectively [24]. Genomic surveillance allows for a framework where sequencing, genomics and bioinformatics can be leveraged in the ways discussed above, and more, to start solving the unpredictable nature of emerging and re-emerging infectious diseases. It can act as an important warning system to

implement necessary actions and reduce the impact of outbreaks. The implementation of genomic surveillance strategies generally have as objectives to understand outbreak dynamics, to perform molecular mapping of viral spread and ultimately to execute a genomics-informed outbreak response. To achieve this objective, genomic surveillance is poised to happen at several points of the disease emergence timeline, each one happening on a different cumulative scale and each having a specific goal towards mitigation of public health emergencies. These are, for instance, wildlife surveillance of pathogen vectors or hosts to identify emergence likelihood hotspots, such as bat surveillance efforts [57], rapid-response metagenomics diagnostics of outbreaks of unknown causative agent [58], and finally in the context of genomic epidemiology. Genomic epidemiology refers to the use of genome sequencing to understand infectious disease transmission and epidemiology. It involves reconstructing transmission events from genomic data to form a picture of epidemic dynamics, often utilising methods of phylogenetics and phylodynamics and contributes to knowledge both with regards to molecular characterization of the pathogen and to give a picture of how local and global epidemic patterns compare, which can ultimately guide public health response. Phylodynamics refers to analyses where epidemiological and evolutionary dynamics of pathogens are considered simultaneously for powerful investigations towards epidemic preparedness or response. This is a concept first proposed by Grenfell et al. [59] and now key to infectious disease research, which captures how epidemiological, immunological and evolutionary processes act to modulate the evolving phylogeny of pathogens. Genomic surveillance and genomic epidemiology for infectious disease emergence has been utilized a number of times before. This has included, for example, investigation of the epidemiology of Ebola in West Africa [60], Zika in the Americas [31,61,62], and SARS in Asia [63].

However, this implementation has never been at the scale and stakes that it has been during the current SARS-CoV-2 pandemic. In the past, there have been a number of obstacles to implementing a global pathogen genomics surveillance system. Challenges exist along much of the pipeline from sampling a pathogen, sequencing the pathogen, bioinformatics analysis of the genomic data to ultimately making it useful for public health response. An important one has been the general inaccessibility of genomic sequencing in settings where outbreaks of emerging infectious diseases most often occur. Luckily, the advent of more affordable and portable sequencing technologies started to help overcome this challenge. In 2014, the MinION from Oxford Nanopore Technologies was released on a model of a free dongle-type instrument powered by USB connection to a laptop and consumers only pay for reagents and flow-cells [64]. Since then, the MinION has been used for pathogen genomics in the field and on the go, for example during the Ebola epidemic [65,66] and Zika real-time mobile laboratory in Brazil (ZiBRA) [67]. Yet, challenges of scaling up genomic surveillance for communities where it matters more remain. Even more affordable and easier to manage technologies require an equitable distribution of funding, robust capacity building and constructive engagement with public health officials for implementation in low and middle income settings where increasing accessibility to genomic surveillance would have

truly valuable benefits to disease control. Strategies to optimise the scaling up of genomic surveillance are elaborated upon in the next subsection.

For genomic data to become useful to public health, particularly in emergencies like the current pandemic, it is critical that the genome sequences be openly shared to allow for thorough genomic epidemiology tool development and analyses that will inform rapid and accurate responses. For the genomic epidemiology investigation of extensive outbreaks, epidemic and pandemics, this publicly-shared data must ideally cover a sampling from the whole geographical and temporal extent of the outbreak and must be accompanied by informative clinical metadata. In the past however, there have been serious concerns around the feasibility and ethics of such an endeavour. Genomic data come directly from patient samples and thus completely unrestricted access poses issues related to patient privacy, safety, consent, concerns around inadequate data regulations, reuse by third parties, and scientific or political disincentives to releasing data [68–70]. While rapid data sharing had happened to some extent during previous epidemics, the current pandemic demonstrated a real commitment to safe and ethical open access and open data from the academic and public health institutions and other stakeholders of the pathogen genomics community, upheld by GISAID [71], the main database for genomic data sharing during the pandemic. As of the beginning of October 2022, two and a half years roughly since the recognition of the SARS-CoV-2 outbreak as a public-health emergency and pandemic, GISAID had facilitated the public sharing of more than 13 million SARS-CoV-2 genomic sequences from all corners of the globe, 215 countries and territories to be exact. This unprecedented momentum of data sharing provided an opportunity to track the evolution of the SARS-CoV-2 like never before.

Aims and Objectives

The main objectives of research conducted as part of this thesis were to apply innovative methods of genomic surveillance and genomic epidemiology during the SARS-CoV-2 pandemic in South Africa and Africa to answer key questions for public health response that relied on an understanding of the ongoing molecular evolution of the virus. The specific questions to be answered included the following:

- What are the sources of introduction of the virus in South Africa and Africa?
- When was the virus introduced?
- How many introductions of the virus are occuring from abroad?
- When did community transmission start?
- Which lineages are circulating in a region?
- Are any mutations emerging in local outbreaks?
- Are we detecting any known or emerging variants?
- How are emerging variants emerging, evolving and spreading?

Study Design & Methodology

Genomic data was obtained from the whole genome sequencing of remnant diagnostics samples that tested positive for SARS-CoV-2. These samples were collected during the course of the evolving pandemic. Research conducted towards this thesis analysed the genetic characteristics and evolution of the virus in order to investigate the transmission dynamics within outbreaks and to determine how molecular changes in the virus are affecting the progression of epidemic waves. This was achieved using three main arms of methodological techniques, as outlined below. More in-depth methodological description accompanies each chapter where respective results are presented.

Genome Assembly

The raw reads from sequencing were assembled by aligning to a reference genome, consensus variant calling and correcting for sequencing artefacts to generate complete or near complete pathogen genomes. Raw reads coming from both Nanopore and Illumina sequencing were assembled using Genome Detective (https://www.genomedetective.com/) and the embedded Coronavirus Typing Tool [72]. As required, the initial assemblies obtained from Genome Detective were polished by aligning mapped reads to the references and filtering out low-quality mutations, and those were confirmed visually with bam files using Geneious software (Biomatters Ltd, New Zealand). All of the sequences will be deposited in GISAID (https://www.gisaid.org/).

Genomic Epidemiology

Phylogenetic analysis (using Maximum Likelihood and Bayesian methods) of these genomes was carried out in conjunction with a representative reference set to determine clades and lineages of interest and determine the relatedness and molecular characteristics of samples. This allowed the generated genomic data to be analysed in context of prevailing outbreaks or epidemics.

Reference dataset

Sequences were downloaded from the GISAID sequence database (https://www.gisaid.org/) as of the specific dates of analysis to serve as reference dataset for the phylogenetic analysis of genomes assembled as part of this thesis. Metadata associated with these sequences were also obtained from GISAID. To ensure high accuracy of results, only the highest quality of genomic data were included. Sequences that were <25kbp in length as well as sequences with a high proportion of ambiguous sites (>5%) were filtered out. Additionally, sequences that lacked any geographic and or sampling date information were also not considered. Due to the large size of the GISAID dataset, specific schemes were used for genomic dataset subsampling as per the needs of the research questions being answered. One strategy is to scale the number of sequences in the reference set by the size of the epidemic in sampling countries. These are more thoroughly described in each chapter.

Lineage, clade and variant classification

The best available lineage classification methods at each respective point along the thesis research were used for sequence analysis. A dynamic lineage classification method was proposed by Rambaut et al. early on in the pandemic [73] via the Phylogenetic Assignment of named Global Outbreak LINeages (PANGOLIN) software suite (https://github.com/hCoV-2019/pangolin) [74]. This is aimed at identifying the most epidemiologically important lineages of SARS-CoV-2, allowing researchers to monitor the epidemic in a particular geographical region more effectively. Two main SARS-CoV-2 lineages were initially recognized; lineage A, defined by Wuhan/WH04/2020 and lineage B, defined by Wuhan-Hu-1 strain. Although Wuhan-Hu-1 was the first published genome from SARS-CoV-2, it is classified as lineage B. Phylogenetic analyses of SARS-CoV-2 identified sequences from lineage A to be more closely related to a bat coronavirus, which suggest this to be the first lineage (hence A). Sub-lineages can further diversify into sub sub-lineages (e.g. A·1·1). The Nextstrain and Nextclade systems of SARS-CoV-2 clade nomenclature [75] were also used where appropriate, particularly when variants of concern emerged and circulated as bigger clades of interest. Following the global circulation of VOCs and associated infection waves, the WHO designated official Greek letter nomenclature to clades and lineages of public health concern or interest [76]. Classification of sequences as lineages, clades or variants were often interchangeable and the use of nomenclature highly specific to the intended audience of related research or scientific communication.

Phylogenetic analysis

In order to reconstruct the evolutionary history of sampled SARS-CoV-2 genomes, phylogenetic trees were inferred using a variety of relevant techniques. First, sequences were aligned to one another and to the reference using either MAFFT [77] or Nextalign [78]. For preliminary analyses, and those requiring the incorporation of a large set of genomes, maximum likelihood (ML) tree topologies were constructed, either in IQ-TREE (GTR+G+I, no support) or Fasttree. Resulting ML tree topologies were usually transformed into a time scaled phylogenies where necessary using TreeTime [79] with a clock rate of 8x10-4 and rooted along the branch of Wuhan-WH04 (GISAID: hCoV19/Wuhan/WH04/2020) and Wuhan Hu1 (Genbank: MN908947). At times, the Nextstrain build pipeline was used to construct interactive SARS-CoV-2 phylogenies [80]. SARS-CoV-2 phylogenetic trees were used to identify clusters of community transmission in a certain location, identify lineages of interest and track viral transmission chains in South Africa or Africa.

Based on these large phylogenies of SARS-CoV-2, it was also possible to infer viral importation and exportation routes, in order to identify sources of introductions for instance. To do this, a *mugration* model was fitted on the resulting time-scaled tree topology in TreeTime [79], mapping country locations to tips and internal nodes. The resulting annotated tree topology was used to infer the number and origin of viral introductions into target countries (e.g. South Africa) through time.

For clusters of interest or emerging lineages, it may become necessary to estimate precise dates of origin and reconstruct their spatiotemporal dispersal pattern accurately. For this purpose, Bayesian coalescent analyses were performed. The purpose of these analyses were mainly to: (i) estimated date of origin for SARS-CoV-2 lineages and variants, (ii) infer the estimated date to the most recent common ancestor (MRCA) for major lineages and variants, (iii) infer the estimated dates of viral introductions into specified countries or regions in a Bayesian framework, and (iv) reconstruct the phylodynamics and phylogeography of viral dispersal.

First, preliminary ML tree topologies were constructed as described above and used to analyse the molecular clock signal for the viral lineage in TempEst software suite [81]. Coalescent molecular clock analyses were performed on clusters with significantly high temporal signal in BEAST software [82,83]. In short, analyses were run either under a strict assumption at a constant evolutionary rate of 8x10-4 nucleotide substitutions per site per year, or a relaxed molecular clock, depending on the purpose and context of the analysis, and an exponential growth coalescent tree prior. The Markov Chains were usually run in duplicate for a total length of 100 million steps sampling every 10,000 iterations in the chains, or until the runs converged. Runs were assessed in Tracer for good convergence (ESS>200) and TreeAnnotator after discarding 10% of runs as burn-in.

Data Visualisation

Throughout research conducted towards this thesis, data visualisation concepts were developed and used to carefully and creatively present results of genomic epidemiology. Methods used included data analysis and visualisation software such as R ggplot [84] and ggtree [85], Seraphim [86], Figtree (http://tree.bio.ed.ac.uk/software/figtree/). Genomics and phylogenetics results were integrated with various epidemiological datasets for context and optimal communication of genomic epidemiology significance of these results. Epidemiological datasets used in this research included Our World in Data COVID-19 repository (https://github.com/owid/covid-19-data/tree/master/public/data), aggregated COVID-19 incidence data from South Africa by the Data Science for Social Impact Research Group @ University of Pretoria group (https://github.com/dsfsi/covid19za), excess death reports from the South Africa Medical Research Council (SA-MRC) (https://www.samrc.ac.za/reports/report-weekly-deaths-south-africa), and estimates of effective daily reproductive values (Re) from the COVID-19-Re data repository (https://github.com/covid-19-Re/dailyRe-Data) [87].

References

1. Hibbs, J. & Young, N. S. Viruses, virulence and pathogenicity. *Baillieres Clin Haematol* **8**, 1–23 (1995).

2. Drake, J. W. & Holland, J. J. Mutation rates among RNA viruses. *Proc Natl Acad Sci*

USA **96**, 13910–13913 (1999).

3. Mansky, L. M. In vivo analysis of human T-cell leukemia virus type 1 reverse transcription accuracy. *J. Virol.* **74**, 9525–9531 (2000).

4. Crotty, S., Cameron, C. E. & Andino, R. RNA virus error catastrophe: direct molecular test by using ribavirin. *Proc Natl Acad Sci USA* **98**, 6895–6900 (2001).

5. Holmes, E. C. Error thresholds and the constraints to RNA virus evolution. *Trends Microbiol.* **11**, 543–546 (2003).

6. Belshaw, R., Pybus, O. G. & Rambaut, A. The evolution of genome compression and genomic novelty in RNA viruses. *Genome Res.* **17**, 1496–1504 (2007).

7. Li, Q. *et al.* Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus-Infected Pneumonia. *N. Engl. J. Med.* **382**, 1199–1207 (2020).

8. Zhou, P. *et al.* A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* **579**, 270–273 (2020).

9. Cherry, J. D. & Krogstad, P. SARS: the first pandemic of the 21st century. *Pediatr. Res.* **56**, 1–5 (2004).

10. de Wit, E., van Doremalen, N., Falzarano, D. & Munster, V. J. SARS and MERS: recent insights into emerging coronaviruses. *Nat. Rev. Microbiol.* **14**, 523–534 (2016).

11. Cucinotta, D. & Vanelli, M. WHO Declares COVID-19 a Pandemic. *Acta Biomed* **91**, 157–160 (2020).

12. Wu, F. *et al.* A new coronavirus associated with human respiratory disease in China. *Nature* **579**, 265–269 (2020).

13. Boni, M. F. *et al.* Evolutionary origins of the SARS-CoV-2 sarbecovirus lineage responsible for the COVID-19 pandemic. *Nat. Microbiol.* **5**, 1408–1417 (2020).

14. Korber, B. *et al.* Tracking Changes in SARS-CoV-2 Spike: Evidence that D614G Increases Infectivity of the COVID-19 Virus. *Cell* **182**, 812-827.e19 (2020).

15. Volz, E. *et al.* Evaluating the Effects of SARS-CoV-2 Spike Mutation D614G on Transmissibility and Pathogenicity. *Cell* **184**, 64-75.e11 (2021).

16. Martin, D. P. *et al.* The emergence and ongoing convergent evolution of the SARS-CoV-2 N501Y lineages. *Cell* **184**, 5189-5200.e7 (2021).

17. Tao, K. *et al.* The biological and clinical significance of emerging SARS-CoV-2 variants. *Nat. Rev. Genet.* **22**, 757–773 (2021).

18. GBD Collaborators. Global burden of 369 diseases and injuries in 204 countries and territories, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019. *Lancet* **396**, 1204–1222 (2020).

19. Armitage, C. The high burden of infectious disease. *Nature* **598**, S9 (2021).

20. Carlson, C. J. *et al.* Climate change increases cross-species viral transmission risk. *Nature* **607**, 555–562 (2022).

21. Buseh, A. G., Stevens, P. E., Bromberg, M. & Kelber, S. T. The Ebola epidemic in West Africa: challenges, opportunities, and policy priority areas. *Nurs. Outlook* **63**, 30–40 (2015).

22. Petersen, L. R., Jamieson, D. J., Powers, A. M. & Honein, M. A. Zika Virus. *N. Engl. J. Med.* **374**, 1552–1563 (2016).

23. One Health | CDC. https://www.cdc.gov/onehealth/index.html.

24. Gardy, J. L. & Loman, N. J. Towards a genomics-informed, real-time, global pathogen surveillance system. *Nat. Rev. Genet.* **19**, 9–20 (2018).

25. Traynor, B. J. The era of genomic epidemiology. *Neuroepidemiology* **33**, 276–279

(2009).

26. Wang, Y., Kang, H., Liu, X. & Tong, Z. Combination of RT-qPCR testing and clinical features for diagnosis of COVID-19 facilitates management of SARS-CoV-2 outbreak. *J. Med. Virol.* **92**, 538–539 (2020).

27. Vogels, C. B. F. *et al.* Multiplex qPCR discriminates variants of concern to enhance global surveillance of SARS-CoV-2. *PLoS Biol.* **19**, e3001236 (2021).

28. Kames, J. *et al.* Sequence analysis of SARS-CoV-2 genome reveals features important for vaccine design. *Sci. Rep.* **10**, 15643 (2020).

29. Kyriakidis, N. C., López-Cortés, A., González, E. V., Grimaldos, A. B. & Prado, E. O. SARS-CoV-2 vaccines strategies: a comprehensive review of phase 3 candidates. *npj Vaccines* **6**, 28 (2021).

30. Abat, C., Chaudet, H., Rolain, J.-M., Colson, P. & Raoult, D. Traditional and syndromic surveillance of infectious diseases and pathogens. *Int. J. Infect. Dis.* **48**, 22–28 (2016).

31. Faria, N. R. *et al.* Establishment and cryptic transmission of Zika virus in Brazil and the Americas. *Nature* **546**, 406–410 (2017).

32. Campbell, F., Strang, C., Ferguson, N., Cori, A. & Jombart, T. When are pathogen genome sequences informative of transmission events? *PLoS Pathog.* **14**, e1006885 (2018).

33. San, J. E. *et al.* Transmission dynamics of SARS-CoV-2 within-host diversity in two major hospital outbreaks in South Africa. *Virus Evol.* **7**, veab041 (2021).

34. Shen, Z. *et al.* Genomic Diversity of Severe Acute Respiratory Syndrome-Coronavirus 2 in Patients With Coronavirus Disease 2019. *Clin. Infect. Dis.* **71**, 713–720

(2020).

35. Lythgoe, K. A. *et al.* SARS-CoV-2 within-host diversity and transmission. *Science* **372**, (2021).

36. Popa, A. *et al.* Genomic epidemiology of superspreading events in Austria reveals mutational dynamics and transmission properties of SARS-CoV-2. *Sci. Transl. Med.* **12**, (2020).

37. Karim, F. *et al.* Persistent SARS-CoV-2 infection and intra-host evolution in association with advanced HIV infection. *medRxiv* (2021) doi:10.1101/2021.06.03.21258228.

38. Lu, L. *et al.* Adaptation, spread and transmission of SARS-CoV-2 in farmed minks and associated humans in the Netherlands. *Nat. Commun.* **12**, 6802 (2021).

39. Aggarwal, D. *et al.* Genomic epidemiology of SARS-CoV-2 in a UK university identifies dynamics of transmission. *Nat. Commun.* **13**, 751 (2022).

40. Tegally, H. *et al.* A Novel and Expanding SARS-CoV-2 Variant, B.1.1.318, dominates infections in Mauritius. *medRxiv* (2021) doi:10.1101/2021.06.16.21259017.

41. Capobianchi, M. R., Giombini, E. & Rozera, G. Next-generation sequencing technology in clinical virology. *Clin. Microbiol. Infect.* **19**, 15–22 (2013).

42. Starr, T. N. *et al.* Deep Mutational Scanning of SARS-CoV-2 Receptor Binding Domain Reveals Constraints on Folding and ACE2 Binding. *Cell* **182**, 1295-1310.e20 (2020).

43. Khan, K. *et al.* Omicron sub-lineages BA.4/BA.5 escape BA.1 infection elicited neutralizing immunity. *medRxiv* (2022) doi:10.1101/2022.04.29.22274477.

44. Cao, Y. R. *et al.* B.1.1.529 escapes the majority of SARS-CoV-2 neutralizing

antibodies of diverse epitopes. *BioRxiv* (2021) doi:10.1101/2021.12.07.470392.

45. Cele, S. *et al.* SARS-CoV-2 Omicron has extensive but incomplete escape of Pfizer BNT162b2 elicited neutralization and requires ACE2 for infection. *medRxiv* (2021) doi:10.1101/2021.12.08.21267417.

46. Greaney, A. J. *et al.* Comprehensive mapping of mutations in the SARS-CoV-2 receptor-binding domain that affect recognition by polyclonal human plasma antibodies. *Cell Host Microbe* **29**, 463-476.e6 (2021).

47. Riou, C. *et al.* Escape from recognition of SARS-CoV-2 variant spike epitopes but overall preservation of T cell immunity. *Sci. Transl. Med.* **14**, eabj6824 (2022).

48. Keeton, R. *et al.* T cell responses to SARS-CoV-2 spike cross-recognize Omicron. *Nature* **603**, 488–492 (2022).

49. Madhi, S. A. *et al.* Efficacy of the ChAdOx1 nCoV-19 Covid-19 Vaccine against the B.1.351 Variant. *N. Engl. J. Med.* **384**, 1885–1898 (2021).

50. Hacisuleyman, E. *et al.* Vaccine Breakthrough Infections with SARS-CoV-2 Variants. *N. Engl. J. Med.* **384**, 2212–2218 (2021).

51. Baden, L. R. *et al.* Efficacy and safety of the mRNA-1273 SARS-CoV-2 vaccine. *N. Engl. J. Med.* **384**, 403–416 (2021).

52. Zuckerkandl, E., Pauling, L., Kasha, M. & Pullman, B. Horizons in biochemistry. *Horizons in Biochemistry* 97–166 (1962).

53. Koonin, E. V. A half-century after the molecular clock: new dimensions of molecular evolution. *EMBO Rep.* **13**, 664–666 (2012).

54. Firth, C. *et al.* Using time-structured data to estimate evolutionary rates of double-stranded DNA viruses. *Mol. Biol. Evol.* **27**, 2038–2051 (2010).

55. Bogich, T. L. *et al.* Preventing pandemics via international development: a systems approach. *PLoS Med.* **9**, e1001354 (2012).

56. Daszak, P. A Call for "Smart Surveillance": A Lesson Learned from H1N1. *Ecohealth* **6**, 1–2 (2009).

57. Temmam, S. *et al.* Bat coronaviruses related to SARS-CoV-2 and infectious for human cells. *Nature* **604**, 330–336 (2022).

58. Pendleton, K. M. *et al.* Rapid Pathogen Identification in Bacterial Pneumonia Using Real-Time Metagenomics. *Am. J. Respir. Crit. Care Med.* **196**, 1610–1612 (2017).

59. Grenfell, B. T. *et al.* Unifying the epidemiological and evolutionary dynamics of pathogens. *Science* **303**, 327–332 (2004).

60. Dudas, G. *et al.* Virus genomes reveal factors that spread and sustained the Ebola epidemic. *Nature* **544**, 309–315 (2017).

61. Faria, N. R. *et al.* Zika virus in the Americas: Early epidemiological and genetic findings. *Science* **352**, 345–349 (2016).

62. Grubaugh, N. D. *et al.* Genomic epidemiology reveals multiple introductions of Zika virus into the United States. *Nature* **546**, 401–405 (2017).

63. Braden, C. R., Dowell, S. F., Jernigan, D. B. & Hughes, J. M. Progress in global surveillance and response capacity 10 years after severe acute respiratory syndrome. *Emerging Infect. Dis.* **19**, 864–869 (2013).

64. Jain, M., Olsen, H. E., Paten, B. & Akeson, M. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biol.* **17**, 239 (2016).

65. Hoenen, T. *et al.* Nanopore sequencing as a rapidly deployable Ebola outbreak tool. *Emerging Infect. Dis.* **22**, (2016).

66. Quick, J. *et al.* Real-time, portable genome sequencing for Ebola surveillance. *Nature* **530**, 228–232 (2016).

67. Faria, N. R. *et al.* Mobile real-time surveillance of Zika virus in Brazil. *Genome Med.* **8**, 97 (2016).

68. Sane, J. & Edelstein, M. Overcoming barriers to data sharing in public health. *A global perspective. Chatham House* (2015).

69. Ross, E. Perspectives on data sharing in disease surveillance.

70. Raza, S. & Luheshi, L. Big data or bust: realizing the microbial genomics revolution. *Microb. Genom.* **2**, e000046 (2016).

71. Shu, Y. & McCauley, J. GISAID: Global initiative on sharing all influenza data - from vision to reality. *Euro Surveill.* **22**, 30494 (2017).

72. Cleemput, S. *et al.* Genome Detective Coronavirus Typing Tool for rapid identification and characterization of novel coronavirus genomes. *Bioinformatics* **36**, 3552–3555 (2020).

73. Rambaut, A. *et al.* A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nat. Microbiol.* **5**, 1403–1407 (2020).

74. O'Toole, Á. *et al.* Assignment of epidemiological lineages in an emerging pandemic using the pangolin tool. *Virus Evol.* **7**, veab064 (2021).

75. Aksamentov, I., Roemer, C., Hodcroft, E. & Neher, R. Nextclade: clade assignment, mutation calling and quality control for viral genomes. *JOSS* **6**, 3773 (2021).

76. Subissi, L. *et al.* An early warning system for emerging SARS-CoV-2 variants. *Nat. Med.* **28**, 1110–1115 (2022).

77. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software

version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).

78. GitHub - neherlab/nextalign: 🧬 Viral genome reference alignment. https://github.com/neherlab/nextalign.

79. Sagulenko, P., Puller, V. & Neher, R. A. TreeTime: Maximum-likelihood phylodynamic analysis. *Virus Evol.* **4**, vex042 (2018).

80. Hadfield, J. *et al.* Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics* **34**, 4121–4123 (2018).

81. Rambaut, A., Lam, T. T., Max Carvalho, L. & Pybus, O. G. Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus Evol.* **2**, vew007 (2016).

82. Drummond, A. J., Suchard, M. A., Xie, D. & Rambaut, A. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol. Biol. Evol.* **29**, 1969–1973 (2012).

83. Suchard, M. A. *et al.* Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evol.* **4**, vey016 (2018).

84. Wickham, H. ggplot2. *WIREs Comp Stat* **3**, 180–185 (2011).

85. Yu, G. Using ggtree to Visualize Data on Tree-Like Structures. *Curr. Protoc. Bioinformatics* **69**, e96 (2020).

86. Dellicour, S., Rose, R., Faria, N. R., Lemey, P. & Pybus, O. G. SERAPHIM: studying environmental rasters and phylogenetically informed movements. *Bioinformatics* **32**, 3204–3206 (2016).

87. Huisman, J. S. *et al.* Estimation and worldwide monitoring of the effective reproductive number of SARS-CoV-2. *eLife* **11**, (2022).

Chapter 2: Sixteen novel lineages of SARS-CoV-2 in South Africa

This study, a collaboration with colleagues from the Network for Genomic Surveillance in South Africa (NGS-SA), retrospectively analyses viral genomes of SARS-CoV-2 collected in South Africa over the first wave of the pandemic as part of routine genomic surveillance in all provinces of the country. This was the first genomic epidemiology investigation of national scale to come out of South Africa. It uncovers unique transmission dynamics that unfolded during this time, including early introductions from Europe, followed by a period of local amplification of certain viral lineages. This study reveals sixteen lineages unique to South Africa during those early stages of the pandemic, three of which spread nationally and dominated most infections in that first wave. As the first author, I conceptualised and led the analysis from sequence assembly, curation of a genomic dataset, phylogenetic and phylogeography analysis, data visualisation and writing. The impact of this paper was critical, as it formed the very baseline and methodology that allowed for the quick identification of the Beta and Omicron variants of concern by the same team later on in the pandemic, efforts which then became globally recognized as early warning systems for SARS-CoV-2 evolution.

This chapter was published as a peer-reviewed research article in Nature Medicine in February 2021 and can be accessed at the following DOI: 10.1038/s41591-021-01255-3. The chapter is presented in a similar format as the journal article. A full list of authors and affiliations is shown below.

Houriiyah Tegally[1]*, Eduan Wilkinson[1]*, Richard R Lessells[1], Jennifer Giandhari[1], Sureshnee Pillay[1], Nokukhanya Msomi[2], Koleka Mlisana[3], Jinal N. Bhiman[4], Anne von Gottberg [4,15], Sibongile Walaza [4, 17] ,Vagner Fonseca[1], Mushal Allam[4], Arshad Ismail[4], Allison J. Glass[16,15] , Susan Engelbrecht[5], Gert Van Zyl[5], Wolfgang Preiser[5], Carolyn Williamson[6], Francesco Pettruccione[1], Alex Sigal[1], Inbal Gazy[1], Diana Hardie[6], Nei-yuan Hsiao[6], Darren Martin[7], Denis York[8], Dominique Goedhals[9], Emmanuel James San[1], Marta Giovanetti[10], José Lourenço[11], Luiz Carlos Junior Alcantara[10,12] and Tulio de Oliveira[#1,13,14]

*joint first authors

Affiliations: [1]KwaZulu-Natal Research Innovation and Sequencing Platform (KRISP), Nelson R Mandela School of Medicine, University of KwaZulu-Natal, Durban, South Africa [2]Discipline of Virology, University of KwaZulu-Natal, Durban, South Africa; [3]National Health Laboratory Service, South Africa; [4]National Institute For Communicable Diseases (NICD) of the National Health Laboratory Service (NHLS),,Johannesburg, South Africa; [5]Division of Medical Virology at NHLS Tygerberg Hospital, Stellenbosch University, South Africa; [6]Division of Medical Virology at NHLS Groote Schuur Hospital, University of Cape Town, Cape Town, South Africa; [7]Division of Computational

Biology, Department of Integrative Biomedical Sciences, Institute of Infectious Diseases and Molecular medicine, The University of Cape Town, Cape Town, South Africa; [8]Molecular Diagnostics Services, Durban, South Africa; [9]Division of Virology at NHLS Universitas Academic Laboratories, University of The Free State, Bloemfontein, South Africa; [10]Laboratorio de Flavivirus, Fundacao Oswaldo Cruz, Rio de Janeiro, Brazil, [11]Department of Zoology, University of Oxford, Oxford, United Kingdom, [12]Laboratorio de Genetica Celular e Molecular, ICB, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil; [13]Department of Global Health, University of Washington, Seattle, USA, [14]Centre for the AIDS Programme of Research in South Africa (CAPRISA), Durban, South Africa; [15]School of Pathology, Faculty of Health Sciences, University of the Witwatersrand [16]Department of Molecular Pathology, Lancet Laboratories, Johannesburg, South Africa; [167]School of Public Health, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa.

Abstract

The first SARS-Co-V-2 infection in South Africa was identified on 5 March, 2020 and by 26 March, the country was in full lockdown (Oxford stringency index of 90)[1]. In spite of the early response, by November 2020, over 785,000 people in South Africa were infected, which accounted for approximately 50% of all known African infections[2]. Here, we analyze 1,365 near-whole genomes and report the identification of 16 new lineages of SARS-CoV-2 isolated between 6 March and 26 August 2020. Most of these lineages have unique mutations that have not been identified elsewhere. We also show that three lineages (B.1.1.54, B.1.1.56, and C.1) spread widely in South Africa during the first wave, comprising ~42% of all infections in the country at the time. The newly identified C lineage of SARS-CoV-2, C.1, which has 16 nucleotide mutations as compared with the original Wuhan sequence, including 1 amino acid change on the spike protein, D614G[3], was the most geographically widespread lineage in South Africa by the end of August 2020. An early South-African specific lineage, B.1.106, which was identified in April 2020[4], became extinct after nosocomial outbreaks were controlled in KZN. Our findings show that genomic surveillance can be implemented on a large scale in Africa to identify new lineages and inform measures to control the spread of SARS-CoV-2. Such genomic surveillance presented in this study has revealed to be crucial in the identification of the 501Y.V2 variant in South Africa in December 2020[5].

Main text

Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) is a novel betacoronavirus, first detected in China in December 2019[6,7]. Since then, the coronavirus disease (COVID-19) has developed into a global pandemic, resulting in several waves of epidemics, infecting over 85 million people, and causing > 1.9 millions death**s** by 9 January 2021, globally. Lockdown and travel restriction measures have varied from country to country, which has dictated the profile of local outbreaks. Through the sharing of SARS-CoV-2 sequences during this pandemic, including from one of the first cases in Wuhan, China (*MN908947.3*)[7], genomic epidemiology investigations are playing a major role in characterizing and understanding this emerging virus[8–13]. SARS-CoV-2 has typically been classified into two main phylogenetic lineages, lineage A and lineage B. While

both lineages originated in China, lineage A spread from Asia to the rest of the world, whereas lineage B predominantly spread from Europe[14].

The COVID-19 epidemic in South Africa is by far the largest in Africa, with > 785,000 individuals infected and >20,000 deaths by end of November 2020. The first case of SARS-CoV-2 infection in South Africa (SA) was recorded in KwaZulu-Natal (KZN) on 5 March 2020 in a traveler returning from Italy[15]. Around mid-March, cases of community transmission were reported across the country. The profile of SARS-CoV-2 epidemiological progression in South Africa was largely influenced by the implementation of lockdown measures in the early phases of the epidemic and the subsequent easing of these measures. On 26 March 2020, the government-imposed, nationwide lockdown included the prohibition of all gatherings, travel restrictions, and closure of schools and non-essential businesses (Oxford stringency index of 90, or commonly known in South Africa as level 5 lockdown; Supplementary Table 1)[16]. Although the epidemic was growing, lockdown measures were progressively eased on 1 May, 2020 (level 4) and on 1 June, 2020 (level 3) to mitigate negative impacts on the country's economy. For example, by 1 June, interprovincial travel was allowed and there was no curfew on the movement of people. Restrictions were further relaxed on 17 August (level 2), allowing restaurants and bars to open. More restrictions were lifted on 1 October (Fig 2.1A) once the initial peak of new daily infections had passed, allowing students to return to university campuses and South Africa to return to normality. The epidemic in South Africa can generally be characterized by two phases, one dominated by travel-related early introductions, and the second being the period of peak infections (Fig 2.1A).

We monitored the likelihood of SARS-CoV-2 transmission by estimating the effective reproduction number, Re, which provides a measure of the average number of secondary infections caused by an infected person[17]. Typically, a growing epidemic is characterized by Re > 1 and Re < 1 indicates a slowed progression. At the start of the epidemic, in mid-March 2020, we estimated the Re value to be > 3, quickly falling after the start of lockdown to a value of < 1 in late-March 2020. A subsequent jump in the Re value to > 1 in April 2020 was found to be concurrent with the timing of a number of localized outbreaks in the country, including nosocomial outbreaks[18]. The Re value again dropped to < 1 at the beginning of August 2020, coinciding with a decrease in the daily number of positive cases recorded (Fig 2.1A).

Genomic epidemiology is important to understand SARS-CoV-2 evolution and track the dynamics of transmission across the world[8–13]. By 15 September 2020, at the tail end of the first epidemiological peak in the country, we had produced 1365 SARS-CoV-2 genomic sequences, 2020 (>90% coverage; publicly shared on GISAID[19]) in our laboratories as part of the Network for Genomic Surveillance (NGS-SA) consortium[20]. These genomes were sampled between 6 March and 26 August, 2020 in eight of the nine provinces of South Africa and in all the districts of KZN province, (Appendix A - Extended Data Figure 1), and represented consistent sampling

from the beginning of the epidemic and corresponding to important events of the epidemiological progression (Fig 2.1A).

We estimated maximum likelihood (ML) and molecular clock phylogenies for a dataset containing 7213 global genomes, including 1365 South African genomes, sampled from 24 December 2019 to 26 August 2020 (Fig 2.1C). Time-measured phylogeographic analyses estimated at least 101 introductions into South Africa. The bulk of imported introductions happened before lockdown (26 March, 2020) from Europe, when the epidemic was most quickly progressing (Figure 2.1B). Although at least 67 introduction events are inferred to have occurred after lockdown, these represent only 5% of the genomes that were sampled following lockdown (Fig 2.1C). In the early phases of the epidemic, before 1st of April. 34 introductions were inferred from 35 genomes sampled (97.1%), which we call early introductions (Fig 2.1B). The small number of apparent introductions after lockdown could be explained by more intensive genomic sampling at later stages, which likely revealed introduction events linked to previously undetected transmission chains.

The early introductions were mostly isolated cases with a few occurrences of small onward transmission clusters, by contrast with large transmission clusters during the peak infection phase (Fig 2.1D). The period between these two phases was inferred to be characterized by localized transmission events, which saw the emergence and spread of new lineages, which were later amplified during the first peak of the epidemic. The South African genomes in this study were assigned to 42 different lineages based on the proposed dynamic nomenclature for SARS-CoV-2 lineages[14]. This included 16 South Africa specific lineages, defined as being lineages that are presently predominant in South Africa by cov-lineages.org as of 15 September 2020[21] (Appendix A - Extended Data Figure 2). One of these has been assigned a novel SARS-CoV-2 main lineage classification, lineage C, the parent of which is lineage B.1.1.1.

Extensive SARS-CoV-2 genomic sampling, which has spanned the duration of the epidemic to date, and analyzed until the end of the first wave in this study, enabled for such lineage emergence to be observed, similar to the genomic investigation of SARS-CoV-2 in the United Kingdom[22]. During the first wave of the epidemic in South-Africa, until 15 September 2020, a total of 42 detectable SARS-CoV-2 phylogenetic lineages were circulating in the country, with an average of around 10 lineages circulating per epidemiological week, peaking to 24 in the weeks of highest infections. During the same timeframe, >1000 such transmission lineages were circulating in the UK[23]. We focused on the three largest monophyletic lineage clusters (C.1, B.1.1.54, B.1.1.56,) that spread in South Africa during lockdown and then grew into large transmission clusters during the peak infection phase of the epidemic (Fig 2.1D).

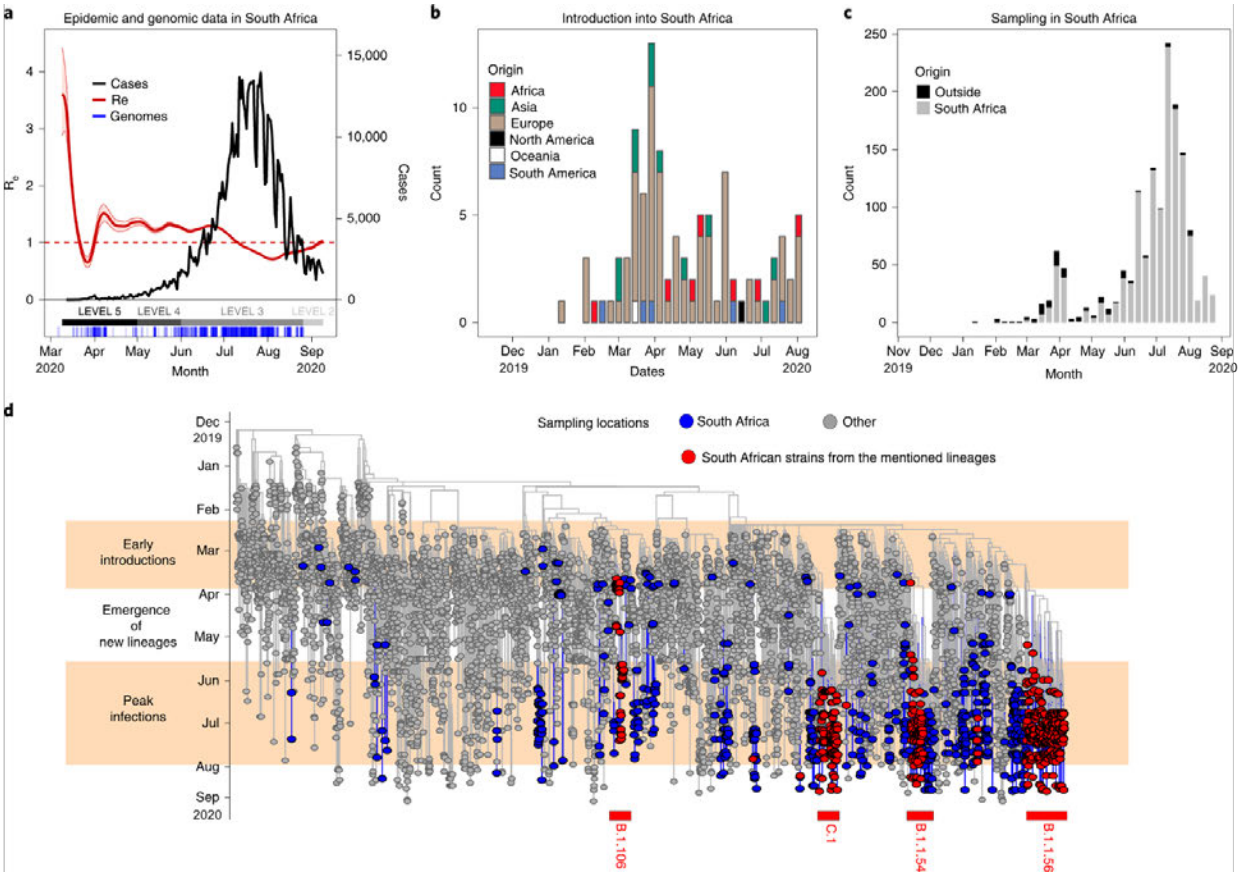


Figure 2.1: Monitoring SARS-CoV-2 Epidemic in South Africa using genomic sequencing. A) Epidemiological curve showing the progression of daily COVID-19 numbers in South Africa, changes in Re estimations (mean estimated median Re with upper and lower bounds of the 95% confidence interval shown), lockdown levels and the timing of genomic sampling in South Africa from the beginning of the epidemic to 15 September 2020. B) Estimated numbers of introductions into SA coloured by region of origin. C) Overall sampling of genomes in South Africa coloured by whether the genomes are associated with introduction events (origins outside South Africa) or not (origins in South Africa). D) Maximum clade credibility (MCC) tree of 7213 global genomes including 1365 South African sequences, indicating a period of early introductions and a period of peak infection separated by a period of emergence of new lineages. The three largest monophyletic lineage clusters in South Africa, along with the early B.1.106 South African lineage, are labelled.

B.1.1.54, B.1.1.56, and C.1 were the three largest monophyletic clusters of observed South-African lineages that emerged and spread in the country following lockdown and into the first peak of the epidemic. They contain 320, 104, and 151 genomes, respectively, which represents 42.1% of the total genomes in this study (Appendix A - Extended Data Figure 2), with a clear over-representation from mid-May to September 2020 (Fig 2.2D). Genomes belonging to these lineages were sampled in five adjacent provinces of South Africa and in all 11 districts of KZN province

(Fig 2.2B, 2.2C, Appendix A - Extended Data Figure 3), and corresponded to timepoints spanning from 31 March 2020 to 26 August 2020 (Figure 2.2B, 2.2C). We compared Ct scores, as approximations of viral loads, for genomes for which this was measured (n=653) and found no significant difference between the Ct scores of sequences belonging to these three lineages and the others (Appendix A - Extended Data Figure 4). This suggests that the fast spread of the lineages of interest is likely a result of localized outbreaks and expected transmission dynamics, rather than caused by any fitness advantage, with the caveat that samples with Ct scores measured might have been collected at different times during the course of infection, which could obscure lineage-associated differences.

In order to provide details on the spatiotemporal diffusion of South African specific lineages, we used a continuous phylogeographic model that maps the phylogenetic nodes to their inferred geographical origin locations (Fig 2.2A). Bayesian Markov chain Monte Carlo (MCMC) analysis suggests that these lineages emerged between 15 February and 24 May 2020 (Appendix A - Extended Data Figure 6). Our phylogeographic reconstruction suggests that lineage B.1.1.56 emerged in the city of Durban (eThekwini, ETH) around mid-March 2020 (95% HPD 2020-02-15 – 2020-03-30). It appears that from June onwards, this lineage quickly disseminated to all of the districts in KZN. This occurred when the country moved from lockdown level 4 to 3, which allowed for increased movement of people and goods between districts. Lineage C.1 most likely emerged in early May 2020 (95% HPD 2020-04-24 – 2020-05-24) in the city of Johannesburg, located in Gauteng province, from where it quickly spread to the adjacent North-West province, causing a large nosocomial outbreak[24]. Furthermore, the lineage spread through two independent events to the northern province of Limpopo and to northwestern KZN. From this location, the lineage further spread into all districts of KZN and to the adjacent Free State Province. Unfortunately, lineage B.1.1.54 showed poor temporal signaling (Appendix A - Extended Data Figure 5) and therefore Bayesian spatiotemporal analyses could not be performed for this cluster. A closer look at the cluster (from the ML timetree) is, however, shown in Appendix A - Extended Data Figure 6 and indicates that this lineage was first sampled in KZN and Gauteng and later spread in large numbers in the provinces of KZN, North West and the Free State.

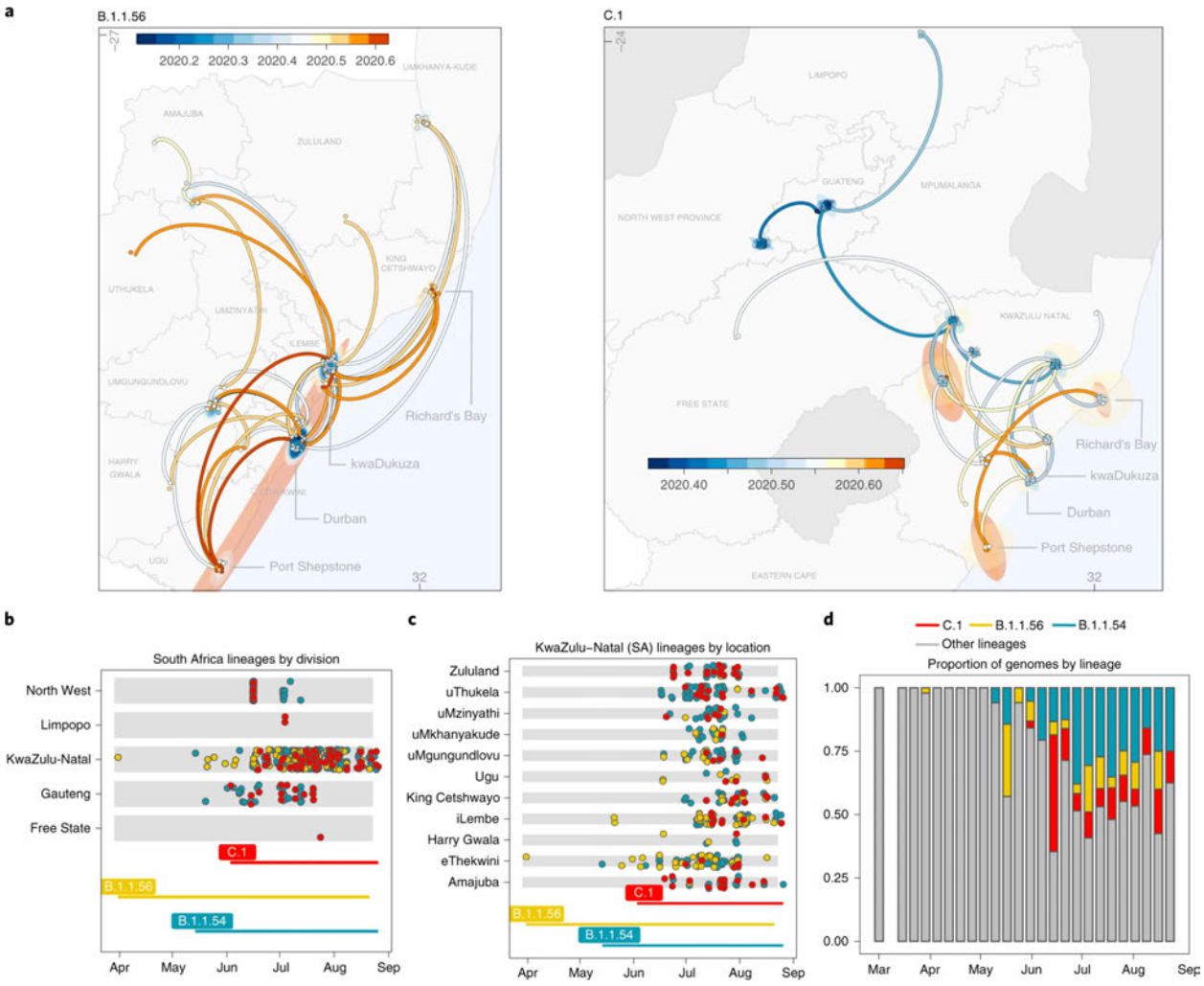


Figure 2.2: Geographical distribution and spread of lineage clusters in (5) five provinces and all districts of KZN. A) Mapping the spread of the B.1.1.56 cluster (left) and the C.1 cluster (right) from phylogeographic reconstructions. Time scale is specified in decimal dates from 2020.2 (March 2020) to 2020.6 (July 2020). B) Sampling timeline and locations of genomes belonging to each lineage cluster in (5) five provinces. C) Sampling timeline and locations of genomes belonging to each lineage cluster in all 11 districts of KZN province. D) The progression of the proportions of genomes belonging to the main lineage clusters over time.

We analyzed the sequences of the three main lineage clusters in order to determine their lineage-defining mutations, if any. On average, sequences in the C.1 cluster accumulated roughly 16 mutations, while B.1.1.56 and B.1.1.54 have approximately 13-14 mutations relative to the Wuhan reference (*MN908947.3*) (Fig 2.3A). This is relatively higher than the number of acquired mutations in other sequences as of 26 of August 2020, which is consistent with these three lineages having emerged more recently than others in the study, hence accumulating more genomic changes. Sequences are assigned lineages based on the presence of certain lineage-defining

mutations (Appendix A - Extended Data Figure 7). The sequences belonging to B.1.1.54, B.1.1.56 and C.1 all have the mutations that define their B.1.1 parental lineage (C.1 was previously known as B.1.1.1.1) (Fig 2.3B), including the 23403 A>G (Spike D614G) mutation, with additional mutations that differentiate them (Fig 2.3B). Sequences in B.1.1.54 have the *12503T>C (NSP8: Y138H)* and *29721C>T* mutations in > 90% frequency, similar to *22675C*>T for B.1.1.56, and *4002C>T (NSP3: T428I)*, *10097G>A (3C-like proteinase: G15S)*, *13536C>T*, *18747C>T* and *23731C>T* for C.1 (Fig 2.3B). The early hospital-linked lineage B.1.106 was defined by the *16376C>T (helicase: P47L)* mutation. Five of these mutations, *12503T>C*, *16376C>T, 18747C>T 29721C>T* and *22675C>T*, are predominantly present in South African SARS-CoV-2 genomes, with just a few occurrences found elsewhere globally (Fig 2.3C and Appendix A - Extended Data Figure 8), whereas other lineage-defining mutations are also common in other countries (Appendix A - Extended Data Figure 8). There are two other highly prevalent nucleotide mutations on the spike protein in the B.1.56 and C.1. lineages, 22675C>T and 23731C>T; however these are synonymous mutations and are distinct from those identified in 501Y.V2[5].

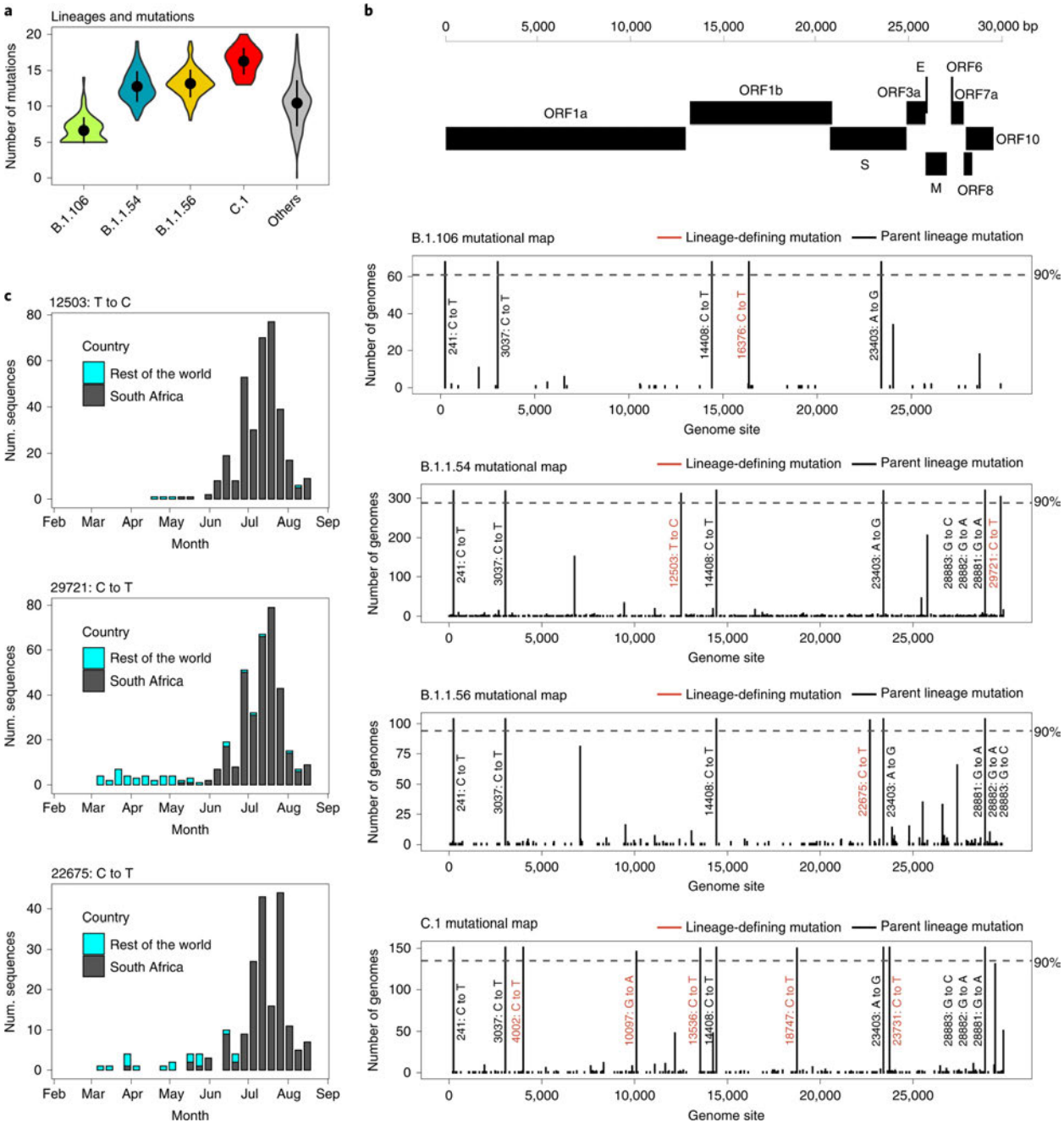


Figure 2.3: Lineage-defining mutations of the three main SA lineage clusters. A) Violin plot showing the number of mutations in each cluster. B) Variant maps of the most common mutations in each cluster mapped against the SARS-CoV-2 genomes. Most common mutations, defined as mutations present in >90% of the genomes in that group. C) Change in frequency of some unique South African mutations over time in South Africa vs the rest of the world.

Major contributors to lineage amplifications in South Africa were hospital outbreaks. For example, lineage C.1 was amplified in a nosocomial outbreak in the North West Province in April 2020[24] before spreading to KZN and other provinces. Another South African lineage, B.1.106, also

emerged in a nosocomial outbreak in KZN in April 2020. This was a large outbreak that infected 100 healthcare staff and 65 patients, and dominated most of the early infections in Durban, South Africa (Figure 2.4B). This nosocomial outbreak attracted national attention as it was responsible for 14% of the infections in KZN and over 45% of the national deaths in early April 2020. We used genetic sequencing, together with active outbreak investigation to understand how the virus entered and spread in this hospital[13]. This lineage also spread to the population and caused a second nosocomial outbreak in a nearby hospital that infected 15 health care workers (Figure 2.4A). These two nosocomial outbreaks were identified within days of the first infection and were followed with active infection and prevention control measures[4,18]. The B.1.106 lineage largely subsided following the outbreak investigations and isolation of all infected individuals. The B.1.106 lineage's prevalence at the population level decreased quickly after June 2020 (Figure 2.4B).

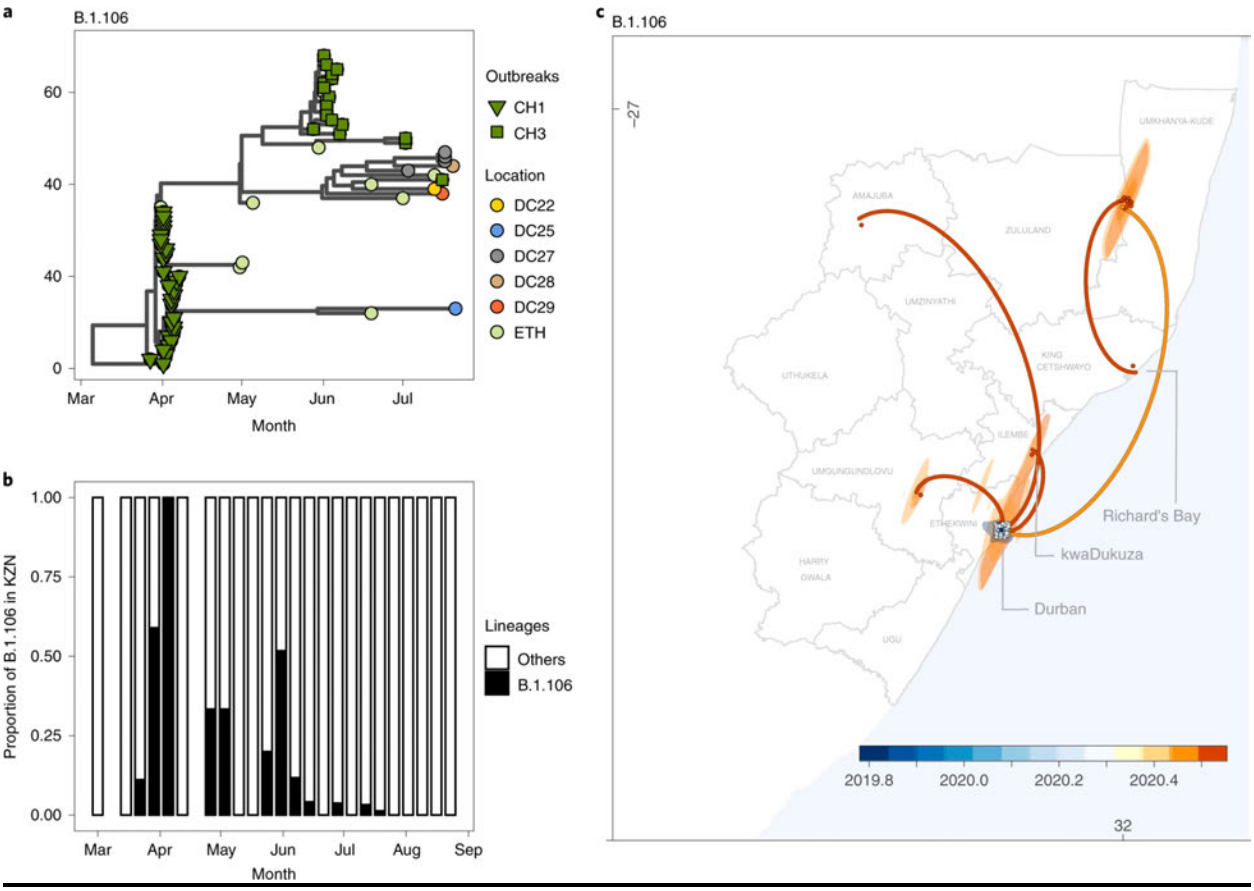


Figure 2.4: Lineage B.1.106 phylogenetic tree and dispersion throughout KwaZulu-Natal (KZN). A) Phylogenetic tree of B.1.1.106 sequences by nosocomial outbreak (Clustered Hospital 1, CH1 and CH3) and district location. B) Proportion of sequences classified as B.1.106 over time in KZN. C) Mapping the spread of the B.1.106 lineage from phylogeographic reconstructions. Time scale is specified in decimal dates from 2019.8 (November October 2020) to 2020.5 (June 2020). Shade circular patterns represent confidence intervals of location estimation.

We report an in-depth analysis of the spread of SARS-CoV-2 in South Africa from 6 March 2020 to 26 August 2020 showing that the bulk of introductions happened before lockdown and travel restrictions were implemented on March 26 2020. However, despite drastic lockdown measures, the pandemic spread quickly, causing over 785,000 laboratory confirmed infections by November 2020. In order to track the evolution of the virus in real-time, we formed the NGS-SA[25], a consortium of genomics and bioinformatics scientists who worked with national government laboratories to quickly generate and analyze data in the country. We produced 1,365 SARS-CoV-2 genomes and mapped the emergence of 16 novel lineages in South Africa. We found that three main lineages were responsible for almost half of all the infections in South Africa as of 15 September 2020. Despite a relative sequencing bias in KZN, we were able to detect these major lineages across multiple provinces. It is therefore likely that more extensive sampling throughout the country could detect the spread of these lineages nationally, especially during the period when lockdown levels were eased and mobility increased. Indeed, data from Cape Town also later identified 27 sequences available in GISAID of the C.1 lineage[19] (*EPI_ISL_660121-EPI_ISL_660150, EPI_ISL_660158*). B.1.1.54, B.1.1.56 and C.1 were the most geographically widespread lineage in South Africa during the time of this study.

Genomic data was also used in real-time to identify and control nosocomial outbreaks. The B.1.106 lineage, which was the first South African lineage to be identified, was leveraged to document how the virus spread inside a large hospital in Durban, KZN. The lessons learned in this outbreak were used to quickly control a second nosocomial outbreak in a nearby hospital. The active outbreak response, investigation and isolation of positive cases may have limited the spread of this lineage. Our analysis therefore shows that a number of SARS-CoV-2 lineages, each with unique mutations, emerged within localized epidemics during lockdown even as the introduction of new lineages from outside South Africa was being curbed.

That many of the mutations in our analysis are synonymous and that differences in Ct values do not seem to be affected by the infecting viral strain argues against selection for fitter variants, which contrast reported characteristics of variant 501Y.V2[5]. All four of the main lineages reported in the current study contain the D614G mutation in the spike gene. Furthermore, the D614G mutation is found in 1350 (95%) of the South African sequences and >99% of sequences (1241 of 1244) sampled after May 2020. Although we are currently investigating any fitness cost associated with the different lineages, we found only other three non-synonymous mutations in Spike (A688V, G769V, A1078S) with frequency ranging from 1.2 to 3.6% in this dataset, which suggests that the evolutionary stability of SARS-CoV-2 and in particular the spike protein was maintained in South Africa during the first wave of the pandemic. However, during this study, whether any of the low prevalence spike mutations reported could have a fitness advantage in terms of transmission, viral replication, or a reduced immunogenicity, was unknown. That said, we remain vigilant that there remain small recurrent gaps in our genome sequences in potentially important regions, especially in some of our lower quality sequences. We believe these gaps might have been

introduced due to potential primer mismatch in small parts of the ORF1b, S and ORF3a genes (Appendix A - Extended Data Figure 9). However, as we enter the period where re-infections and re-introduction of the viruses from international travelers is becoming more frequent, pre-existing immune responses could exert enough pressure on SARS-CoV-2 to select for resistance mutations. The dynamic nature of the COVID-19 epidemic in South Africa, and globally, supports the case for continued genomic surveillance of SARS-CoV-2. We are currently investigating limits to cross-reactivity between strains. Limited cross-reactivity could lead to effects such as antibody dependent enhancement (ADE) in response to a vaccine with a non-native strain. ADE occurs in infections such as Dengue when a previously infected individual is infected with a second strain of virus, which antibodies from the first infection can bind to but not neutralize viral proteins[20]. There is a chance that this could also happen to SARS-CoV-2 if the pandemic is not controlled over a long time providing a greater opportunity for viral evolution, which could potentially impact efficacy of current vaccines.

This study emphasizes the usefulness of integrating genomic surveillance methods to document and to help control SARS-CoV-2 spread in local and national settings. Genomics data can also be used in real-time to inform and consolidate national outbreak investigation and response strategies widely throughout Africa.

Acknowledgements
This research was funded by The South African Medical Research Council (SAMRC), MRC SHIP and the Department of Science and Innovation (DSI) of South Africa. KRISP is funded by a core award of the South African Technology Innovation Agency (TIA). We also would like to thank Andrew Rambaut and Áine O'Toole for scientific discussion on how to include the South African lineages on PANGOLIN dynamic classification. We would also like to thank all NGS-SA laboratories in South-Africa that were responsible for producing the SARS-CoV-2 genomes that were the focus of the analysis in this paper. A full list of originating laboratories and authors is included in Supplemental Table 2. Finally, we would like to thank all other global laboratories for generating and making public the SARS-CoV-2 sequences (through GISAID) used as reference dataset in this study. A complete list of individual contributors of sequences is provided in the supplemental materials.

Author Contributions

J.G., S.P., S.E., and A.I., produced SARS-CoV-2 genomic data. N.M., K.M., N.H., D.Y., D.G., A.v.G., S.W., A.J.G., I.G., A.S., G.V.Z. W.P, and D.H. collected samples and curated the metadata. R.J.L and T.d.O participated in outbreak response. H.T., E.W., R.J.L., J.G., S.P., E.J.S., S.E., F.P., A.I., J.N.B., V.F., M.G., J.L., L.C.J.A., and T.d.O analysed the data. D.H., N.H, D.M., D.G., E.J.S.,

M.G., J.L., L.C.J.A., T.d.O. helped with data interpretation. H.T., E.W., M.G., J.L., L.C.J.A., T.d.O wrote the initial manuscript. All authors participated in review of the manuscript at all stages.

Competing Interests Statement

The authors declare no competing interests.

References

1. Hale, T., Webster, S., Petherick, A., Phillips, T. & Kira, B. Oxford COVID-19 Government Response Tracker, Blavatnik School of Government.(2020). *URL www. bsg. ox. ac. uk/covidtracker.[Online](Accessed 18 Novemb. 2020)*.
2. Marivate, V. *et al.* Coronavirus disease (COVID-19) case data - South Africa. (2020) doi:10.5281/ZENODO.3819126.
3. Korber, B. *et al.* Tracking Changes in SARS-CoV-2 Spike: Evidence that D614G Increases Infectivity of the COVID-19 Virus. *Cell* **182**, 812-827.e19 (2020).
4. Lessells, R., Moosa, Y. & De Oliveira, T. *Report into a nosocomial outbreak of coronavirus disease 2019 (COVID-19) at Netcare St. Augustine's Hospital*. (2020).
5. Tegally, H. *et al.* Emergence and rapid spread of a new severe acute respiratory syndrome-related coronavirus 2 (SARS-CoV-2) lineage with multiple spike mutations in South Africa. *medRxiv* (2020).
6. Andersen, K. G., Rambaut, A., Lipkin, W. I., Holmes, E. C. & Garry, R. F. The proximal origin of SARS-CoV-2. *Nature Medicine* vol. 26 450–452 (2020).
7. Wu, F. *et al.* A new coronavirus associated with human respiratory disease in China. *Nature* **579**, 265–269 (2020).
8. Deng, X. *et al.* Genomic surveillance reveals multiple introductions of SARS-CoV-2 into Northern California. *Science (80-.).* eabb9263 (2020) doi:10.1126/science.abb9263.
9. Gonzalez-Reiche, A. S. *et al.* Introductions and early spread of SARS-CoV-2 in the New York City area. *medRxiv* 2020.04.08.20056929 (2020) doi:10.1101/2020.04.08.20056929.
10. Munnink, B. B. O. *et al.* Rapid SARS-CoV-2 whole genome sequencing for informed public health decision making in the Netherlands. *bioRxiv* 2020.04.21.050633 (2020) doi:10.1101/2020.04.21.050633.
11. Eden, J.-S. *et al.* An emergent clade of SARS-CoV-2 linked to returned travellers from Iran. *Virus Evol.* **6**, (2020).
12. Gudbjartsson, D. F. *et al.* Spread of SARS-CoV-2 in the Icelandic Population. *N. Engl. J. Med.* (2020) doi:10.1056/NEJMoa2006100.

13. Leung, K. S.-S. *et al.* A territory-wide study of early COVID-19 outbreak in Hong Kong community: A clinical, epidemiological and phylogenomic investigation. *medRxiv* 2020.03.30.20045740 (2020) doi:10.1101/2020.03.30.20045740.
14. Rambaut, A. *et al.* A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nat. Microbiol.* 1–5 (2020) doi:10.1038/s41564-020-0770-5.
15. Giandhari, J. *et al.* Early transmission of SARS-CoV-2 in South Africa: An epidemiological and phylogenetic report. *Int. J. Infect. Dis.* **103**, 234–241 (2021).
16. Disaster Management Act: Regulations to address, prevent and combat the spread of Coronavirus COVID-19: Amendment | South African Government. https://www.gov.za/documents/disaster-management-act-regulations-address-prevent-and-combat-spread-coronavirus-covid-19.
17. Huisman, J. S. *et al. A method to monitor the effective reproductive number of SARS-CoV-2*. https://ibz-shiny.ethz.ch/covid-19-re/.
18. Nordling, L. Study tells 'remarkable story' about COVID-19's deadly rampage through a South African hospital. *Science (80-.).* (2020) doi:10.1126/science.abc9593.
19. Jasper J. Koehorst, Jesse C. J. van Dam, Edoardo Saccenti, Vitor A. P. Martins dos Santos, M. S.-D. and P. J. S. GISAID Global Initiative on Sharing All Influenza Data. Phylogeny of SARS-like betacoronaviruses including novel coronavirus (nCoV). *Oxford* **34**, 1401–1403 (2017).
20. KRISP Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) vi... - SRA - NCBI. https://www.ncbi.nlm.nih.gov/sra/SRX8454220[accn].
21. SARS-CoV-2 lineages. https://cov-lineages.org/descriptions.html.
22. Meredith, L. W. *et al.* Rapid implementation of SARS-CoV-2 sequencing to investigate cases of health-care associated COVID-19: a prospective genomic surveillance study. *Lancet Infect. Dis.* (2020) doi:10.1016/S1473-3099(20)30562-4.
23. du Plessis, L. *et al.* Establishment and lineage dynamics of the SARS-CoV-2 epidemic in the UK. *Science (80-.).* eabf2946 (2021) doi:10.1126/science.abf2946.
24. Cluster outbreak at North West hospital: 106 patients and nurses infected with Covid-19. https://www.iol.co.za/news/south-africa/north-west/cluster-outbreak-at-north-west-hospital-106-patients-and-nurses-infected-with-covid-19-49616607.
25. Msomi, N. *et al.* A genomics network established to respond rapidly to public health threats in South Africa. *The Lancet Microbe* **1**, e229–e230 (2020).

Methods

Ethical statement

We obtained deidentified remnant nasopharyngeal and oropharyngeal swab samples from patients testing positive for SARS-CoV-2 by RT-qPCR from public health and private medical diagnostics laboratories (Supplemental Table 3. The project was approved by University of KwaZulu-Natal Biomedical Research Ethics Committee. Protocol reference number: BREC/00001195/2020. Project title: COVID-19 transmission and natural history in KwaZulu-Natal, South Africa: Epidemiological Investigation to Guide Prevention and Clinical Care. This project was also approved by University of the Witwatersrand Human Research Ethics Committee. Clearance certificate number: M180832. Project title: Surveillance for outpatient influenza-like illness and asymptomatic virus colonization in South Africa. Sequence data from the Western Cape was approved by the Stellenbosch University HREC Reference No: N20/04/008_COVID-19. Project Title: COVID-19: sequencing the virus from South African patients. Patient consent was not required for the genomic surveillance. This requirement was waived by the Research Ethics Committees.

Epidemiological data

We analyzed COVID-19 cases counts in South Africa from publicly released data up to 15th September 2020 from the National Department of Health (NDoH) and the National Institute for Communicable Diseases (NICD) in South Africa. This was accessible through the repository of the Data Science for Social Impact Research Group at the University of Pretoria (https://github.com/dsfsi/covid19za)[26]. The NDoH releases daily updates on the number of new confirmed cases, deaths and recoveries, with a breakdown by province. For correlation with government epidemic control measures, information from government press releases and speech transcripts was extracted. To illustrate the epidemic progression, the daily number of confirmed cases for South Africa was plotted alongside a timeline of lockdown levels and variation in estimated virus reproduction number until the 15th of September 2020.

Estimation of reproduction number

The estimations for effective daily reproduction number, Re, of SARS-CoV-2 in South Africa were obtained from the covid-19-re data repository (https://github.com/covid-19-Re/dailyRe-Data)[17] as at 15th September 2020. The effective reproductive number describes the average number of secondary infections caused by an infected individual. As described previously[17], the relevant method of calculation of Re builds upon another method developed by Cori et al.[27], accessible through EpiEstim R package. Instead of using a time series of infection incidence, which cannot be observed directly, the relevant method infers the infection incidence time series based secondary sources of information such as COVID-19 confirmed case data, hospital admissions, and deaths. This was considered in combination with two other sets of time variables:

i) the duration of SARS-CoV-2 incubation period and ii) the time delays between onset of symptoms and a positive test, a hospital admission or the death of a patient. The relevant method infers infection time series from the stated observed incidence data by deconvolution[28,29].

SARS-CoV-2 samples and metadata

Residual samples from nasopharyngeal and oropharyngeal swabs collected from COVID-19 positive patients obtained from all 11 districts for the province of KwaZulu-Natal (KZN), were used for SARS-CoV-2 WGS. We obtained samples either in the form of primary swabs or extracted RNA. The swab samples were heat inactivated in a water bath at 60°C for 30 minutes, in biosafety level 3 laboratory, prior to RNA extraction. RNA was extracted using the Viral NA/gDNA Kit on the Chemagic 360 system (Perkin Elmer, Hamburg, Germany) using the automated Chemagic 360 insturment (Perkin Elmer, Hamburg, Germany) or manually using the Qiagen Viral RNA Mini Kit (QIAGEN, California, USA). Associated metadata for the samples included date and location (district) of sampling, and sex and age of the patients.

Real Time RT-PCR

In order to detect the SARS-CoV-2 virus by PCR, the TaqPath COVID-19 CE-IVD RT-PCR Kit (Life Technologies, Carlsbad, CA) was used according to the manufacturer's instructions. The assays target genomic regions (ORF1ab, S protein and N protein) of the SARS-CoV-2 genome. RT-PCR was performed on a QuantStudio 7 Flex Real-Time PCR instrument (Life Technologies, Carlsbad, CA). Cycle thresholds (Ct) values were analyzed using auto-analysis settings with the threshold lines falling within the exponential phase of the fluorescence curves and above any background signal.

Whole genome sequencing and genome assembly

cDNA synthesis was performed on the RNA using random primers followed by gene specific multiplex PCR using the ARTIC protocol [30]. Briefly, extracted RNA was converted to cDNA using the Superscript IV First Strand synthesis system (Life Technologies, Carlsbad, CA) and random hexamer primers. SARS-CoV-2 whole genome amplification by multiplex PCR was carried out using primers designed on Primal Scheme (http://primal.zibraproject.org/) to generate 400bp amplicons with an overlap of 70bp that covers the 30Kb SARS-CoV-2 genome. PCR products were cleaned up using AmpureXP purification beads (Beckman Coulter, High Wycombe, UK) and quantified using the Qubit dsDNA High Sensitivity assay on the Qubit 4.0 instrument (Life Technologies Carlsbad, CA).

The Illumina® Nextera Flex DNA Library Prep kit was used according to the manufacturer's protocol to prepare uniquely indexed paired end libraries of genomic DNA. Sequencing libraries

were normalized to 4nM, pooled and denatured with 0.2N sodium acetate. 12pM sample library was spiked with 1% PhiX (PhiX Control v3 adapter-ligated library used as a control). Libraries were loaded onto a 500-cycle v2 MiSeq Reagent Kit and run on the Illumina MiSeq instrument (Illumina, San Diego, CA, USA).
Raw reads coming from Illumina sequencing were assembled using Genome Detective 1.126 (https://www.genomedetective.com/) and the Coronavirus Typing Tool [31,32]. The initial assembly obtained from Genome Detective was polished by aligning mapped reads to the references and filtering out low-quality mutations using bcftools 1.7-2 mpileup method. All mutations were confirmed visually with bam files using Geneious software (Biomatters Ltd, New Zealand). All of the sequences were deposited in GISAID (https://www.gisaid.org/)[19], and the GISAID accession included as part of the Supplementary Table Data S1.

Compilation of SARS-CoV-2 South Africa dataset

To present a comprehensive analysis of the genomic epidemiology of SARS-CoV-2 in South Africa, the genomes generated as of 15th September 2020 (n=1111) were combined with all other South African genomes available in GISAID as at the same date (n=298). Appropriate acknowledgement was given to the sequencing laboratories (Supplementary Data S2), and this resulted in a dataset of 1409 genomes. Sampling locations of genomes in this dataset included all provinces in South Africa, and all districts in KZN, the most sampled province (Appendix A - Extended Data Figure 1), and collection dates spanned from 6th of March 2020 (the first cases in SA) to 26th August 2020.

Quality control of genome sequences

Prior to phylogenetic reconstruction we filtered out low quality sequences from the dataset. We retrieved all South African SARS-CoV-2 genotypes from the GISAID database as of 26th of August 2020 (N=1,409). We filtered out all genotypes that met any of the following criteria: (1) Sequences with <90% genotype coverage; (2) genotypes with too many mutations (defined as having >20 nucleotide mutations relative to the Wuhan reference), which would violate the SARS-CoV-2 molecular clock at the time of study; (3) genotypes with >10 ambiguous bases; and (4) genotypes with clustered mutations defined as mutations in close proximity to one another. These are the standard quality assessment parameters employed in NextClade (https://clades.nextstrain.org). To this end we analyzed all 1,409 South African genotypes. A total of 16 South African genotypes were filtered out due to low coverage, while a further 28 were removed due to poor sequence quality. All the genomes in this dataset had a total coverage of >90%, with 70.4% of them (n=959) having a coverage of >99%, and 94.1% (n=1283) of them having a coverage of >95% relative to the reference, while 53.1% (n=726) of genomes had no missing nucleotides, giving a coverage of 100%". The final dataset of South African sequences (N=1365) were further annotated with additional metadata information (sampling locations, unique lab IDs, and outbreak numbers) (Appendix A - Extended Data Figure 10. The bulk of the

South African sequences (~81%) were sampled within the province of KZN, with sampling from all of the 11 districts within the province.

Global reference dataset

South African sequences were analyzed against a backdrop of globally representative SARS-CoV-2 genotypes. At the time of sequence analysis, more than 90,000 SARS-CoV-2 genotypes have been publicly shared. Due to the sheer size of this dataset and over sampling and in specific countries (e.g. England) we had to down sample this dataset to a manageable size. Important lineage defining genotypes along with ten randomly sampled genotypes per location were included in the phylogenetic reconstruction. The final 5,848 references contained 889 other African genotypes, 1,209 genotypes from Asia, 2,775 genotypes from Europe, 434 and 367 genotypes from North and South America respectively and 174 genotypes from Oceania.

Phylogenetic analysis of SARS-CoV-2 in South Africa

South African genotypes were analyzed against the global reference dataset using a custom build of the SARS-CoV-2 NextStrain build (https://github.com/nextstrain/ncov). The pipeline contains several python scripts that manage the analysis workflow. In short it allows for the filtering of genotypes, the alignment of genotypes in MAFFT[33], phylogenetic tree inference in IQ-Tree[34], tree dating and ancestral state construction and annotation. The resulting time scaled phylogeny can be viewed interactively and has been shared publicly on the NGS-SA NextStrain page (https://nextstrain.org/groups/ngs-sa/COVID19-Africa-2020.09.16).

The raw ML-tree topology that was produced by the NextStrain build was used to estimate the number of viral introductions through time into South Africa. TreeTime[35] was used to transform this ML-tree topology into a dated tree topology using a constant rate of 8.0×10^{-4} nucleotide substitutions per site per year, following the exclusion of outlier sequences. A migration model was fitted on the resulting time scaled tree topology in TreeTime mapping country locations to tips and internal nodes. The resulting annotated tree topology was used to infer the number of viral introductions into South Africa through time.

Lineage & Clade classification

We used the dynamic lineage classification method proposed by Rambault et al.[14] in this study via the Phylogenetic Assignment of named Global Outbreak LINeages (PANGOLIN) software suite (https://github.com/hCoV-2019/pangolin). This is aimed at identifying the most epidemiologically important lineages of SARS-CoV-2 at the time of analysis, allowing researchers to monitor the epidemic in a particular geographical region. Accordingly, with this recently proposed dynamic lineage classification many factors might suggest a new lineage including: i) monophyletic clusters

on a global tree; ii) the presence of a statistically significant support (bootstrap/ultrafast bootstrap) on the node of the new lineages; iii) introduction into a novel geographic region; iv) epidemiological support (location; travel history); v) characteristic Single Nucleotide Polymorphisms. Accordingly, with those characteristics, three main SARS-CoV-2 lineages are currently recognized; lineage A, defined by Wuhan/WH04/2020, lineage B, defined by Wuhan-Hu-1 strain, and lineage C, a sub-classification from the B lineage. We also classified the SARS-CoV-2 genomes in our dataset using the clade classification proposed by Nextstrain, divided into 19A, 19B, 20A, 20B, and 20C clades[36,37].

Dated phylogenetics

To estimate time-calibrated phylogenies dated from time-stamped genome data, we conducted phylogenetic analysis using the Bayesian software package BEASTv.1.10.4[38], on three smaller subsets of data for each of the three lineages identified in the ML phylogeny and containing isolates from South Africa (Cluster B.1.1.54, n = 320; Cluster B.1.1.56, n = 104; Cluster C.1, n = 151; Cluster B.1.106, n = 68).

ML trees from these three data subsets were inspected in TempEst v1.5.3 for the presence of a temporal (i.e. molecular clock) signal[39]. Linear regression of root-to-tip genetic distances against sampling dates indicated that the SARS-CoV-2 sequences evolve in a relatively-strong clock-like manner (r = 9.45e-2; r = 0·34; r=0.74, r = 0·50 from subset B.1.1.54; B.1.1.56 and C.1, respectively) (Appendix A - Extended Data, Figure 5).

For this analysis we employed the strict molecular clock model, the HKY+I, nucleotide substitution model and the exponential growth coalescent model[40]. We computed MCMC (Markov chain Monte Carlo) triplicate runs of 100 million states each, sampling every 10.000 steps for each data set. Convergence of MCMC chains was checked using Tracer v.1.7.1[41]. Maximum clade credibility trees were summarised from the MCMC samples using TreeAnnotator after discarding 10% as burn-in.

Phylogeographic analysis

To model phylogenetic diffusion of South African lineages across the country, we used a flexible relaxed random walk (RRW) diffusion model that accommodates branch-specific variation in rates of dispersal with a Cauchy distribution[42]. For each sequence, latitude and longitude were attributed to a point randomly sampled within the patient's province or district of residence. We discretised sequence sampling locations by considering 5 of 9 provinces in South Africa, and all 11 districts in KZN, the most sampled province, where sequences belonging to the three clusters were sampled (as shown in Appendix A - Extended Data Figure 3).

MCMC chains were run for >100 million generations and sampled every 10000th step, with convergence assessed using Tracer v1.7[43]. Maximum clade credibility trees were summarized using TreeAnnotator after discarding 10% as burn-in. We used the R package "seraphim"[44,45] to extract and map spatiotemporal information embedded in posterior trees.

Data Availability

All the SARS-CoV-2 genomes generated and presented in this study are publicly accessible through the GISAID platform (https://www.gisaid.org/). The GISAID Accession IDs of the South Africa sequences and reference genomes analyzed in this study are provided as part of Supplementary Table 3, which also contains the metadata for the sequences. Other raw data for this study are provided as the supplementary dataset https://github.com/krisp-kwazulu-natal/SARSCoV2_South_Africa_major_lineages.git. The reference SARS-CoV-2 genome (*MN908947.3*) was downloaded from the NCBI database (https://www.ncbi.nlm.nih.gov/).

Code Availability

R code and bash scripts to reproduce the analyses and figures presented in this paper are available at https://github.com/krisp-kwazulu-natal/SARSCoV2_South_Africa_major_lineages.git.

Methods-only References

26. Marivate, V. & Combrink, H. M. Use of Available Data To Inform The COVID-19 Outbreak in South Africa: A Case Study. *Data Sci. J.* **19**, (2020).
27. Cori, A., Ferguson, N. M., Fraser, C. & Cauchemez, S. A new framework and software to estimate time-varying reproduction numbers during epidemics. *Am. J. Epidemiol.* **178**, 1505–1512 (2013).
28. Gostic, K. M. *et al.* Practical considerations for measuring the effective reproductive number, Rt. *medRxiv Prepr. Serv. Heal. Sci.* (2020) doi:10.1101/2020.06.18.20134858.
29. Goldstein, E. *et al.* Reconstructing influenza incidence by deconvolution of daily mortality time series. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 21825–21829 (2009).
30. Quick, J. *ARTIC Coronavirus Method Development Community 1 more workspace*. https://protocols.io/view/ncov-2019-sequencing-protocol-v3-locost-bh42j8ye (2020).
31. Vilsker, M. *et al.* Genome Detective: an automated system for virus identification from high-throughput sequencing data. doi:10.1093/bioinformatics/bty695.
32. Cleemput, S. *et al.* Genome Detective Coronavirus Typing Tool for rapid identification and characterization of novel coronavirus genomes. doi:10.1093/bioinformatics/btaa145.

33. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
34. Nguyen, L.-T., Schmidt, H. A., Von Haeseler, A. & Minh, B. Q. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. doi:10.1093/molbev/msu300.
35. Sagulenko, P., Puller, V. & Neher, R. A. TreeTime: Maximum-likelihood phylodynamic analysis. *Virus Evol.* **4**, vex042–vex042 (2018).
36. James Hadfield, Colin Megill, Sidney M Bell, John Huddleston, Barney Potter, Charlton Callender, Pavel Sagulenko, Trevor Bedford, R. A. N. Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics* **Volume 34**, 4121–4123 (2018).
37. Year-letter Genetic Clade Naming for SARS-CoV-2 on Nextstain.org. https://nextstrain.org/blog/2020-06-02-SARSCoV2-clade-naming.
38. Suchard, M. A. *et al.* Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evol.* **4**, vey016 (2018).
39. Rambaut, A., Lam, T. T., Max Carvalho, L. & Pybus, O. G. Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus Evol.* **2**, vew007 (2016).
40. Griffiths, R. C. & Tavaré, S. Sampling theory for neutral alleles in a varying environment. *Philos. Trans. R. Soc. London. Ser. B, Biol. Sci.* **344**, 403–410 (1994).
41. Rambaut, A., Drummond, A. J., Xie, D., Baele, G. & Suchard, M. A. Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7. *Syst. Biol.* **67**, 901–904 (2018).
42. Lemey, P., Rambaut, A., Welch, J. J. & Suchard, M. A. Phylogeography takes a relaxed random walk in continuous space and time. *Mol. Biol. Evol.* **27**, 1877–1885 (2010).
43. Ferreira, M. A. R. & Suchard, M. A. Bayesian analysis of elapsed times in continuous-time Markov chains. *Can. J. Stat.* **36**, 355–368 (2008).
44. Dellicour, S. *et al.* Phylodynamic assessment of intervention strategies for the West African Ebola virus outbreak. *Nat. Commun.* **9**, 2222 (2018).
45. Dellicour, S., Rose, R., Faria, N. R., Lemey, P. & Pybus, O. G. SERAPHIM: studying environmental rasters and phylogenetically informed movements. *Bioinformatics* **32**, 3204–3206 (2016).

Chapter 3: Detection of a SARS-CoV-2 variant of concern in South Africa

This study was the first to show evolution of the SARS-CoV-2 virus into a variant of concern. The paper describes the discovery, first report and rapid characterization of the Beta variant of concern of the SARS-CoV-2 virus. This paper showcases rapid response from the Network for Genomic Surveillance in South Africa to rapidly rising cases during South Africa's second wave, at a time where most countries were seeing the end of a first wave and believing the pandemic to be on the decline. This work was one of the first indications of the ability of this virus to adapt to human circulation and that calling the pandemic over would be much more difficult than so far imagined. As first author on this study, I was at the forefront of the bioinformatics and data analysis of the sequencing data generated for this investigation. My contribution to this study ranged from the generation of genomic data, integrative data analysis of genomic and epidemiological data, phylogenetic analysis, thorough inspection of mutations, data visualisation to writing. The work culminating into this manuscript had a major impact on South Africa's pandemic response. The country went back into lockdown, experienced one of the highest rates of recorded and excess deaths, and shifted its vaccination plans.

This chapter was published as a peer-reviewed research article in Nature in March 2021 and can be accessed at the following DOI: 10.1038/s41586-021-03402-9. The chapter is presented in a similar format as the journal article. A full list of authors and affiliations is shown below.

Houriiyah Tegally[1]*, Eduan Wilkinson[1]*, Marta Giovanetti[2,3]*, Arash Iranzadeh[4]*, Vagner Fonseca[1,3], Jennifer Giandhari[1], Deelan Doolabh[5], Sureshnee Pillay[1], Emmanuel James San[1], Nokukhanya Msomi[6], Koleka Mlisana[7,8], Anne von Gottberg[9,10], Sibongile Walaza[9,11], Mushal Allam[9], Arshad Ismail[9], Thabo Mohale[9], Allison J Glass[10,12], Susan Engelbrecht[13], Gert Van Zyl[13], Wolfgang Preiser[13], Francesco Petruccione[14,15], Alex Sigal[16,17,18], Diana Hardie[19], Gert Marais[19], Marvin Hsiao[19], Stephen Korsman[19], Mary-Ann Davies[20,21], Lynn Tyers[5], Innocent Mudau[5], Denis York[22], Caroline Maslo[23], Dominique Goedhals[24], Shareef Abrahams[25], Oluwakemi Laguda-Akingba[25,26], Arghavan Alisoltani-Dehkordi[27,28], Adam Godzik[28], Constantinos Kurt Wibmer[9], Bryan Trevor Sewell[29], José Lourenço[30], Luiz Carlos Junior Alcantara[2,3], Sergei L Kosakovsky Pond[31], Steven Weaver[31], Darren Martin[4,5], Richard J Lessells[1,8], Jinal N Bhiman[9,10]*, Carolyn Williamson[5,8,19]*, Tulio de Oliveira[1,8,32]*

[1] KwaZulu-Natal Research Innovation and Sequencing Platform (KRISP), Department of Laboratory Medicine & Medical Sciences, University of KwaZulu-Natal, Durban, South Africa

[2] Laboratorio de Flavivirus, Fundacao Oswaldo Cruz, Rio de Janeiro, Brazil

[3] Laboratório de Genética Celular e Molecular, Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, Brazil

[4] Computational Biology Division, Department of Integrative Biomedical Sciences, University of Cape Town, Cape Town, 7925, South Africa

[5] Division of Medical Virology, Institute of Infectious Disease and Molecular Medicine, University of Cape Town, Cape Town, South Africa

[6] Discipline of Virology, University of KwaZulu-Natal, School of Laboratory Medicine and Medical Sciences and National Health Laboratory Service, Durban, South Africa

[7] National Health Laboratory Service, Johannesburg, South Africa

[8] Centre for the AIDS Programme of Research in South Africa (CAPRISA), Durban, South Africa

[9] National Institute for Communicable Diseases of the National Health Laboratory Service, Johannesburg, South Africa

[10] School of Pathology, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa

[11] School of Public Health, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa

[12] Department of Molecular Pathology, Lancet Laboratories, Johannesburg, South Africa

[13] Division of Medical Virology at NHLS Tygerberg Hospital and Faculty of Medicine and Health Sciences, Stellenbosch University, Cape Town, South Africa

[14] Centre for Quantum Technology, University of KwaZulu-Natal, Durban, South Africa

[15] National Institute for Theoretical Physics (NITheP), KwaZulu-Natal, South Africa

[16] Africa Health Research Institute, Durban, South Africa

[17] School of Laboratory Medicine and Medical Sciences, University of KwaZulu-Natal, Durban, South Africa

[18] Max Planck Institute for Infection Biology, Berlin, Germany

[19] Division of Medical Virology at NHLS Groote Schuur Hospital, University of Cape Town, Cape Town, South Africa

[20] Centre for Infectious Disease Epidemiology and Research, University of Cape Town, Cape Town, South Africa

[21] Western Cape Government: Health, Cape Town, South Africa

[22] Molecular Diagnostics Services, Durban, South Africa

[23] Department of Quality Leadership, Netcare Hospitals, Johannesburg, South Africa

[24] Division of Virology at NHLS Universitas Academic Laboratories, University of The Free State, Bloemfontein, South Africa

[25] National Health Laboratory Service, Port Elizabeth, South Africa

[26] Department of Laboratory Medicine and Pathology, Faculty of Health Sciences, Walter Sisulu University, Mthatha, South Africa

[27] Division of Medical Virology, Department of Pathology, University of Cape Town, Cape Town, South Africa

[28] Division of Biomedical Sciences, University of California Riverside School of Medicine, Riverside, California, USA

[29] Structural Biology Research Unit, Department of Integrative Biomedical Sciences, University of Cape Town, Rondebosch, South Africa

[30] Department of Zoology, University of Oxford, Oxford, United Kingdom

[31] Institute for Genomics and Evolutionary Medicine, Temple University, Philadelphia, Pennsylvania, USA

[32] Department of Global Health, University of Washington, Seattle, USA

* These authors contributed equally

Abstract

Continued uncontrolled transmission of the severe acute respiratory syndrome-related coronavirus 2 (SARS-CoV-2) in many parts of the world is creating the conditions for significant virus evolution. Here, we describe a new SARS-CoV-2 lineage (501Y.V2) characterised by eight lineage-defining mutations in the spike protein, including three at important residues in the receptor-binding domain (K417N, E484K and N501Y) that may have functional significance. This lineage emerged in South Africa after the first epidemic wave in a severely affected metropolitan area, Nelson Mandela Bay, located on the coast of the Eastern Cape Province. This lineage spread rapidly, becoming dominant in the Eastern Cape, Western Cape and KwaZulu-Natal Provinces within weeks. Whilst the full significance of the mutations is yet to be determined, the genomic data, showing the rapid expansion and displacement of other lineages in multiple regions, suggest that this lineage is associated with a selection advantage, most plausibly as a result of increased transmissibility or immune escape.

Introduction

Severe acute respiratory syndrome-related coronavirus 2 (SARS-CoV-2) emerged in 2019 and spread rapidly around the world, causing over 80 million recorded cases of coronavirus disease (COVID-19) and over 1.7 million deaths by the end of 2020. The failure of public health measures to contain the spread of the virus in many countries has given rise to a large number of virus lineages. Open genomic surveillance data sharing and collaborative online platforms have enabled real-time tracking of the emergence and spread of these lineages[1,2].

To date there has been relatively limited evidence of SARS-CoV-2 mutations that have had a significant functional effect on the virus. One mutation in the spike protein (D614G) emerged early in the epidemic and spread rapidly through Europe and North America in particular. Several lines of evidence now suggest that SARS-CoV-2 variants carrying this mutation have increased transmissibility[3-6]. Later in the epidemic, lineages with a N439K mutation in the spike receptor-binding domain (RBD) emerged independently in different European countries and in the United States. This mutation is associated with escape from monoclonal antibody (mAb) and polyclonal serum mediated neutralization[7].

South Africa has been the most severely affected country in Africa, with over 80 000 excess natural deaths having occurred by the end of 2020 (approximately 1400 per million population)[8]. We have previously described the introduction and spread of several SARS-CoV-2 lineages, and the emergence of unique South African lineages during the early phase of the epidemic[9,10]. Here, we now describe the emergence and spread of a new SARS-CoV-2 lineage harbouring multiple nonsynonymous spike mutations, including mutations at key sites in the RBD (K417N, E484K and N501Y) that may have functional significance. We demonstrate that this lineage emerged after

the first epidemic wave in the worst affected metropolitan area within the Eastern Cape (EC) Province. This was followed by rapid spread of this lineage to the extent that by the end of 2020 it had become the dominant lineage in three provinces.

Results

Epidemic dynamics in South Africa

The second SARS-CoV-2 epidemic wave in South Africa began around October 2020, just weeks after a trough in daily recorded cases following the first peak (Fig. 3.1A)[32]. The country's estimated effective reproduction number, Re, increased to above 1 at the end of October, indicating a growing epidemic, coinciding with a steady rise in daily cases. At the peak of the national epidemic in mid-July there were over 13 000 confirmed cases per day and almost 7000 excess deaths per week. The epidemiological profile in the three provinces that are the focus of this analysis (EC, WC and KZN) were broadly similar, although WC had an earlier and flatter peak in the first wave (Fig. 3.1B-D). At the end of the first wave in early September, there had been over 10 000 excess deaths in EC (1510 per million population), the highest for any province (Appendix B - Suppl Fig. S3). Although there was a plateau in cases following the first wave, this was noticeably short in EC and by early October there was a second phase of exponential growth, associated with an increase in deaths at a similar rate to the first wave (Fig. 3.1B). PCR test positivity rate data at a local municipality level shows very high levels (>20%) in Nelson Mandela Bay from mid-October and then rapidly rising levels in surrounding areas through October and November (Appendix B - Suppl Fig. S4). The resurgence of the daily case counts at an exponential rate happened later for WC and KZN than for EC (Fig. 3.1C-D). By early December, all three provinces were experiencing a second wave, and new cases in WC had already surpassed the peak of the first wave.

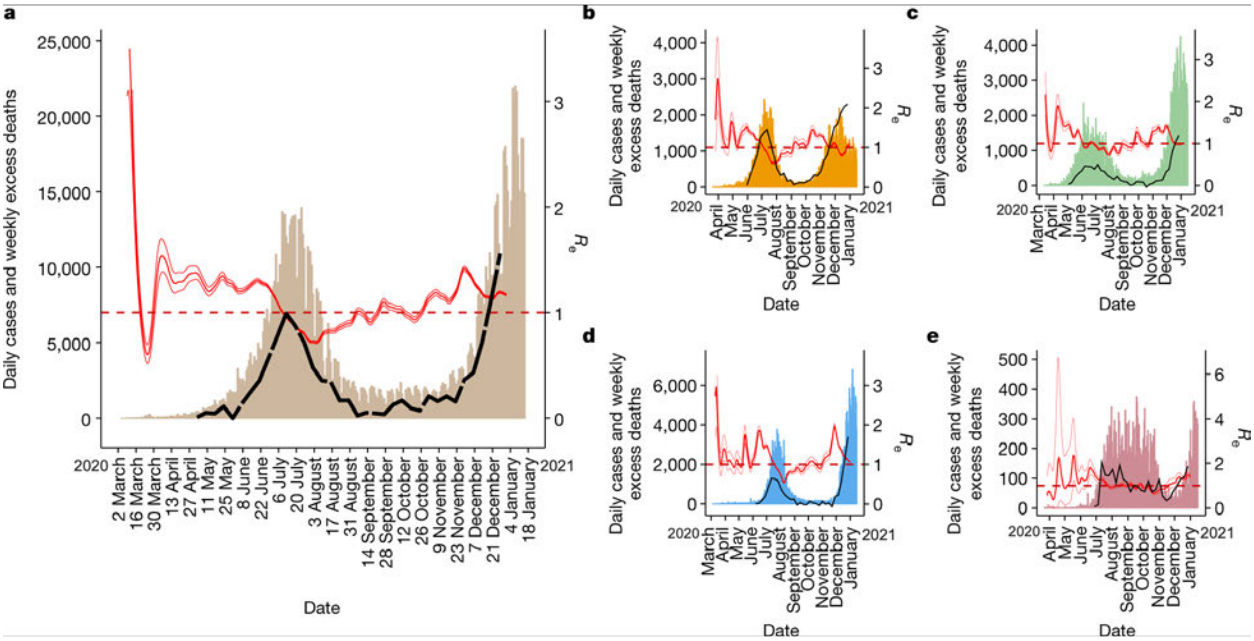


Figure 3.1. SARS-CoV-2 epidemiological dynamics in South-Africa (A), and the four provinces under study, Eastern Cape (B), Western Cape (C), KwaZulu-Natal (D) and Northern Cape (E). The histograms show the number of daily confirmed COVID-19 cases in each region (mapped to left y-axis). Fluctuations to daily Re estimates are shown in red (mapped to right y-axis) (mean estimated median Re with upper and lower bounds of the 95% confidence interval shown), with a cut-off for R=1 shown as the red broken line. Weekly excess deaths in each region are shown with the black broken lines (mapped to the left y-axis).

Phylogenetic and phylogeographic analysis

The early and rapid resurgence of the epidemic in parts of the EC and WC prompted intensification of genomic surveillance by the NGS-SA, including sampling in and around Nelson Mandela Bay in EC, and in the neighbouring Garden Route district of WC (Appendix B - Suppl Fig. S5.). We analysed 2882 SARS-CoV-2 whole genomes from South Africa collected between 5 March and 10 December 2020. We estimated preliminary maximum likelihood (ML) and molecular clock phylogenies for a dataset containing as many global reference genomes (Fig. 3.2A). We identified a new monophyletic cluster (501Y.V2) containing 341 sequences from samples collected between 8 October and 10 December in KZN, EC, WC and NC (Fig. 3.2B). Seven South African sequences are basal to the 501Y.V2 cluster (Fig 3.2A) were sampled in the provinces of the EC, WC, Gauteng and KZN between late June to early September. While these do not have any of the defining mutations of the 501Y.V2 variant, they are basal to the B.1.351 lineage and indicates that the precursor to the new variant had probably been circulating throughout the country before the emergence of the 501Y.V2.

The 501Y.V2 cluster was phylogenetically distinct from the three main lineages (B.1.1.54, B.1.1.56, and C.1) circulating widely in South Africa (>42% of samples sequenced before October 2020) during the first wave of infections (Fig 3.2A)[10]. These three lineages had been circulating in the provinces of KZN, WC, Gauteng, Free State, Limpopo, and North-West). By mid-November, the 501Y.V2 lineage had superseded B.1.1.54, B.1.1.56 and C.1, and rapidly became the dominant lineage in samples from EC, KZN and WC (Fig. 3.2C, Appendix B - Suppl Fig. S6,S7).

Spatiotemporal phylogeographic analysis suggests that the 501Y.V2 lineage emerged in early August (mid July – end August 2020, 95% highest posterior density) in Nelson Mandela Bay. Initial spread to the Garden Route District of WC was then followed by more diffuse spread from both of those areas to other regions of EC, and more recently to the City of Cape Town and several locations in KZN (Fig. 3.2D) From the City of Cape Town the variant has travelled north along the west coast of the country to the Namakwa District in the Northern Cape (NC) province.

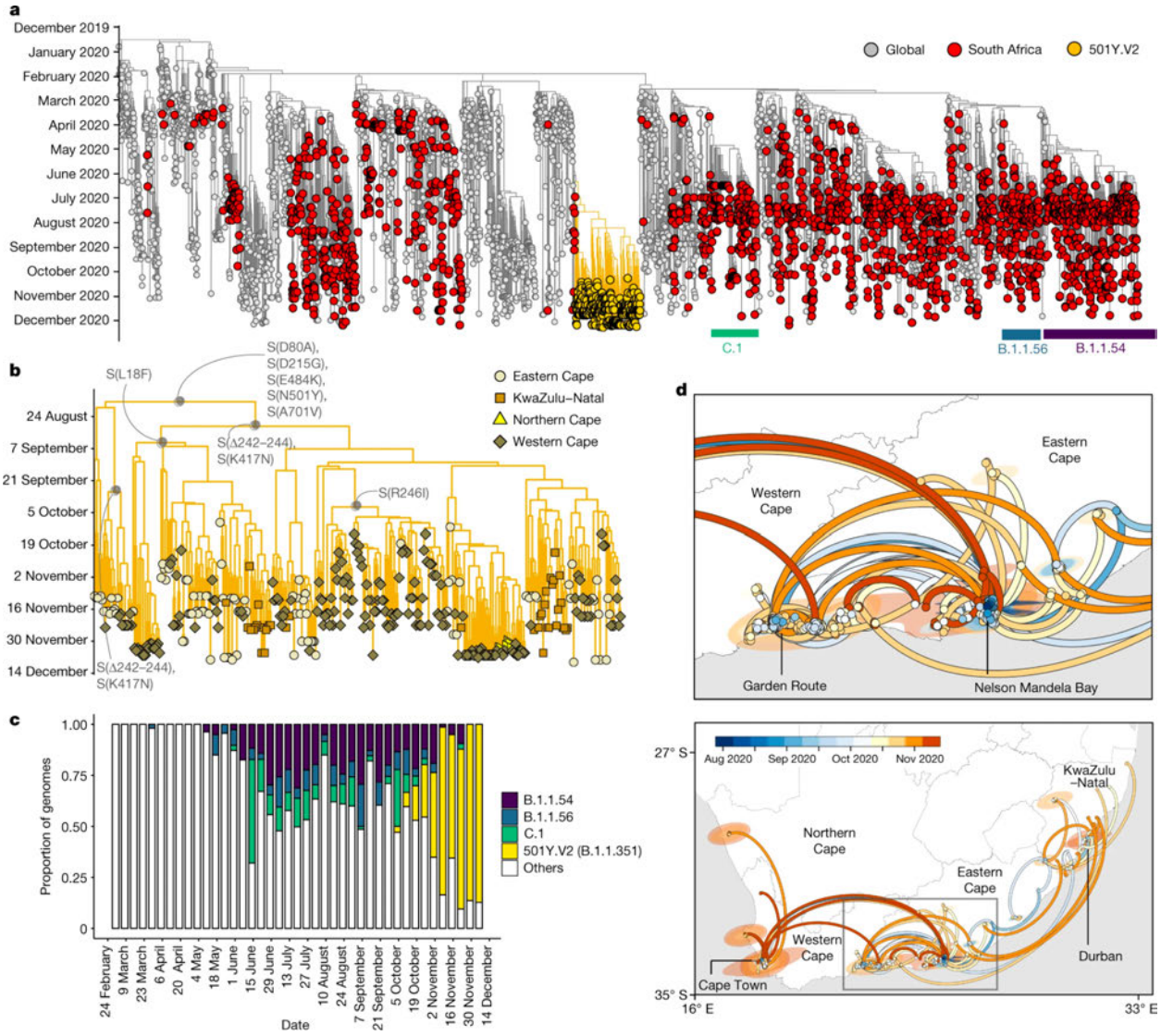


Figure 3.2. Evolution and spread of the 501Y.V2 cluster in South Africa. A) Time-resolved maximum clade credibility phylogeny of 5239 SARS-CoV-2 sequences, 2756 of which are from South Africa (red). The new SARS-CoV-2 cluster is highlighted in yellow. B) Time-resolved maximum clade credibility phylogeny of 501Y.V2 cluster, with province location indicated. Mutations characterizing the cluster are highlighted at each branch where they first emerged. C) Frequency and distribution of SARS-CoV-2 lineages circulating in South Africa over time. D) Spatiotemporal reconstruction of the spread of the 501Y.V2 cluster in South Africa during the second epidemic wave. Circles represent nodes of the maximum clade credibility phylogeny and are colored according to their inferred time of occurrence. Shaded areas represent the 80% highest posterior density interval and depict the uncertainty of the phylogeographic estimates for each node. Solid curved lines denote the links between nodes and the directionality of movement. The date scale goes from August 2020 (2020.6) to November 2020 (2020.9)

Mutational profile

At the point of first sampling on 15 October this lineage had, in addition to D614G, five other non-synonymous mutations in the spike protein, namely D80A, D215G, E484K, N501Y and A701V (Fig. 3.2B, Fig. 3.3A, Appendix B - Suppl Fig. S8). Three further spike mutations emerged by the end of November: L18F, R246I and K417N. We also observe a deletion of three amino acids at 242-244, seen in samples extracted and generated in various laboratories across the network (Because of a hard-to-align repeat region, the deletion could potential also be in amino acids 241-243 but the resulting sequence of both deletions are exactly the same). While the variants appeared in a varying proportion of the sampled genomes and showed changing frequency levels with time, the RBD mutations seem to become fixed in our sampling set, present in almost all the samples, and consistently high in frequency across time (Fig. 3.3A-B). Compared to the three largest lineages circulating in SA previously, 501Y.V2 shows marked hypermutation both in the whole genomes and the spike regions, including nonsynonymous mutations leading to amino acid changes (Fig. 3.3C). The main lineages from the first wave (B.1.1.54, B.1.1.56 and C.1) only contained the single non-synonymous spike mutation (D614G) despite following the expected temporal accumulation of mutations and therefore did not show any concerning mutation pattern like the 501Y.V2. An estimate of the evolutionary rates indicates that substitutions on the 501Y.V2 lineage are happening at 1.917E-3 nucleotide changes/site/year, compared to 5.344E-4, 4.251E-4 and 9.781E-4 respectively for B.1.1.54, B.1.1.56 and C.1 (Appendix B - Suppl Fig S2). Structural modelling of the spike trimer with these mutations reveals that three of the spike mutations are at key residues in the RBD (N501Y, E484K and K417N), three are in the N-terminal domain (L18F, D80A and D215G) and one is in loop 2 (A701V) (Fig 3.3D). The 3-amino acid deletion (242-244) also lies on the NTD. Two of the RBD sites in particular (417 and 484) are key regions for binding of neutralising antibodies (Appendix B - Suppl Fig S11).

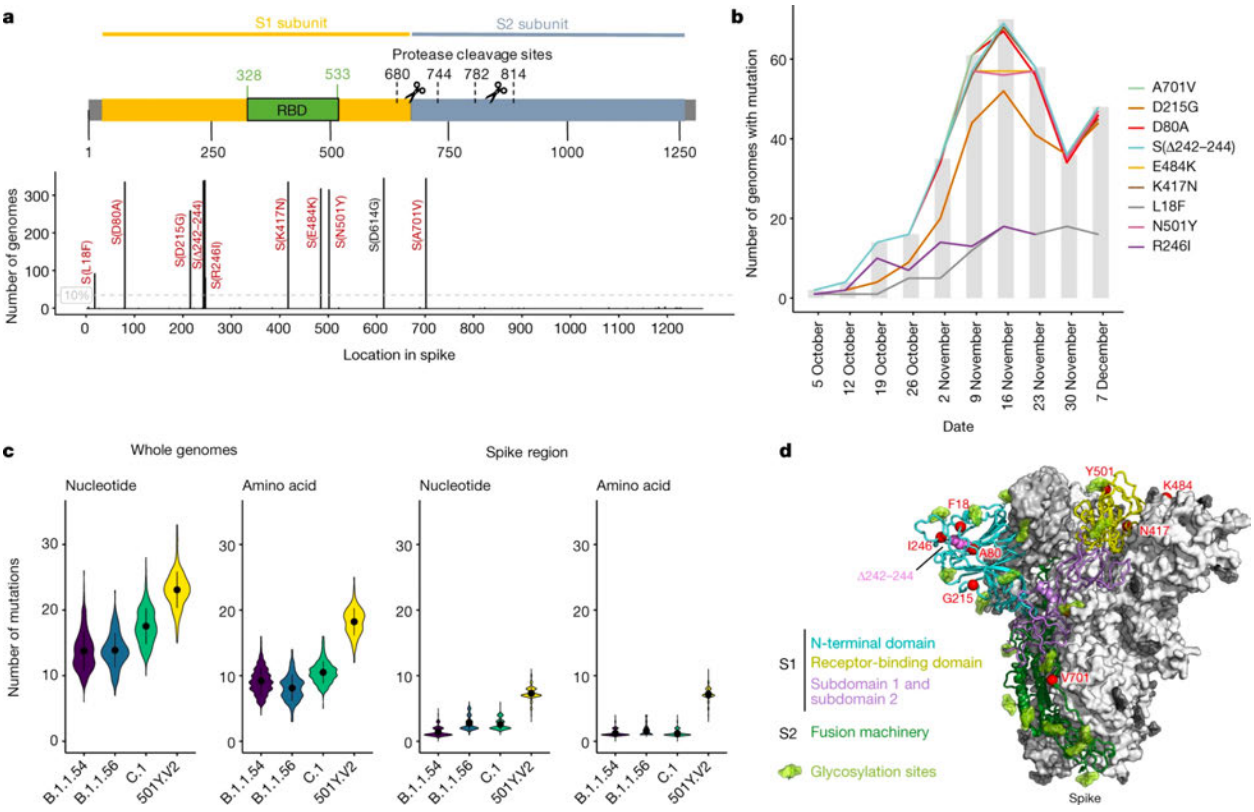


Figure 3.3. A) Amino acid changes in the spike region of the 341 501Y.V2 genomes in this study mapped to the spike protein sequences structure, indicating key regions, such as the RBD. Each spike protein variant is shown at their respective protein locations, with the bar lengths representing the number of genomes harboring the specific mutations (Only mutations that appear in >10% of sequences are shown). The D614G mutation (in black) is already present in the parent lineage. B) Changes in the mutation frequency of each variant observed during the course of sampling. Grey bars show the number of 501Y.V2 sequences sampled at a given time point and the colored lines show the change in the number of those sequences harboring each variant at the respective time points. C) Violin plots showing the numbers of nucleotide substitutions and amino acid changes that have accumulated in both the whole genomes and the spike region of the 501Y.V2 lineage, compared to lineages B.1.1.54, B.1.1.56, and C.1, three major lineages circulating in South Africa during the first wave. D) A complete model of the SARS-CoV-2 Spike (S) trimer is shown, with domains of a single protomer shown in cartoon view and coloured cyan (N-terminal domain, NTD), yellow (C-terminal domain/receptor binding domain, CTD/RBD), purple (subdomain 1 and 2, SD1 and SD2), and dark green (S2), while N-acetylglucosamine moieties are coloured light green. The adjacent protomers are shown in surface view and coloured shades of grey. Eight nonsynonymous mutants (red) and a three amino acid deletion (pink) that together define the Spike 501Y.V2 lineage are shown with spheres.

Selection analysis

We examined patterns of nucleotide variations and fluctuations in mutant frequencies at eight polymorphic spike gene sites (Fig. 3.2B) to determine whether any of the observed polymorphisms might be contributing to changes in viral fitness globally. For this analysis we used 142 037 high quality sequences from GISAID sampled between 24 December 2019 and 14 November 2020,

which represented 5964 unique spike haplotypes. The analysis indicated that two of the three sites in RBD (E484 and N501) display a pattern of nucleotide variation that is consistent with the site evolving under diversifying positive selection. The N501Y polymorphism that first appears in our sequences sampled on 15 October shows indication of positive selection on five global tree internal branches, with codon 501 of the spike gene displaying a significant excess of non-synonymous substitutions globally (dN/dS > 1 on internal branches, p=0.0011 by the FEL method), and mutant viruses encoding Y at this site have rapidly increased in frequency in both the UK and in South Africa (Z-score = 11, trend Jonckheere Terpstra non-parametric trend test). Similarly, at codon 484 there is indication of positive selection on seven global tree internal branches, with an overall significant excess of non-synonymous substitutions globally (p=0.015). Outside RBD, codons 18 (p<0.001), 80 (p=0.0014), and 215 (p<0.001) show evidence of positive diversifying selection globally with the L18F mutation also having increased in frequency in the regions where it has occurred (Z-score = 17). Up until 14 November 2020 there was no statistical evidence of positive selection at codons 417, 246 and 701.

Discussion

We describe and characterise a new SARS-CoV-2 lineage with multiple spike mutations that emerged in a major metropolitan area in South Africa following the first wave of the epidemic and then spread to multiple locations within two other neighbouring provinces. We show that this lineage has rapidly expanded and become dominant in three provinces, at the time of a rapid resurgence in infections. Whilst the full significance of the mutations is not yet clear, the genomic and epidemiological data suggest increased transmissibility associated with the virus. These data highlight the urgent need to refocus the public health response in South Africa on interrupting transmission, not only to reduce hospitalisations and deaths but to limit the national and international spread of this lineage.

This new lineage has three mutations at key sites in the RBD (K417N, E484K and N501Y). Two of these (E484K and N501Y) are within the receptor-binding motif (RBM), the main functional motif that forms the interface with the human ACE2 (hACE2) receptor. The N501Y mutation has recently been identified in a new lineage in the United Kingdom (B.1.1.7), with some preliminary evidence that it may be more transmissible[33,34]. N501 forms part of the binding loop in the contact region of hACE2, forming a hydrogen bond with Y41 in hACE2[35-37]. It also stabilises K353, one of the virus-binding hotspot residues on hACE2[38]. It is one of the key positions that differentiates SARS-CoV-2 from SARS-CoV and contributes to the enhanced binding affinity of SARS-CoV-2 for hACE2[35,38]. The N501Y mutation has been shown through deep mutation scanning and in a mouse model to enhance binding affinity to hACE2[39,40]. The E484K mutation is uncommon, being present in <0.02% of sequences from outside South Africa. E484 is also in the RBM, and interacts with the K31 interaction hotspot residue of hACE2. There is some evidence that the E484K mutation may modestly enhance binding affinity[39]. K417 is a unique hACE-2 interacting residue that forms a salt bridge interaction across the central contact region with D30 of hACE2[35,36]. This is the most striking difference in the RBD-hACE2 complex between SARS-CoV-2 and SARS-CoV, and contributes to the enhanced binding affinity of SARS-CoV-2 to hACE2[35-37]. Deep

mutational scanning suggests that the K417N mutation has minimal impact on binding affinity to hACE2[39].

The spike RBD is the main target of neutralizing antibodies (NAbs) elicited during SARS-CoV-2 infection[41]. NAbs to the RBD can be broadly divided into four main classes[42]. Of these, class 1 and class 2 antibodies appear to be most frequently elicited during SARS-CoV-2 infection, and their epitopes directly overlap the hACE2 binding site[41]. Class 1 antibodies have a VH3-53 restricted mode of recognition centred around spike residue K417. The K417N mutation would abolish key interactions with class 1 NAbs, and likely contributes toward immune evasion at this site. The N501Y mutation may also have a role in escaping class 1 NAbs, some of which make contact at this site. Class 2 antibodies bind to spike residue E484, and the E484K mutation has been shown to confer resistance to NAbs in this class, and to panels of convalescent sera, suggesting that E484 is a dominant neutralizing epitope[43-46].

One hypothesis for the emergence of this lineage, given the large number of mutations relative to the background mutation rate of SARS-CoV-2, is that it may have arisen through intra-host evolution in one or more individuals with prolonged viral replication[47,48]. This hypothesis is supported by the long branch length connecting the lineage to the remaining sequences in our phylogenetic tree (Appendix B - Suppl Fig. S10). The N501Y mutation is one of several spike mutations that emerged in an immunocompromised individual in the US who had prolonged viral replication for over 20 weeks[48]. In South Africa, the country with the world's biggest HIV epidemic, one concern has been the possibility of prolonged viral replication and intra-host evolution in the context of HIV infection, although the limited evidence so far does not suggest that HIV infection is associated with persistent SARS-CoV-2 replication[49]. It should be noted, however, that the observed diversity within this lineage cannot be explained by a single long-term infection in one individual because the lineage contains circulating intermediate mutants with subsets of the main mutations that characterise the lineage. If evolution within long-term infections were the explanation for the evolution of this lineage then one would need to invoke a transmission chain that passes through multiple individuals. Further, antigenic evolution, even within non-immuno-suppressed individuals, could offer an alternative explanation, given that several of the individual sites in spike appear to be under selective pressure worldwide, and that several of the identified mutations have been found in circulating lineages together.

Whilst we have yet to characterise how the mutations (particularly those in the RBM) affect antigenicity, it is plausible that high levels of population immunity could have driven the selection of this lineage. We have very limited SARS-CoV-2 seroprevalence data from South Africa to help understand the true extent of the epidemic. In studies using residual blood samples from routine public sector antenatal and HIV care, seroprevalence in parts of the City of Cape Town was estimated at approximately 40% in July-August, towards the end of the first epidemic wave in that area[50]. We have shown that EC, and Nelson Mandela Bay in particular, were worse affected than City of Cape Town in the first wave, and therefore we believe that population immunity could have been sufficiently high in this region to contribute to population-level selection. Whilst there have been no confirmed re-infections (supported by whole genome sequencing) in South Africa, the true extent of re-infections is unknown and this is now the focus of urgent investigation.

We are now working to understand the phenotypic impact of these SARS-CoV-2 mutations. We are focusing on the following priority studies: we are performing infectivity assays and neutralization assays to understand the effect of these mutations on ACE2 binding and NAb binding; we are analysing clinical data from the three provinces to identify any signals pointing to different disease progression or severity; we have intensified the genomic surveillance in all provinces of South Africa; and we have stepped up monitoring and surveillance for possible re-infections. Whilst the full implications of this new lineage in South Africa are yet to be determined, these findings highlight the importance of coordinated molecular surveillance systems in all parts of the world, to enable early detection and characterisation of new lineages and to inform the global pandemic response.

Acknowledgements

This research reported in this publication was supported by the Strategic Health Innovation Partnerships Unit of the South African Medical Research Council, with funds received from the South African Department of Science and Innovation.

Author Contributions

Produced SARS-CoV-2 genomic data: Jennifer Giandhari, Sureshnee Pillay, Susan Engelbrecht and Arshad Ismail

Collected samples and curated metadata: Nokukhanya Msomi, Koleka Mlisana, Nei-yuan Hsiao, Denis York, Dominique Goedhals, Anne von Gottberg, Sibongile Walaza, Allison J. Glass, Inbal Gazy, Alex Sigal, Gert Van Zyl, Wolfgang Preiser, Diana Hardie

Analysed the data: Houriiyah Tegally, Eduan Wilkinson, Marta Giovanetti, Richard J Lessells, Sergei L Kosakovsky Pond, Steven Weaver, Darren Martin, Jennifer Giandhari, Sureshnee Pillay, Emmanuel James San, Susan Engelbrecht, Francesco Pettruccione, Arshad Ismail, Jinal N. Bhiman, Vagner Fonseca, José Lourenço, Luiz Carlos Junior Alcantara, and Tulio de Oliveira

Helped with data interpretation: Diana Hardie, Nei-yuan Hsiao, Darren Martin, Dominique Goedhals, Emmanuel James San1, Marta Giovanetti, José Lourenço, Luiz Carlos Junior Alcantara, Tulio de Oliveira

Initial Manuscript Writing: Richard J Lessells, Houriiyah Tegally, Eduan Wilkinson, Marta Giovanetti, Darren Martin and Tulio de Oliveira

Review of the manuscript: All authors.

Competing Interests Statement

The authors declare no competing interests.

Methods

Epidemiological dynamics

We analysed daily cases of SARS-CoV-2 in South Africa up 16 January 2020 from publicly released data provided by the National Department of Health (NDoH) and the National Institute for Communicable Diseases. This was accessible through the repository of the Data Science for Social Impact Research Group at the University of Pretoria (https://github.com/dsfsi/covid19za)[11,12]. The NDoH releases daily updates on the number of confirmed new cases, deaths and recoveries, with a breakdown by province. We also mapped excess deaths in each province and in South Africa as a whole on to general epidemiological data to determine the extent of potential under-reporting of case numbers and gauge the severity of the epidemic. Excess deaths here are defined as the excess natural deaths (in individuals aged 1 year and above) relative to the value predicted from 2018 and 2019 data, setting any negative excesses to zero. We obtained the data from the Report on Weekly Deaths from the South Africa Medical Research Council Burden of Disease Research Unit[8]. We generated estimates for the effective reproduction number (Re) of SARS-CoV-2 in South Africa from the covid-19-re data repository (https://github.com/covid-19-Re/dailyRe-Data) as of 14 December 2020[13].

Sampling of SARS-CoV-2

As part of the Network for Genomic Surveillance in South Africa (NGS-SA)[14], five sequencing hubs receive randomly selected samples for sequencing every week according to approved protocols at each site. These samples include remnant nucleic acid extracts or remnant nasopharyngeal and oropharyngeal swab samples from routine diagnostic SARS-CoV-2 PCR testing from public and private laboratories in South Africa. In response to a rapid resurgence of COVID-19 in EC and the Garden Route District of WC in November, we enriched our routine sampling with additional samples from those areas. In total, we received samples from over 50 health facilities in the EC and WC (Appendix B - Suppl Fig. S1).

Ethical Statement

The project was approved by University of KwaZulu-Natal Biomedical Research Ethics Committee. Protocol reference number: BREC/00001510/2020. Project title: Spatial and genomic monitoring of COVID-19 cases in South Africa. This project was also approved by University of the Witwatersrand Human Research Ethics Committee. Clearance certificate number: M180832. Project title: Surveillance for outpatient influenza-like illness and asymptomatic virus colonization in South Africa. Sequence data from the Western Cape was approved by the Stellenbosch University HREC Reference No: N20/04/008_COVID-19. Project Title: COVID-19: sequencing the virus from South African patients. Patient consent was not required for the genomic surveillance. This requirement was waived by the Research Ethics Committees.

Whole genome sequencing and genome assembly

cDNA synthesis was performed on the extracted RNA using random primers followed by gene specific multiplex PCR using the ARTIC V3 protocol[15]. Briefly, extracted RNA was converted to

cDNA using the Superscript IV First Strand synthesis system (Life Technologies, Carlsbad, CA) and random hexamer primers. SARS-CoV-2 whole genome amplification was performed by multiplex PCR using primers designed on Primal Scheme (http://primal.zibraproject.org/) to generate 400bp amplicons with an overlap of 70bp that covers the 30Kb SARS-CoV-2 genome. PCR products were cleaned up using AmpureXP purification beads (Beckman Coulter, High Wycombe, UK) and quantified using the Qubit dsDNA High Sensitivity assay on the Qubit 4.0 instrument (Life Technologies Carlsbad, CA).

We then used the Illumina® Nextera Flex DNA Library Prep kit according to the manufacturer's protocol to prepare indexed paired end libraries of genomic DNA. Sequencing libraries were normalized to 4nM, pooled and denatured with 0.2N sodium acetate. 12pM sample library was spiked with 1% PhiX (PhiX Control v3 adapter-ligated library used as a control). We sequenced libraries on a 500-cycle v2 MiSeq Reagent Kit on the Illumina MiSeq instrument (Illumina, San Diego, CA, USA). We have previously published full details of the amplification and sequencing protocol[16,17].

We assembled paired-end fastq reads using Genome Detective 1.126 (https://www.genomedetective.com) and the Coronavirus Typing Tool[18]. To accurately call mutations and short indels for SARS-CoV-2, Genome Detective software was updated with an additional assembly step after the de novo assembly and strain identification. When the de novo assembly indicates a nucleotide similarity higher than 97% to the reference strain, a new assembly is made by read mapping against the reference. In this process, for strains satisfying this criterion, reads are mapped using minimap2[19] against the reference rather than the de novo consensus sequence, and subsequently final mutations and indels are called using GATK HaplotypeCaller[20], with low quality variants (with QD < 10) filtered using GATK VariantFiltration[20]. To call the consensus sequence, GATK HaplotypeCaller is used with default settings, followed by GATK VariantFiltration to select only variants with a variant confidence normalized by unfiltered depth of variant samples of at least 10 (QualByDepth >= 10). Mutations were confirmed visually with bam files using Geneious software (Biomatters Ltd, New Zealand). The reference genome used throughout the assembly process was *NC_045512.2* (numbering equivalent to *MN908947*.3). All of the sequences were deposited in GISAID (https://www.gisaid.org/), and the GISAID accession IDs are included as part of Suppl Table S2. Raw reads for our sequences have also been deposited at the National Center for Biotechnology Information Sequence Read Archive (BioProject accession PRJNA694014).

In some samples, the K417N mutation was covered by the sequencing but not called. To avoid an assembly concern, these samples were also analyzed using the ARTIC Illumina pipeline [connor-lab/ncov2019-artic-nf, git revision 9ac3119a87]. Results between the two pipelines were highly consistent with respect to the lineage defining mutations, but also consistent with respect to the missing 22813G>T (K417N) mutation in these samples despite being considered covered by both pipelines (Supplementary Table S1).

LoFreq was used to detect minor viral variants to study the intra-host heterogeneity of viral variants (quasi-species)[21] (Appendix B - Suppl Fig S12). Variants were called with at minimum coverage

of 10% and conservative false discovery rate (FDR) p-value of 0.1. LoFreq models sequencing error rate and implements a Poisson distribution to probe the statistical significance of nucleotide variants at each position filtering out all variants falling below the p-value threshold.

Quality control of South African genomic sequences

We retrieved all South African SARS-CoV-2 genomes from the GISAID database as of 4th January 2021 (N=2882). Prior to phylogenetic reconstruction, we removed low quality sequences from this dataset. We filtered out genomes that did not pass standard quality assessment parameters employed in NextClade (https://clades.nextstrain.org). We filtered out 105 South African genomes due to low coverage, and a further 18 due to poor sequence quality. Poor sequence quality was defined as sequences with clustered SNPs and ambiguous bases at >10% of sites, and low coverage genomes were anything with <90% genome coverage against the reference. We therefore analyzed a total of 2759 South African genomes. We also retrieved a global reference dataset (N=2573). This was selected from the Nextstrain global reference dataset, plus the five most similar sequences to each of the South African sequences as defined by a local BLAST search.

Phylogenetic analysis

We initially analyzed South African genomes against the global reference dataset using a custom pipeline based on a local version of NextStrain[2]. The pipeline contains several python scripts that manage the analysis workflow. It performs alignment of genomes in MAFFT[22], phylogenetic tree inference in IQ-Tree[23], tree dating and ancestral state construction and annotation (https://github.com/nextstrain/ncov). The full nextstrain build can be viewed at: https://nextstrain.org/groups/ngs-sa/COVID19-ZA-2021.01.18.

The initial phylogenetic analysis allowed us to identify a large cluster of sequences (n=341) with multiple spike mutations. We extracted this cluster and constructed a preliminary maximum likelihood (ML) tree in IQ-tree, together with eight basal sequences from the region sampled June-September 2020. We inspected this ML tree in TempEst v1.5.3 for the presence of a temporal (i.e. molecular clock) signal. Linear regression of root-to-tip genetic distances against sampling dates indicated that SARS-CoV-2 sequences evolved in a relatively strong clock-like manner (correlation coefficient=0.34, R^2=0.11) (Appendix B - Suppl Fig. S2).

We then estimated time-calibrated phylogenies using the Bayesian software package BEASTv.1.10.4. For this analysis, we employed the strict molecular clock model, the HKY+I, nucleotide substitution model and the exponential growth coalescent model[24]. We computed Markov chain Monte Carlo (MCMC) in duplicate runs of 100 million states each, sampling every 10 000 steps. Convergence of MCMC chains was checked using Tracer v.1.7.1[25]. Maximum clade credibility trees were summarised from the MCMC samples using TreeAnnotator after discarding 10% as burn-in.

Phylogeographic analysis

To model phylogenetic diffusion of the new cluster across the country, we used a flexible relaxed random walk (RRW) diffusion model that accommodates branch-specific variation in rates of

dispersal with a Cauchy distribution[26]. For each sequence, latitude and longitude were attributed to the health facility at which the diagnostic sample was obtained, or, if that information was not available, to a point randomly sampled within the local area or district of origin. Given that we don't have access to residential geolocators within the genomic surveillance, the location of the health facility serves as a reasonable proxy, especially as two-thirds of the population live within 2km of their nearest health facility[27].

As above, MCMC chains were run in duplicate for 100 million generations and sampled every 10 000 steps, with convergence assessed using Tracer v1.7.1. Maximum clade credibility trees were summarized using TreeAnnotator after discarding 10% as burn-in. We used the R package "seraphim" to extract and map spatiotemporal information embedded in posterior trees.

Lineage classification

We used the dynamic lineage classification method proposed by Rambault et al. via the Phylogenetic Assignment of named Global Outbreak LINeages (PANGOLIN) software suite (https://github.com/hCoV-2019/pangolin)[28]. This is aimed at identifying the most epidemiologically important lineages of SARS-CoV-2 at the time of analysis, allowing researchers to monitor the epidemic in a particular geographical region. A lineage is a linear chain of viruses in a phylogenetic tree showing connection from the ancestor to the last descendant. Variant refers to a genetically distinct virus with different mutations to other viruses. For the new variant identified in South Africa in this study, we have assigned it the name 501Y.V2; the corresponding PANGO lineage classification is B.1.351 (lineages version 2021-01-06).

Selection analysis

To identify which, if any, of the observed mutations in the spike protein was most likely to increase viral fitness, we used the natural selection analysis of SARS-CoV-2 pipeline (https://observablehq.com/@spond/revised-sars-cov-2-analytics-page). This pipeline examines the entire global SARS-CoV-2 nucleotide sequence dataset for evidence of: (i) polymorphisms having arisen in multiple epidemiologically unlinked lineages that have statistical support for non-neutral evolution (Mixed Effects Model of Evolution, MEME)[29], (ii) sites where these polymorphisms have support for a greater than expected ratio of non-synonymous:synonymous nucleotide substitution rates on internal branches of the phylogenetic tree (Fixed Effects Likelihood, FEL)[30], and (iii) whether these polymorphisms have increased in frequency in the regions of the world where they have occurred.

Structural modelling

We modelled the spike protein based on the Protein Data Bank coordinate set 7a94, showing the first step of the S protein trimer activation with one RBD domain in the up position, bound to the hACE2 receptor[31]. We used the Pymol program (The PyMOL Molecular Graphics System, Version 2.2.0, Schrödinger, LLC.) for visualization.

Data Availability

All the SARS-CoV-2 501Y.V2 genomes generated and presented in this study are publicly accessible through the GISAID platform (https://www.gisaid.org/), along with all other SARS-CoV-2 genomes generated by the Network for Genomic Surveillance in South-Africa (NGS-SA). The GISAID Accession IDs of the 501Y.V2 sequences analyzed in this study are provided as part of Supplementary Table S2, which also contains the metadata for the sequences. The raw reads for the 501Y.V2 have been deposited at the NCBI SRA (BioProject accession PRJNA694014). Other raw data for this study are provided as supplementary dataset on our GitHub repository: https://github.com/krisp-kwazulu-natal/SARSCoV2_South_Africa_501Y_V2_B_1_351. The reference SARS-CoV-2 genome (*MN908947.3*) was downloaded from the NCBI database (https://www.ncbi.nlm.nih.gov/).

Code Availability

R code and bash scripts to reproduce the analyses and figures presented in this paper are available at https://github.com/krisp-kwazulu-natal/SARSCoV2_South_Africa_501Y_V2_B_1_351.

References

1. Shu Y, McCauley J. GISAID: Global initiative on sharing all influenza data - from vision to reality. *Euro Surveill.* 2017;22(13).

2. Hadfield J, Megill C, Bell SM, et al. Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics.* 2018;34(23):4121-4123.

3. Korber B, Fischer WM, Gnanakaran S, et al. Tracking Changes in SARS-CoV-2 Spike: Evidence that D614G Increases Infectivity of the COVID-19 Virus. *Cell.* 2020;182(4):812-827 e819.

4. Volz E, Hill V, McCrone JT, et al. Evaluating the Effects of SARS-CoV-2 Spike Mutation D614G on Transmissibility and Pathogenicity. *Cell.* 2020.

5. Plante JA, Liu Y, Liu J, et al. Spike mutation D614G alters SARS-CoV-2 fitness. *Nature.* 2020.

6. Yurkovetskiy L, Wang X, Pascal KE, et al. Structural and Functional Analysis of the D614G SARS-CoV-2 Spike Protein Variant. *Cell.* 2020;183(3):739-751 e738.

7. Thomson EC, Rosen LE, Shepherd JG, et al. The circulating SARS-CoV-2 spike variant N439K maintains fitness while evading antibody-mediated immunity. *bioRxiv.* 2020:2020.2011.2004.355842.

8. Bradshaw D, Laubscher R, Dorrington R, Groenewald P, Moultrie T. *Report on Weekly Deaths in South Africa: 1 January - 8 December 2020 (Week 49).* Burden of Disease Research Unit, South African Medical Research Council;2020.

9. Giandhari J, Pillay S, Wilkinson E, et al. Early transmission of SARS-CoV-2 in South Africa: An epidemiological and phylogenetic report. *Int J Infect Dis.* 2020.

10. Tegally H, Wilkinson E, Lessells RR, et al. Major new lineages of SARS-CoV-2 emerge and spread in South Africa during lockdown. *medRxiv.* 2020:2020.2010.2028.20221143. Nature Medicine in press.

11. Marivate V, Combrink HM. Use of available data to inform the COVID-19 outbreak in South Africa: a case study. *Data Science Journal.* 2020;19:19.

12. Marivate V, Arbia R, Combrink H, et al. Coronavirus disease (COVID-19) case data - South Africa. In: Pretoria DrgatUo, ed2020.

13. Huisman JS, Scire J, Angst DC, Neher RA, Bonhoeffer S, Stadler T. Estimation and worldwide monitoring of the effective reproductive number of SARS-CoV-2. *medRxiv.* 2020:2020.2011.2026.20239368.

14. Msomi N, Mlisana K, de Oliveira T, Network for Genomic Surveillance in South Africa writing g. A genomics network established to respond rapidly to public health threats in South Africa. *Lancet Microbe.* 2020;1(6):e229-e230.

15. Quick J. nCoV-2019 sequencing protocol v3 (LoCost). *protocolsio.* 2020.

16. Pillay S, Giandhari J, Tegally H, et al. Whole Genome Sequencing of SARS-CoV-2: Adapting Illumina Protocols for Quick and Accurate Outbreak Investigation during a Pandemic. *Genes (Basel).* 2020;11(8).

17. Giandhari J, Pillay S, Tegally H, et al. NEBnext library construction and sequencing for SARS-CoV-2: Adapting COVID-19 ARTIC protocol. *protocolsio.* 2020.

18. Cleemput S, Dumon W, Fonseca V, et al. Genome Detective Coronavirus Typing Tool for rapid identification and characterization of novel coronavirus genomes. *Bioinformatics.* 2020;36(11):3552-3555.

19. Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics.* 2018;34(18):3094-3100.

20. McKenna A, Hanna M, Banks E, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 2010;20(9):1297-1303.

21. Wilm A, Aw PP, Bertrand D, et al. LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. *Nucleic Acids Res.* 2012;40(22):11189-11201.

22. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol.* 2013;30(4):772-780.

23. Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol.* 2015;32(1):268-274.

24. Griffiths RC, Tavare S. Sampling theory for neutral alleles in a varying environment. *Philos Trans R Soc Lond B Biol Sci.* 1994;344(1310):403-410.

25. Rambaut A, Drummond AJ, Xie D, Baele G, Suchard MA. Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7. *Syst Biol.* 2018;67(5):901-904.

26. Lemey P, Rambaut A, Welch JJ, Suchard MA. Phylogeography takes a relaxed random walk in continuous space and time. *Mol Biol Evol.* 2010;27(8):1877-1885.

27. McLaren ZM, Ardington C, Leibbrandt M. Distance decay and persistent health care disparities in South Africa. *BMC Health Serv Res.* 2014;14(1):541.

28. Rambaut A, Holmes EC, O'Toole Á, et al. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nature Microbiology.* 2020.

29. Murrell B, Wertheim JO, Moola S, Weighill T, Scheffler K, Kosakovsky Pond SL. Detecting individual sites subject to episodic diversifying selection. *PLoS Genet.* 2012;8(7):e1002764.

30. Kosakovsky Pond SL, Frost SD. Not so different after all: a comparison of methods for detecting amino acid sites under selection. *Mol Biol Evol.* 2005;22(5):1208-1222.

31. Benton DJ, Wrobel AG, Xu P, et al. Receptor binding and priming of the spike protein of SARS-CoV-2 for membrane fusion. *Nature.* 2020;588(7837):327-330.

32. Blumberg L, Frean J. COVID-19 Second Wave in South Africa. National Institute of Communicable Diseases. https://www.nicd.ac.za/covid-19-second-wave-in-south-africa/. Published 2020. Accessed 21 December, 2020.

33. Kemp S, Datir R, Collier D, et al. Recurrent emergence and transmission of a SARS-CoV-2 Spike deletion ΔH69/V70. *bioRxiv.* 2020:2020.2012.2014.422555.

34. Rambaut A, Loman N, Pybus O, et al. Preliminary genomic characterisation of an emergent SARS-CoV-2 lineage in the UK defined by a novel set of spike mutations. https://virological.org/t/preliminary-genomic-characterisation-of-an-emergent-sars-cov-2-lineage-in-the-uk-defined-by-a-novel-set-of-spike-mutations/563. Published 2020. Accessed 21 December, 2020.

35. Lan J, Ge J, Yu J, et al. Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. *Nature.* 2020;581(7807):215-220.

36. Wang Y, Liu M, Gao J. Enhanced receptor binding of SARS-CoV-2 through networks of hydrogen-bonding and hydrophobic interactions. *Proc Natl Acad Sci U S A.* 2020;117(25):13967-13974.

37. Yi C, Sun X, Ye J, et al. Key residues of the receptor binding motif in the spike protein of SARS-CoV-2 that interact with ACE2 and neutralizing antibodies. *Cell Mol Immunol.* 2020;17(6):621-630.

38. Shang J, Ye G, Shi K, et al. Structural basis of receptor recognition by SARS-CoV-2. *Nature.* 2020;581(7807):221-224.

39. Starr TN, Greaney AJ, Hilton SK, et al. Deep Mutational Scanning of SARS-CoV-2 Receptor Binding Domain Reveals Constraints on Folding and ACE2 Binding. *Cell.* 2020;182(5):1295-1310 e1220.

40. Gu H, Chen Q, Yang G, et al. Adaptation of SARS-CoV-2 in BALB/c mice for testing vaccine efficacy. *Science.* 2020;369(6511):1603-1607.

41. Piccoli L, Park YJ, Tortorici MA, et al. Mapping Neutralizing and Immunodominant Sites on the SARS-CoV-2 Spike Receptor-Binding Domain by Structure-Guided High-Resolution Serology. *Cell.* 2020;183(4):1024-1042 e1021.

42. Barnes CO, Jette CA, Abernathy ME, et al. SARS-CoV-2 neutralizing antibody structures inform therapeutic strategies. *Nature.* 2020.

43. Weisblum Y, Schmidt F, Zhang F, et al. Escape from neutralizing antibodies by SARS-CoV-2 spike protein variants. *Elife.* 2020;9:e61312.

44. Greaney AJ, Starr TN, Gilchuk P, et al. Complete Mapping of Mutations to the SARS-CoV-2 Spike Receptor-Binding Domain that Escape Antibody Recognition. *Cell Host Microbe.* 2020.

45. Liu Z, VanBlargan LA, Rothlauf PW, et al. Landscape analysis of escape variants identifies SARS-CoV-2 spike mutations that attenuate monoclonal and serum antibody neutralization. *bioRxiv.* 2020:2020.2011.2006.372037.

46. Baum A, Fulton BO, Wloga E, et al. Antibody cocktail to SARS-CoV-2 spike protein prevents rapid mutational escape seen with individual antibodies. *Science.* 2020;369(6506):1014-1018.

47. Avanzato VA, Matson MJ, Seifert SN, et al. Case Study: Prolonged Infectious SARS-CoV-2 Shedding from an Asymptomatic Immunocompromised Individual with Cancer. *Cell.* 2020.

48. Choi B, Choudhary MC, Regan J, et al. Persistence and Evolution of SARS-CoV-2 in an Immunocompromised Host. *N Engl J Med.* 2020;383(23):2291-2293.

49. Karim F, Gazy I, Cele S, et al. HIV infection alters SARS-CoV-2 responsive immune parameters but not clinical outcomes in COVID-19 disease. *medRxiv.* 2020:2020.2011.2023.20236828.

50. Hsiao M, Davies MA, Kalk E, et al. SARS-CoV-2 seroprevalence in the Cape Town Metropolitan sub-districts after the peak of infections. *NICD COVID-19 Special Public Health Surveillance Bulletin.* 2020;18:1-9.

Chapter 4: Rapid epidemic expansion of the SARS-CoV-2 Omicron variant in southern Africa

This chapter describes the discovery, first report and rapid characterization of the Omicron variant of concern of the SARS-CoV-2 virus. The paper showcases rapid response from the Network for Genomic Surveillance in South Africa to yet another public health emergency of the COVID-19 epidemic in South Africa and globally when recorded cases and test positivity rates were again rapidly rising during South Africa's fourth wave. As co-first author on this study, I was at the forefront of the bioinformatics and data analysis of the sequencing data associated with Omicron, which allowed for real-time genomic surveillance and response. My work towards this study built on previous work characterising other lineages and variants in South Africa, described in Chapters 2 and 3 and involved assembly of genomic data, integrative data analysis of genomic and epidemiological data, phylogenetic analysis to estimate the date of emergence of the variant, careful verification of mutations in the genomic data, data visualisation and manuscript writing. The work culminating into this chapter had a major impact on the world, being the variant that most rapidly progressed to being concerning. Within three days of the first genome being uploaded, it was designated a variant of concern (Omicron, B.1.1.529) by the World Health Organization and, within three weeks, had been identified in 87 countries.

This chapter was published as a peer-reviewed research article in Nature in January 2022 and can be accessed at the following DOI: 10.1038/s41586-022-04411-y. The chapter is presented in a similar format as the journal article. A full list of authors and affiliations is shown below.

Raquel Viana[1]*, Sikhulile Moyo[2,3]*, Daniel Amoako[4]*, Houriiyah Tegally[5]*, Catherine Scheepers[4,6]*, Christian L Althaus[7], Phillip A Bester[8,9], Maciek F Boni[10], Mohammed Chand[11], Wonderful Choga[2], Rachel Colquhoun[12], Michaela Davids[13], Koen Deforche[14], Deelan Doolabh[15], Josie Everatt[4], Jennifer Giandhari[5], Marta Giovanetti[16,17], Diana Hardie[15], Verity Hill[12], Nei-Yuan Hsiao[15,18,19], Arash Iranzadeh[20], Arshad Ismail[4], Charity Joseph[11], Rageema Joseph[15], Legodile Koopile[2], Sergei L Kosakovsky Pond[21], Lesego Kuate-Lere[22], Oluwakemi Laguda-Akingba[23,24], Pamela Lawrence-Smith[22], Onalethatha Lesetedi-Mafoko[25], Richard J Lessells[5], Shahin Lockman[2,26], Alexander Lucaci[21], Arisha Maharaj[5], Boitshoko Mahlangu[4], Tongai Maponga[27], Zinhle Makatini[28], Gert Marais[15], Dorcas Maruapula[2], Kereng Masupu[29], Mogomotsi

Matshaba[30,31], Simnikiwe Mayaphi[32], Nokuzola Mbhele[15], Mpaphi B Mbulawa[33], Adriano Mendes[13], Koleka Mlisana[34,35], Anele Mnguni[4], Thabo Mohale[4], Monika Moir[36], Kgomotso Moruisi[22], Mosepele Mosepele[31,37], Gerald Motsatsi[4], Modisa S Motswaledi[38], Nokukhanya Msomi[39], Peter N Mwangi[9,40], Yeshnee Naidoo[5], Noxolo Ntuli[4], Martin Nyaga[9,40], Lucier Olubayo[19,20], Sureshnee Pillay[5], Botshelo Radibe[2], Yajna Ramphal[5], Upasana Ramphal[5], James E San[5], Lesley Scott[41], Roger Shapiro[2,26], Lavanya Singh[5], Wendy Stevens[41], Amy Strydom[13], Kathleen Subramoney[28], Naume Tebeila N[4,42], Derek Tshiabuila[5], Jacob A Ugochukwu[5], Stephanie van Wyk[36], Steven Weaver[21], Constantinos K Wibmer[4], Eduan Wilkinson[36], Nicole Wolter[4,43], Boitumelo Zuze[2], Dominique Goedhals[9,44], Wolfgang Preiser[27], Florette Treurnicht[28], Marietje Venter[13], Carolyn Williamson[18,19,45], Jinal Bhiman[4,43], Allison Glass[1,43], Darren P Martin[19,45], Andrew Rambaut[12], Simani Gaseitsiwe[2,3]**, Anne von Gottberg[4,43]**, Tulio de Oliveira[5,35,46]**✉

[1]Lancet Laboratories, Johannesburg, South Africa
[2]Botswana Harvard AIDS Institute Partnership, Botswana Harvard HIV Reference Laboratory, Gaborone, Botswana
[3]Harvard T.H. Chan School of Public Health, Boston, Massachusetts
[4]National Institute for Communicable Diseases (NICD) of the National Health Laboratory Service (NHLS), Johannesburg, South Africa
[5]KwaZulu-Natal Research Innovation and Sequencing Platform (KRISP), Nelson R Mandela School of Medicine, University of KwaZulu-Natal, Durban, South Africa
[6]South African Medical Research Council Antibody Immunity Research Unit, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa
[7]Institute of Social and Preventive Medicine, University of Bern, Bern, Switzerland
[8]Division of Virology, National Health Laboratory Service, Bloemfontein, South Africa
[9]Division of Virology, University of the Free State, Bloemfontein, South Africa
[10]Center for Infectious Disease Dynamics, Department of Biology, Pennsylvania State University, University Park, PA, USA
[11]Diagnofirm Medical Laboratories, Gaborone, Botswana
[12]Institute of Evolutionary Biology, University of Edinburgh, Edinburgh, UK
[13]Zoonotic Arbo and Respiratory Virus Program, Department of Medical Virology, University of Pretoria, Pretoria, South Africa
[14]Emweb bv, Herent, Belgium
[15]Division of Medical Virology, Faculty of Health Sciences, University of Cape Town, Cape Town, South Africa
[16]Laboratorio de Flavivirus, Fundacao Oswaldo Cruz, Rio de Janeiro, Brazil
[17]Laboratório de Genética Celular e Molecular, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil
[18]NHLS Groote Schuur Laboratory, Cape Town, South Africa
[19]Wellcome Centre for Infectious Diseases Research in Africa (CIDRI-Africa)
[20]Division of Computational Biology, Faculty of Health Sciences, University of Cape Town
[21]Institute for Genomics and Evolutionary Medicine, Department of Biology, Temple University, Pennsylvania, USA
[22]Health Services Management, Ministry of Health and Wellness, Gaborone, Botswana
[23]NHLS Port Elizabeth Laboratory, Port Elizabeth, South Africa
[24]Faculty of Health Sciences, Walter Sisulu University, Eastern Cape, South Africa
[25]Public Health Department, Integrated Disease Surveillance and Response, Ministry of Health and Wellness, Gaborone, Botswana
[26]Department of Immunology and Infectious Diseases, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, USA

[27]Division of Medical Virology, Faculty of Medicine and Health Sciences, Stellenbosch University, Tygerberg, Cape Town, South Africa
[28]Department of Virology, Charlotte Maxeke Johannesburg Academic Hospital, Johannesburg, South Africa
[29]Covid-19 Taskteam, Gaborone, Botswana
[30]Botswana-Baylor Children's Clinical Centre of Excellence
[31]Baylor College of Medicine, Houston, Texas, USA
[32]Department of Medical Virology, University of Pretoria, Pretoria, South Africa
[33]National Health Laboratory, Health Services Management, Ministry of Health and Wellness, Gaborone, Botswana
[34]National Health Laboratory Service (NHLS), Johannesburg, South Africa
[35]Centre for the AIDS Programme of Research in South Africa (CAPRISA), Durban, South Africa
[36]Centre for Epidemic Response and Innovation (CERI), School of Data Science and Computational Thinking, Stellenbosch University, Stellenbosch, South Africa
[37]Department of Medicine, Faculty of Medicine, University of Botswana, Gaborone, Botswana
[38]Department of Medical Laboratory Sciences, School of Allied Health Professions, Faculty of Health Sciences, University of Botswana, Gaborone, Botswana
[39]Discipline of Virology, School of Laboratory Medicine and Medical Sciences and National Health Laboratory Service (NHLS), University of KwaZulu–Natal, Durban, South Africa
[40]Next Generation Sequencing Unit, University of the Free State, Bloemfontein, South Africa
[41]Department of Molecular Medicine and Haematology, University of the Witwatersrand, Johannesburg, South Africa
[42]Department of Veterinary Tropical Diseases, Faculty of Veterinary Science, University of Pretoria, Onderstepoort, South Africa
[43]Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa
[44]PathCare Vermaak, Pretoria, South Africa
[45]Institute of Infectious Disease and Molecular Medicine, University of Cape Town, Cape Town, South Africa
[46]Department of Global Health, University of Washington, Seattle, WA, USA

*These authors contributed equally: Raquel Viana, Sikhulile Moyo, Daniel Amoako, Houriiyah Tegally, Cathrine Scheepers
**These authors jointly supervised the work: Simani Gaseitsiwe, Anne von Gottberg, Tulio de Oliveira

Abstract
In southern Africa, the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) epidemic has been characterised by three distinct waves. The first wave was associated with a mix of SARS-CoV-2 lineages, whilst the second and third waves were driven by the Beta and Delta variants respectively[1–3]. In November 2021, genomic surveillance teams in South Africa and Botswana detected a new variant associated with a rapid resurgence of infections in Gauteng Province, South Africa. This new variant is characterised by over 30 mutations in the spike glycoprotein, predicted to influence antibody neutralization and spike function[4]. Within three days this was designated a variant of concern (Omicron) by the World Health Organization and, within three weeks, it had been identified in 87 countries. Here, we describe the genomic profile and early transmission dynamics of Omicron, highlighting the rapid spread in regions with high levels of population immunity.

Introduction
Since the onset of the COVID-19 pandemic in December 2019, variants of SARS-CoV-2 have continuously emerged with some spreading around the world; in many cases making major contributions to the cyclical infection waves that occur asynchronously in different regions. Between October and December 2020, the world witnessed the emergence of the first variants of concern (VOC): variants exhibiting increased transmissibility and/or immune evasion properties that threatened global efforts to control the pandemic. Although the Alpha, Beta and Gamma

VOCs[2,5] that emerged in November and December 2020 spread widely around the world and showed early signs of driving resurgences in different regions, it was the highly transmissible Delta variant that displaced all other VOC in most regions of the world[6]. During its spread, the Delta variant evolved into multiple sub-lineages[7], some of which demonstrated signs of having a growth advantage in certain locations, prompting speculation that the next VOC driving resurgences of infections would likely be derived from Delta. However, in October 2021, while Delta was driving continued high-level transmission in the Northern hemisphere, in southern Africa a large Delta wave was subsiding. The culmination of this wave coincided with the emergence of a novel SARS-CoV-2 variant that, within four days of its almost simultaneous discovery in four individuals in Botswana, a traveler from South Africa in Hong Kong, and 54 individuals in South Africa, was designated by the World Health Organization as Omicron: the fifth VOC.

Results & Discussion

Epidemic dynamics and detection of Omicron

The three distinct epidemic waves experienced by southern African countries were each driven by different variants: the first by offshoots of the B.1 lineage[1], the second by the Beta VOC[2,8], and the third by the Delta VOC[3], with an estimated 2-5% of third wave cases attributed to the C.1.2 lineage[9] (**Fig. 4.1A**). Serosurveys conducted even before the Delta wave suggested high levels of exposure to SARS-CoV-2 (40-60%) in South Africa[10,11], Malawi[12], and Zimbabwe[13]. Modelled estimates suggested seroprevalence of 70-80% across South Africa by October 2021[14]. Accordingly, the weeks following the third wave in Southern Africa between 10 October and 15 November 2021 were marked by a period of lower-level transmission as indicated by low incidence of reported COVID-19 cases (100-200 new cases per day) and low (<2%) test positivity rates (**Fig. 4.1A-1C**).

A rapid increase in COVID-19 cases was noted in the week beginning 15 November 2021 in Gauteng province. Specifically, rising case numbers and test positivity rates were first noticed in the Tshwane Metropolitan area, initially associated with cluster outbreaks in higher education settings. This resurgence of cases was accompanied by the identification of multiple instances of S-gene target failure (SGTF) during TaqPath-based (Thermo Fisher Scientific) diagnostic PCR testing: a phenomenon previously observed with the Alpha variant due to a deletion at positions 69 and 70 (Δ69-70) in the spike protein[15]. Given the low prevalence of Alpha in South Africa (**Fig. 4.1A**), targeted whole-genome sequencing of these specimens was prioritized.

On 19 November 2021, sequencing results of an initial batch of 8 SGTF samples collected between 14-16 November 2021 indicated that all were caused by a novel variant of SARS-CoV-2. Further rapid sequencing identified the same variant in 29 of 32 routine diagnostic samples from multiple locations in Gauteng Province and suggested widespread circulation of this new variant by the second week of November. Crucially, this rise immediately preceded a sharp increase in reported case numbers (**Fig. 4.1C, Appendix C - Extended Data Fig. 1**). In the following four days the presence of Omicron was confirmed by sequencing in another two provinces: KwaZulu-Natal (KZN) and the Western Cape (**Fig. 4.1B**).

Concurrently in Gaborone, Botswana (~360km from Tshwane), four genomes generated from samples collected on 11 November 2021, and sequenced on 17-18 November 2021 as part of weekly surveillance, displayed an unusual set of mutations. These were reported to the Botswana

Ministry of Health and Wellness on 22 November 2021, as “unusual sequences” that were linked to a group of visitors (non-residents) on a diplomatic mission. The sequences were uploaded to GISAID [16,17] on 23 November 2021, and it became apparent that they belonged to a new lineage. A further 15 cases were identified within the same week from various other locations in Botswana, confirming the circulation of the new lineage, with most being travellers from other countries, including South Africa, and linked to local transmission chains.

On 24 November 2021, this novel variant was designated as a new PANGO lineage (B.1.1.529)[18]. On November 26, 2021, it was designated a VOC and named Omicron by the WHO on the recommendation of the Technical Advisory Group on SARS-CoV-2 Virus Evolution[19]. By the first week of December 2021, Omicron was driving a rapid and sustained increase in cases in South Africa and Botswana (**Fig. 4.1C, Appendix C - Extended Data Fig. 2** for Botswana). In Gauteng, weekly test positivity rates increased from <1% in the week beginning 31 October, to 16% in the week beginning 21 November 2021, and 35% in the week beginning 28 November, simultaneously with an exponential rise in COVID-19 incidence (**Fig. 4.1C, Appendix C - Extended Data Fig. 1**). Nationally, the daily case numbers exceeded 22 000 (84% of the peak of the previous wave of infections) by 9 December 2021. At the same time, the proportion of TaqPath PCR tests with SGTF increased rapidly in all provinces of South Africa reaching ~90% nationally by the week beginning 21 November 2021, giving a strong indication that the fourth wave was being driven by Omicron: an indication that has now been confirmed by genomic sequencing in all provinces (**Fig. 4.1C**). Similarly, Botswana experienced a sharp increase in cases, doubling every 2-3 days late November to early December 2021, transitioning from a 7-day moving average of <10 cases/100 000 to above 25 cases/100,000 in less than 10 days (**Appendix C - Extended Data Fig. 2)**.

By 16 December 2021, Omicron had been detected in 87 countries, both in samples from travelers returning from southern Africa, and in samples from routine community testing (**Appendix C - Extended Data Fig. 3**).

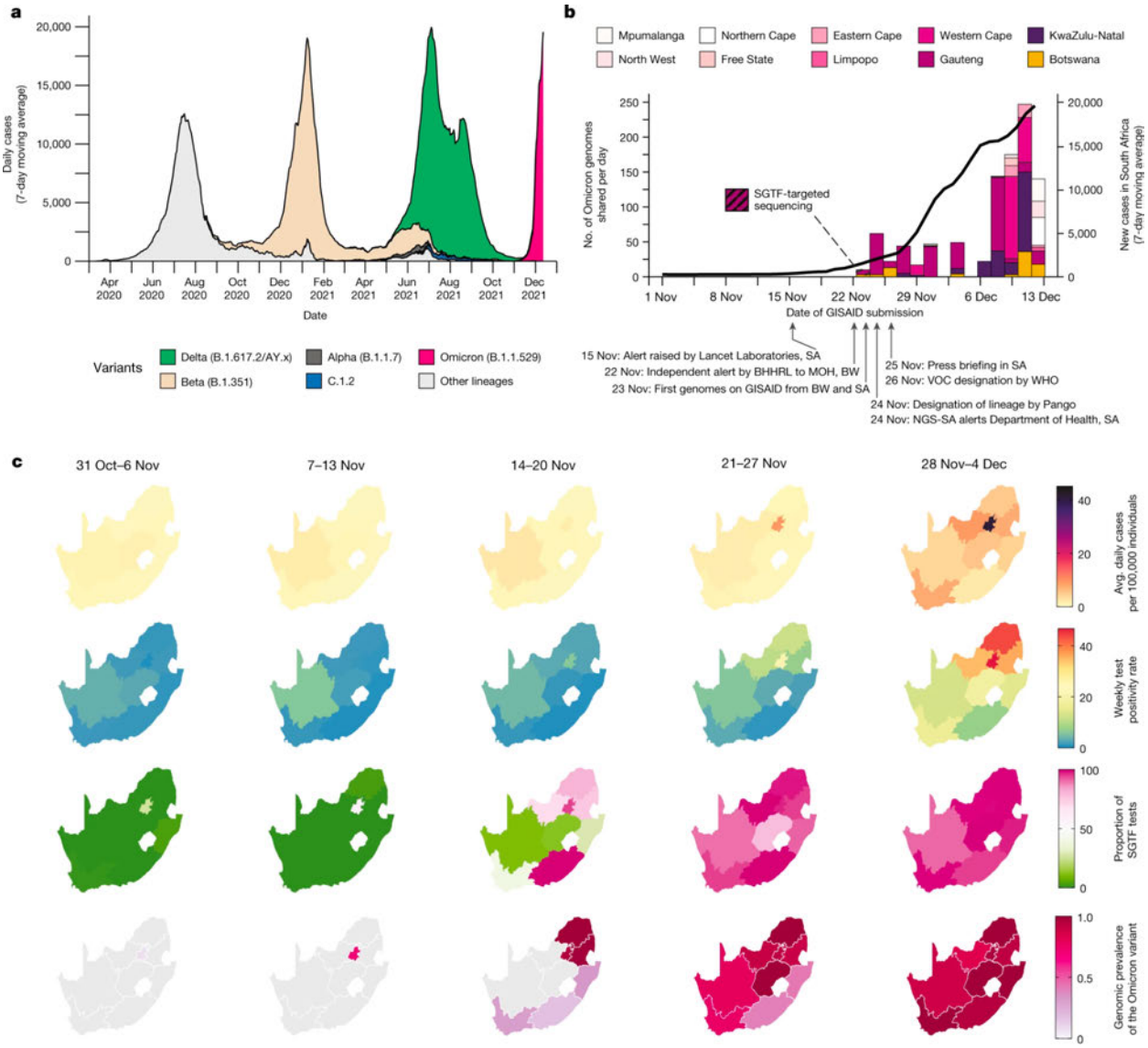


Figure 4.1: Detection of Omicron variant A) The progression of daily cases in South Africa from March 2020 to December 2021 where the 7-day rolling average of daily case numbers is further coloured by the inferred proportion of variants responsible for the infections, as calculated by genomic surveillance data on GISAID. B) Timeline of Omicron detection in Botswana and South Africa. Bars represent the number of Omicron genomes shared per day, according to the date they were uploaded to GISAID, while the line represents the 7-day moving average of daily new cases in South Africa. C) Weekly progression of average daily cases per 100,000, test positivity rates, proportion of SGTF tests (on the TaqPath COVID-19 PCR assay) and genomic prevalence of Omicron in nine provinces of South Africa for five weeks from 31 October to 4 December 2021.

The evolutionary origins of Omicron

To determine when and where Omicron likely originated, we analyzed all 686 available Omicron genomes (including 248 from southern Africa and 438 from elsewhere in the world) retrieved from

GISAID (date of access 7 December 2021)[16,17], against a global reference set of representative SARS-CoV-2 genomes (n=12 609) collected between December 2019 and November 2021. Preliminary maximum-likelihood phylogenies identified the BA.1/Omicron sequences as a monophyletic clade rooted within the B.1.1 lineage (20B clade), with no clear basal progenitor **(Fig. 4.2A)**. Importantly, the BA.1/Omicron cluster is phylogenetically distinct from any known VOC or variants of interest (VOI) or any other lineages known to be circulating in southern Africa (e.g. C.1.2) **(Fig. 4.2A)**. More recently, two related lineages have emerged (BA.2 and BA.3), both sharing many, but not all of the characteristic mutations of BA.1/Omicron and both having many unique mutations of their own. We primarily focus here on the BA.1 lineage which is rapidly spreading in multiple countries around the world and is the lineage officially designated as the Omicron VOC.

Time-calibrated Bayesian phylogenetic analysis of all BA.1 assigned genomes from southern Africa (as of 11 December 2021, n=553) estimated the time when the most recent common ancestor of the analysed BA.1 lineage sequences existed as 9 October 2021 (95% credible intervals 30 September - 20 October) with a per-day growth rate of 0.136 (95% confidence interval (CI) 0.100 – 0.173) reflecting a doubling time of 5.1 days (95% CI 4.0 – 6.9) **(Fig 4.2B)**. These estimates are robust to whether the evolutionary rate is estimated from the data or fixed to previously estimated values (**Appendix C - Extended Data Table S1**). Limiting the analysis to a subset of genomes from Gauteng Province only (279 genomes) yields a faster growth rate estimate with a doubling time of 1.8 days (95% CI 1.4 – 3.0) (**Appendix C - Extended Data S1**). Spatiotemporal phylogeographic analysis indicates that the BA.1/Omicron variant spread from the Gauteng province of South Africa to seven of the eight other provinces and to two regions of Botswana from late October to late November 2021, and shows more recent evidence of transmission within and between other South African provinces (**Fig 4.2C**).

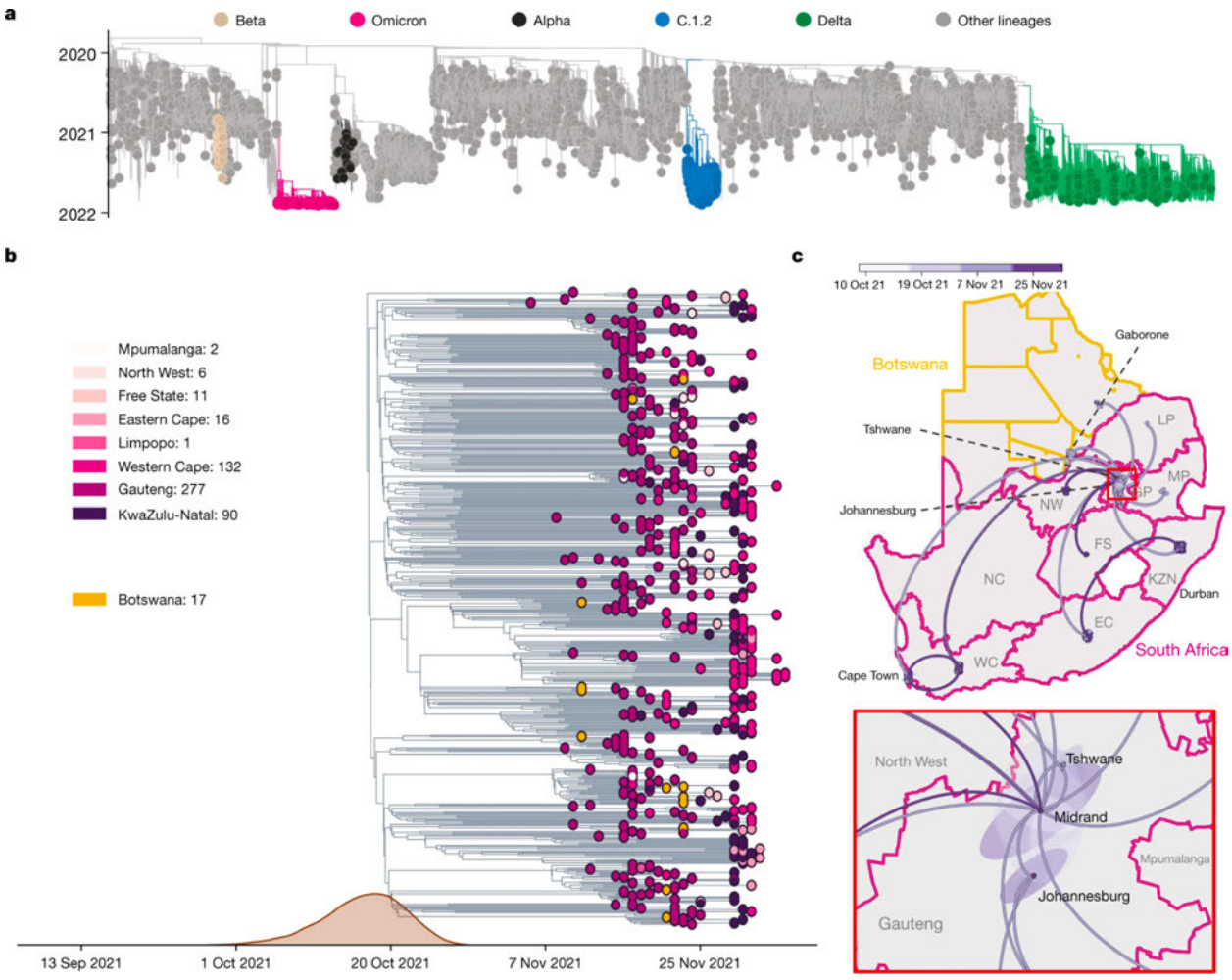


Figure 4.2: Evolution of Omicron. A) Time-resolved maximum likelihood phylogeny of 13,295 SARS-CoV-2 sequences; 9,944 of these are from Africa (denoted with tip point circle shapes). Alpha, Beta and Delta VOCs and the C.1.2 lineage, recently circulating in South Africa, are denoted in black, brown, green and blue respectively. The newly identified SARS-CoV-2 Omicron variant is shown in pink. Genomes of other lineages are shown in grey. B) Time-resolved maximum clade credibility phylogeny of the Omicron cluster of southern African genomes (n = 553), with locations indicated. The distribution of estimated time of origin is also shown. C) Spatiotemporal reconstruction of the spread of the Omicron variant in Southern Africa with an inset of Gauteng province. Circles represent nodes of the maximum clade credibility phylogeny, coloured according to their inferred time of occurrence (scale in top panel). Shaded areas represent the 80% highest posterior density interval and depict the uncertainty of the phylogeographic estimates for each node. Solid curved lines denote the links between nodes and the directionality of movement is anticlockwise along the curve.

Molecular profile of Omicron

Omicron carries 15 mutations in the spike receptor-binding domain (RBD) (**Fig. 4.3**), five of which (G339D, N440K, S477N, T478K, N501Y) have been shown individually to enhance hACE2

binding[20]. A further six of these RBD mutations (K417N, G446S, E484A, Q493R, G496S, Q498R and N501Y) are expected to have moderate to strong impacts on binding of at least three of the four major classes of spike-targeted neutralizing antibodies (NAbs)[21–23]. These RBD mutations coupled with four amino acid substitutions (A67V, T95I, G142D, and L212I), three deletions (69-70, 143-145 and 211) and an insertion (EPE between 214 and 215) in the N-terminal domain (NTD)[24], are predicted to underlie the substantially reduced sensitivity of Omicron to neutralization by anti-SARS-CoV-2 antibodies induced by either infection or vaccination[25,26]. These mutations also involve key structural epitopes targeted by some of the currently authorized monoclonal antibodies, particularly bamlanivimab + etesevimab and casirivimab + imdevimab[27–29]. Preliminary analysis suggests that although the spike mutations involve a number of T cell and B cell epitopes, the majority of epitopes (>70%) remain unaffected[30].

Omicron also has a cluster of three mutations (H655Y, N679K and P681H) adjacent to the S1/S2 furin cleavage site (FCS) which are likely to enhance spike protein cleavage and fusion with host cells[31,32] and which could also contribute to enhanced transmissibility[33] (**Appendix C - Extended Data Fig. 4**).

Outside of the spike protein, a deletion in nsp6 (105-107del), in the same region as deletions seen in Alpha, Beta, Gamma and Lambda, may have a role in evasion of innate immunity[34]; and the double mutation in nucleocapsid (R203K, G204R), also present in Alpha, Gamma and C.1.2, has been associated with enhanced infectivity in human lung cells [35].

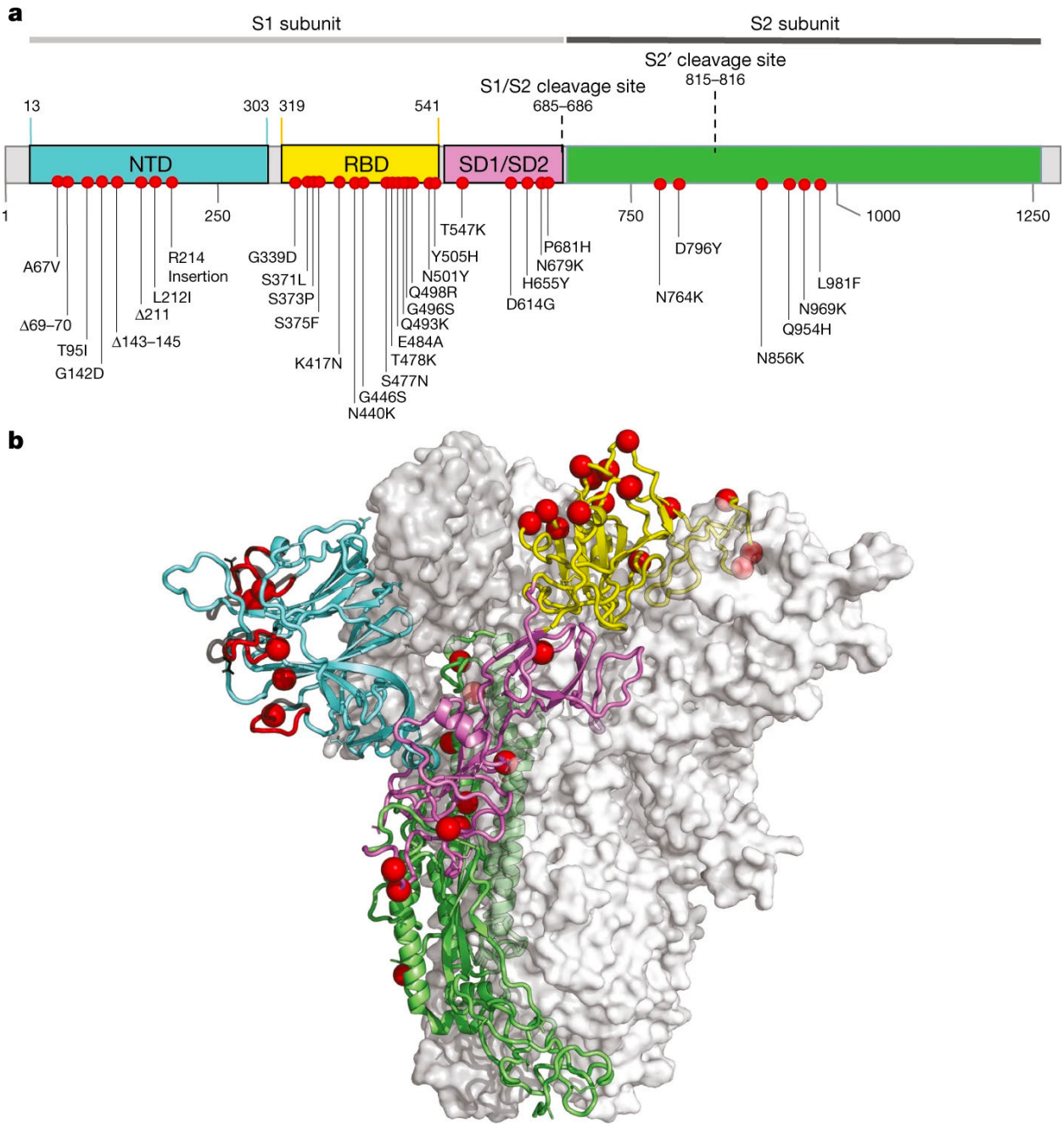


Figure 4.3. Molecular profile of Omicron A) Amino-acid mutations on the spike gene of the BA.1/Omicron variant. B) Structure of the SARS-CoV-2 Spike trimer, showing a single spike protomer in cartoon view. The N terminal domain, receptor binding domain, subdomains 1 and 2, and the S2 protein are shown in cyan, yellow, pink, and green respectively. Red spheres indicate the alpha carbon positions for each omicron variant residue. NTD-specific loop insertions/deletions are shown in red, with the original loop shown in transparent black.

Omicron is not obviously recombinant

Given the large number of mutations differentiating BA.1/Omicron and BA.2 from other known SARS-CoV-2 lineages it was considered plausible that either (i) both of these lineages might have descended from a common recombinant ancestor, or (ii) that one of the BA lineages might have originated via recombination between a virus in the other BA lineage and a virus in a non-BA lineage. We tested these hypotheses using a variety of recombination detection approaches (implemented in the programs GARD[36]; 3SEQ[37]; and RDP5[38]) to identify potential signals of recombination in sequence datasets containing BA.1 and BA.2 sequences together with sequences representative of global SARS-CoV-2 genomic diversity.

3SEQ, GARD and RDP5 all identified potential evidence of recombination in these datasets. The most likely recombination breakpoint locations were located between nucleotide positions 20520 and 21619 (near the start of the S-gene; supported by GARD, RDP5 and 3SEQ) and between 23609 and 23614 (near the middle of the S-gene; supported by GARD and 3SEQ). 3SEQ identified an additional breakpoint at nucleotide position 24513 (toward the end of the S-gene). Phylogenetic analysis of the genome regions bounded by these breakpoints (genome coordinates 1450-20520, 21619-23609 and 23614-24513) revealed no support for a recombinant origin for either the BA.1 or BA.2 lineages (**Appendix C - Extended Data Fig. 5**). Although one BA.1 isolate (Botswana/R43B66) displayed evidence of having potentially inherited nucleotides 23614-24513 from a Delta virus by recombination, there was no strong phylogenetic support for the clustering of this sequence with Delta viruses. Further, read coverage in this region of the Botswana/R43B66 sequence was so low that we were unable to exclude the possibility that the apparent recombination signal was attributable to a combination of miscalled/uncalled nucleotides and alignment uncertainty.

Although we found no convincing phylogenetic or statistical evidence of either the most recent common ancestor of BA.1 and BA.2 being recombinant, or of the most recent common ancestors of either the BA.1 or BA.2 lineages having been derived through recombination, it should be noted that recombination tests in general will not have sufficient statistical power to reliably identify evidence of individual recombination events that result in transfers of less than ~5 contiguous polymorphic nucleotide sites between genomes. Further, if BA.1 and/or BA.2 are the products of a series of multiple partially overlapping recombination events occurring across multiple temporally clustered replication cycles, the complex patterns of nucleotide variation that might result could be extremely difficult to interpret as recombination using the methods applied here[39].

Signs of strong selection and epistasis during the origin and ongoing evolution of the Omicron lineage

We applied a selection analysis pipeline to all available sequences designated as BA.1 in GISAID as of 8 December 2021. The analysis followed the procedure described previously[34], and downsampled alignments of individual protein encoding regions to obtain a median of 25 unique Omicron haplotype sequences and 107 unique haplotype sequences for each gene/ORF from a representative selection of other SARS-CoV-2 lineages (used as background sequences to contextualize evolution within the Omicron sub-clade).

We detected evidence of gene-wide positive selection (using the BUSTED method[40]) acting on six genes/ORFs since the ancestral BA.1/Omicron and BA.2 lineage split from the B.1.1 lineage: S-

gene ($p < 0.0001$), exonuclease ($p < 0.0001$), nsp6 ($p = 0.001$), M-gene ($p = 0.002$), N-gene ($p = 0.006$), and E-gene ($p = 0.05$). In all six genes, this selection was strong ($dN/dS > 10$), and occurred in bursts (≤ 6% of branch/site combinations selected). The branch separating BA.1/Omicron from its most recent B.1.1 ancestor had the most prominent selection signal (which was strongest in the S-gene; BUSTED p-value < 0.0001 with $dN/dS > 100$ at ~0.5% of S-gene codon sites[41]), strongly supporting the hypothesis that adaptive evolution played a significant role in the mutational divergence of Omicron from other B.1.1 SARS-CoV-2 lineages. Relative to the intensity of selection evident within the background B.1.1 lineages, selection in three genes was likely significantly intensified in the ancestral Omicron lineage: S-gene (intensification factor $K = 2.0$; $p < 0.0001$ [42]), exonuclease ($K = 4.0$; $p < 0.0001$), and nsp6 ($K = 5.1$; $p = 0.02$).

Among 294 codon sites that are polymorphic among the BA.1/Omicron sequences analysed, 32 were found to have experienced episodic positive selection since BA.1 split from the B.1.1 lineage (MEME $p \leq 0.01$, **Appendix C - Extended Data Table S2** [43]). Sixteen (50%) of these codon sites are in the S-gene, 13 of which contain BA.1 lineage-defining mutations (i.e. these selection signals reflect mutations that occurred within the ancestral Omicron lineage). The three positively selected codon sites that did not correspond to sites of lineage-defining mutations (S/346, S/452, and S/701) are particularly notable as these are attributable to mutations that have occurred since the MRCA of the analysed BA.1 sequences. The mutations driving the positive selection signals at these three sites in the Omicron S-gene converge on mutations seen in other VOCs or VOIs (R346K in Mu, L452R in Delta, and A701V in Beta and Iota). The A701V mutation, the precise impact of which is currently unknown, is one of 19 in a proposed "501Y lineage Spike meta-signature" comprising the set of mutations that were most adaptive during the evolution of the Alpha, Beta and Gamma VOC lineages[34]. Further, both R346K and L452R are known to impact antibody binding[22] and both of the codon sites where these mutations occur display evidence for directional selection (using the FADE method[44]). These selective patterns suggest that, during its current explosive spread, Omicron may be undergoing additional evolution to modify its neutralization profile.

Potential for increased transmissibility and immune evasion

We estimated that Omicron had a growth advantage of 0.24 (95% CI: 0.16-0.33) per day over Delta in Gauteng, South Africa (**Fig. 4.4A**). This corresponds to a 5.4-fold (95% CI: 3.1-10.1) weekly increase in cases compared to Delta. The growth advantage of Omicron is likely to be mediated by (i) an increase relative to other variants of its intrinsic transmissibility (i.e., the basic reproduction number R_0), (ii) an increase relative to other variants in its capacity to infect, and be transmitted from, previously infected and vaccinated individuals; or (iii) both.

The predicted combination of transmissibility and immune evasion for Omicron strongly depends on the assumed level of current population immunity against infection by, and transmission of, the competing variant Delta that is afforded by prior-infections with wild-type Wuhan, Beta, Delta, and other strains, and/or vaccination (**Fig. 4.4B**). For moderate levels of population immunity against Delta ($\Omega = 0.4$), immune evasion alone cannot explain the observed growth advantage of Omicron (**Fig. 4.4C**). For medium levels of immunity against Delta ($\Omega = 0.6$), very high levels of immune evasion could explain the observed growth advantage without an additional increase in transmissibility (**Fig. 4.4D**). For high levels of population immunity against Delta ($\Omega = 0.8$), even moderate levels of immune evasion (~25-50%) can explain the observed growth advantage without an additional increase in transmissibility (**Fig. 4.4E**). The results of seroprevalence studies and

vaccination coverage (~40% of the adult population) in South Africa suggest that the proportion of the population with potential immunity against Delta and earlier variants is likely to be above 60%[10,11]. We thus argue that the population level of protective immunity against Delta is high, and that partial immune evasion is a major driver for the observed dynamics of Omicron in South Africa. This notion is supported by recent findings that show an increased risk of SARS-CoV-2 reinfection associated with the emergence of Omicron in South Africa[45] and the initial results from neutralization assays [25,26]. An increase, or decrease, in the transmissibility of Omicron compared to Delta cannot, however, be ruled out.

There are a number of limitations to this analysis. First, we estimated the growth advantage of Omicron based on early sequence data only. These data could be biased due to targeted sequencing of SGTF samples and stochastic effects (e.g., superspreading) in a low incidence setting, which can lead to overestimates of the growth advantage, and consequently of the increased transmissibility and immune evasion. Second, without reliable estimates of the level of protective immunity against Delta in South Africa, we cannot obtain precise estimates of transmissibility or immune evasion of Omicron.

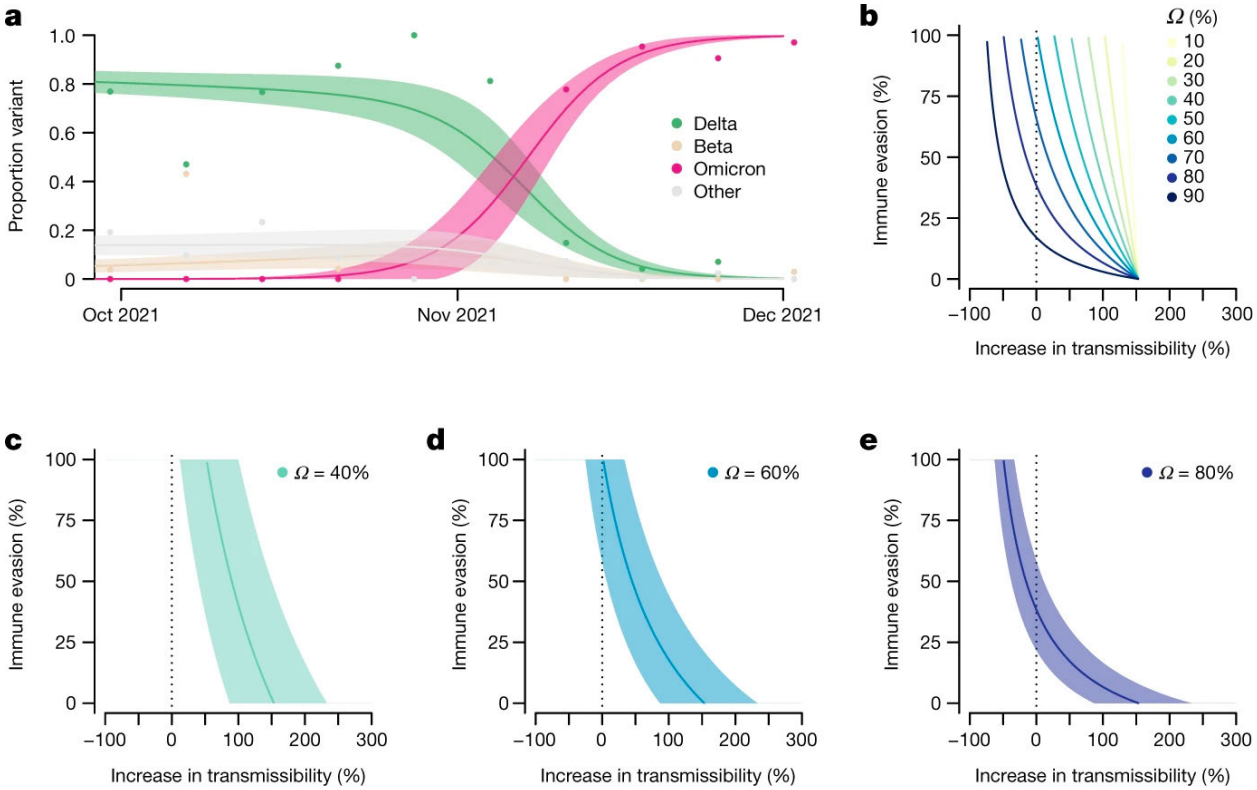


Figure 4.4: Growth of Omicron in Gauteng, South Africa, and relationship between potential increase in transmissibility and immune evasion. (A) Omicron rapidly outcompeted Delta in November 2021. Model fits are based on a multinomial logistic regression. Dots represent the weekly proportions of variants. (B) The relationship between the potential increase in transmissibility and immune evasion strongly depends on the assumed level of current population immunity against Delta (Ω). (C-E) Relationship for a population immunity of 40%, 60%, and 80% against infection and transmission with Delta. The dark vertical dashed line indicates equal

transmissibility of Omicron compared to Delta. Shaded areas correspond to the 95% CIs of the model estimates.

Conclusion

Omicron is now driving a fourth wave of the SARS-CoV-2 epidemic in South Africa and Botswana, and is now already spreading rapidly in several other countries. Genotypic and phenotypic data suggest that Omicron has the capacity for substantial evasion of neutralizing antibody responses, and modelling suggests that immune evasion could be the major driver of the observed transmission dynamics. Close monitoring of the spread of Omicron in countries outside southern Africa will be necessary to better understand its transmissibility and the potential of this variant to evade post-infection and vaccine-elicited immunity. Neutralizing antibodies are only one component of the immune protection from vaccines and prior infection, and cellular immunity is predicted to be less affected by the mutations in Omicron. Vaccination therefore remains critical to protect those at highest risk of severe disease and death. The emergence and rapid spread of Omicron poses a threat to the world and a particular threat in Africa, where fewer than one in ten people is fully vaccinated.

Methods

Epidemiological dynamics

We analyzed daily cases of SARS-CoV-2 in South Africa up to 14 December 2021 from publicly released data provided by the National Department of Health and the National Institute for Communicable Diseases. This was accessible through the repository of the Data Science for Social Impact Research Group at the University of Pretoria (https://github.com/dsfsi/covid19za)[46,47]. The National Department of Health releases daily updates on the number of confirmed new cases, deaths and recoveries, with a breakdown by province. Daily case numbers for Botswana were obtained via Our World in Data (OWID) COVID-19 data repository (https://github.com/owid/covid-19-data). We consulted estimates for the *R*e of SARS-CoV-2 in South Africa and Botswana from the 'covid-19-Re' data repository (https://github.com/covid-19-Re/dailyRe-Data)[48]. We obtained test positivity data from weekly reports from the National Institute for Communicable Diseases (NICD)[49]. Data to calculate the proportion of positive Thermo Fisher TaqPath COVID-19 PCR tests with SGTF in South Africa was obtained from the National Health Laboratory Service and Lancet Laboratories. Test positivity data for Botswana was obtained from the National Health Laboratory through 6 December 2021. All data visualization was generated through the ggplot package in R[50].

SARS-CoV-2 sampling

As part of the NGS-SA, seven sequencing hubs in South Africa receive randomly selected samples for sequencing every week according to approved protocols at each site[51]. These samples include remnant nucleic acid extracts or remnant nasopharyngeal and oropharyngeal swab samples from routine diagnostic SARS-CoV-2 PCR testing from public and private laboratories in South Africa. In response to a focal resurgence of COVID-19 in the City of Tshwane Metropolitan Municipality in Gauteng Province in November, we enriched our routine sampling with additional samples from the affected area, including initial targeted sequencing of SGTF samples. In Botswana, all public and private laboratories submit randomly selected residual nasopharyngeal and oropharyngeal

PCR positive samples weekly to the National Health Laboratory (NHL) and the Botswana Harvard HIV Reference Laboratory (BHHRL) for sequencing.

Ethical statement

The genomic surveillance in South Africa was approved by the University of KwaZulu–Natal Biomedical Research Ethics Committee (BREC/00001510/2020), the University of the Witwatersrand Human Research Ethics Committee (HREC) (M180832), Stellenbosch University HREC (N20/04/008 COVID-19), University of Cape Town HREC (383/2020), University of Pretoria HREC (H101/17) and the University of the Free State Health Sciences Research Ethics Committee (UFS-HSD2020/1860/2710). The genomic sequencing in Botswana was conducted as part of the national vaccine roll-out plan and was approved by the Health Research and Development Committee (Health Research Ethics body, HRDC#00948 and HRDC#00904). Individual participant consent was not required for the genomic surveillance. This requirement was waived by the Research Ethics Committees.

Ion Torrent Genexus Integrated Sequencer methodology for rapid whole genome sequencing of SARS-CoV-2

Viral RNA was extracted using the MagNA Pure 96 DNA and Viral Nucleic Acid kit on the automated MagNA Pure 96 system (Roche Diagnostics, USA) as per the manufacturer's instructions. Extracts were then screened by qPCR to acquire the mean cycle threshold (Ct) values for the SARS-CoV-2 N-gene and ORF1ab-gene using the TaqMan 2019-nCoV assay kit v1 (ThermoFisher Scientific, USA) on the ViiA7 Real-time PCR system (ThermoFisher Scientific, USA) as per the manufacturer's instructions. Extracts were sorted into batches of N=8 within a Ct range difference of 5 for a maximum of two batches per run. Extracts with <200 copies were sequenced using the low viral titer protocol. Next-generation sequencing was performed using the Ion AmpliSeq SARS-CoV-2 Research Panel on the Ion Torrent Genexus Integrated Sequencer (ThermoFisher Scientific, USA) which combines automated cDNA synthesis, library preparation, templating preparation and sequencing within 24 hours. The Ion Ampliseq SARS-CoV-2 Research Panel consists of 2 primer pools targeting 237 amplicons tiled across the SARS-CoV-2 genome providing >99% coverage of the SARS-CoV-2 genome (~30 kb) and an additional 5 primer pairs targeting human expression controls. The SARS-CoV-2 amplicons range from 125 to 275 bp in length. TRINITY was utilised for de novo assembly and the Iterative Refinement Meta-Assembler (IRMA) for genome assisted assembly as well as FastQC for quality checks.

Whole-genome sequencing and genome assembly

RNA was extracted on an automated Chemagic 360 instrument, using the CMG-1049 kit (Perkin Elmer, Hamburg, Germany). The RNA was stored at −80 ◦C prior to use. Libraries for whole genome sequencing were prepared using either the Oxford Nanopore Midnight protocol with Rapid Barcoding or the Illumina COVIDseq Assay.

Illumina Miseq/NextSeq

For the Illumina COVIDseq assay, the libraries were prepared according to the manufacturer's protocol. Briefly, amplicons were tagmented, followed by indexing using the Nextera UD Indexes Set A. Sequencing libraries were pooled, normalized to 4 nM and denatured with 0.2 N sodium acetate. A 8 pM sample library was spiked with 1% PhiX (PhiX Control v3 adaptor-ligated library used as a control). We sequenced libraries on a 500-cycle v2 MiSeq Reagent Kit on the Illumina

MiSeq instrument (Illumina). On the Illumina NextSeq 550 instrument, sequencing was performed using the Illumina COVIDSeq protocol (Illumina Inc, USA), an amplicon-based next-generation sequencing approach. The first strand synthesis was carried using random hexamers primers from Illumina and the synthesized cDNA underwent two separate multiplex PCR reactions. The pooled PCR amplified products were processed for tagmentation and adapter ligation using IDT for Illumina Nextera UD Indexes. Further enrichment and cleanup was performed as per protocols provided by the manufacturer (Illumina Inc). Pooled samples were quantified using Qubit 3.0 or 4.0 fluorometer (Invitrogen Inc.) using the Qubit dsDNA High Sensitivity assay according to manufacturer's instructions. The fragment sizes were analyzed using TapeStation 4200 (Invitrogen). The pooled libraries were further normalized to 4nM concentration and 25 µl of each normalized pool containing unique index adapter sets were combined in a new tube. The final library pool was denatured and neutralized with 0.2N sodium hydroxide and 200 mM Tris-HCL (pH7), respectively. 1.5 pM sample library was spiked with 2% PhiX. Libraries were loaded onto a 300-cycle NextSeq 500/550 HighOutput Kit v2 and run on the Illumina NextSeq 550 instrument (Illumina, San Diego, CA, USA).

Midnight Protocol

For Oxford Nanopore sequencing, the Midnight primer kit was used as described by Freed and Silander[52]. cDNA synthesis was performed on the extracted RNA using LunaScript RT mastermix (New England BioLabs) followed by gene-specific multiplex PCR using the Midnight Primer pools which produce 1200bp amplicons which overlap to cover the 30-kb SARS-CoV-2 genome. Amplicons from each pool were pooled and used neat for barcoding with the Oxford Nanopore Rapid Barcoding kit as per the manufacturer's protocol. Barcoded samples were pooled and bead-purified. After the bead clean-up, the library was loaded on a prepared R9.4.1 flow-cell. A GridION X5 or MinION sequencing run was initiated using MinKNOW software with the base-call setting switched off.

Genome assembly

We assembled paired-end and nanopore .fastq reads using Genome Detective 1.132 (https://www.genomedetective.com) which was updated for the accurate assembly and variant calling of tiled primer amplicon Illumina or Oxford Nanopore reads, and the Coronavirus Typing Tool[53]. In addition, we also used the wf artic (ARTIC SARS-CoV-2) pipeline as built using the nextflow workflow framework [54]. For Illumina assembly, GATK HaploTypeCaller --min-pruning 0 argument was added to increase mutation calling sensitivity near sequencing gaps. For Nanopore, low coverage regions with poor alignment quality (<85% variant homogeneity) near sequencing/amplicon ends were masked to be robust against primer drop-out experienced in the Spike gene, and the sensitivity for detecting short inserts using a region-local global alignment of reads, was increased. In some instances, mutations were confirmed visually with .bam files using Geneious software V2020.1.2 (Biomatters). The reference genome used throughout the assembly process was NC_045512.2 (numbering equivalent to MN908947.3).

Raw reads from the Illumina COVIDSeq protocol were assembled using the Exatype NGS SARS-CoV-2 pipeline v1.6.1, (https://sars-cov-2.exatype.com/). This pipeline performs quality control on reads and then maps the reads to a reference using Examap. The reference genome used throughout the assembly process was NC 045512.2 (Accession number: MN908947.3).

Several of the initial Ion Torrent genomes contained a number of frameshifts, which caused unknown variant calls. Manual inspection revealed that these were likely to be sequencing errors resulting in mis-assembled regions (likely due to the known error profile of Ion Torrent sequencers)[55]. To resolve this, the raw reads from the IonTorrent platform were assembled using the SARSCoV2 RECoVERY (REconstruction of COronaVirus gEnomes & Rapid analYsis) pipeline implemented in the Galaxy instance ARIES (https://aries.iss.it). This pipeline fixed the observed frameshifts, confirming that they were artefacts of mis-assembly; this subsequently resolved the variant calls. The Exatype and RECoVERY pipelines each produce a consensus sequence for each sample. These consensus sequences were manually inspected and polished using Aliview v1.27 (http://ormbunkar.se/aliview/).

All of the sequences were deposited in GISAID (https://www.gisaid.org/) [16,17], and the GISAID accession identifiers are included as part of **Supplementary Table S3**. Raw reads for our sequences have also been deposited at the NCBI Sequence Read Archive (BioProject accession PRJNA784038).

The number and position of the Omicron mutations has affected a number of primers and caused primer drop-outs across a range of sequencing protocols, especially within the RBD (https://primer-monitor.neb.com/lineages). These primer drop-outs have resulted in a number of genomes missing stretches of the RBD, and can affect estimates of mutation prevalence and the determination of the true set of lineage-defining mutations. Given this, bam files of all initial genomes were inspected with IG Viewer to confirm mutation calls where reference calls were suspected to be from low coverage at primer dropout sites[56].

Lineage classification

We used the widespread dynamic lineage classification method from the 'Phylogenetic Assignment of Named Global Outbreak Lineages' (PANGOLIN) software suite (https://github.com/hCoV-2019/pangolin)[18]. This is aimed at identifying the most epidemiologically important lineages of SARS-CoV-2 at the time of analysis, enabling researchers to monitor the epidemic in a particular geographic region. For the Omicron variant described in this study, the corresponding PANGO lineage designation is BA.1 (lineages v1.2.106). When first characterized the lineage was designated as B.1.1.529 but the emergence of a sibling lineage to Omicron resulted in the split into two sublineages (B.1.1.529.1 and B.1.1.529.2, aliased as BA.1 and BA.2). BA.1 now contains all the genomes with the mutational constellation that was designated as Omicron (https://github.com/cov-lineages/constellations/blob/main/constellations/definitions/cBA.1.json).

Recombination testing

To test for the possibility that the Omicron lineage is a recombinant of other SARS-CoV-2 lineages, we used a global subsample of sequences spanning January 2021 to August 2021. Using the NCBI SARS-CoV-2 Data hub[60,61], we constructed a dataset containing 221 sequences by randomly sampling five sequences from each month for each continent. No Oceania samples were available from July or August, and no South American sequences were available from July 2021[62]. These sequences were aligned together with a set of five high quality BA.1 and seven BA.2 sequences (representing the known diversity of these clades on 5 December 2021) using MAFFT[63] with default settings. Whereas 3SEQ[37], and RDP5[38] were used to analyse this dataset, a subsample

of the 39 most divergent sequences from the dataset was analysed using the GARD recombination detection method[36]. Default program settings were used throughout for recombination analyses, with the exception of RDP5 analysis, in which sequences were treated as linear and the window sizes for the SiScan and BootScan methods (two of the seven recombination detection methods applied in RDP5) were changed to 2000 nucleotides.

Selection analyses

We investigated the nature and extent of selective forces acting on BA.1 genes encoding individual protein products (a median of 25 unique BA.1 sequences per protein product encoding genome region). A subset of publicly available sequences (from the Virus Pathogen Database and Analysis Resource (ViPR) (https://www.viprbrc.org/) were included as background sequences to contextualize selection signals detectable within the BA.1 lineage at the levels of complete protein product encoding regions, and individual codons (a median of 106 sequences per protein coding region). Sequences were selected quality checked, aligned and subjected to BUSTED, RELAX, MEME, FADE, FEL, and BGM selection analyses (all implemented in HyPhy v2.5.31[64]) using the automated RASCL pipeline as outlined previously [2,8,34].

Structure modeling

We modelled the spike protein on the basis of the Protein Data Bank coordinate set 7A94, showing the first step of the spike protein trimer activation with one RBD domain in the up position, bound to the human ACE2 receptor[65]. We used the Pymol program (The PyMOL Molecular Graphics System, version 2.2.0) for visualization.

Phylogenetic analysis

All sequences on GISAID[16,17] designated Omicron (n=686; date of access: 7 December 2021) were analyzed against a globally representative reference set of SARS-CoV-2 genotypes (n=12 609) spanning the entire genetic diversity observed since the start of the pandemic. In short, the reference set included: 1. All genomes from Africa assigned to PANGO lineage B.1.1 or any of its descendents, excluding those belonging to a VOC clade; 2. A representative subsampling of global data from the publicly maintained global build of Nexstrain (https://nextstrain.org/ncov/gisaid/global); 3. The top thirty BLAST hits when querying GISAID BLAST for BA.1 and BA.2 sequences. This sampling scheme ensures that we analyze Omicron against the closest variants of the virus. Omicron and reference sequences were aligned with Nextalign[66]. A maximum-likelihood (ML) tree topology was inferred in FastTree[67] under the following parameters: a General Time Reversible (GTR) model of nucleotide substitution and a total of 100 bootstrap replicates[68]. The resulting ML-tree topology was transformed into a time-calibrated phylogeny where branches along the tree are scaled in calendar time using TreeTime[69]. The resulting tree was then visualized and annotated in ggtree in R[70].

Time-calibrated BEAST analysis

To estimate a time-scale and growth rate from the genome sequence data, BEAST v1.10.4[71] was used to sample phylogenetic trees under an exponential growth coalescent model using a strict molecular clock. All BA.1 assigned genomes from South Africa and Botswana (as of 11 December 2021) were included with some lower coverage genomes removed leaving a total of 553 genomes. The single South African BA.2 (CERI-KRISP-K032307, EPI ISL 6795834) was included to help stabilize the root of the BA.1 clade but the exponential growth coalescent model was only applied

to BA.1 (a constant population size coalescent was used for the rest of the tree). The rate of molecular evolution was estimated from the data. Two runs of 100 million iterations were compared to assess convergence and then post-burnin samples pooled to summarize parameter estimates.

Phylogeographic analysis

Markov Chain Monte Carlo (MCMC) analyses were run in duplicate in BEAST v1.10.4[72] for a total of 100 million iterations sampling every 10,000 steps in the chain. Convergence of runs was assessed in Tracer v1.7.1 [73] based on high effective sample sizes (>200) and good mixing in the chains. Maximum clade credibility trees for each run were summarized in TreeAnnotator after discarding the first 10% of the chain as burn in. Finally, the spatiotemporal dispersal of Omicron was mapped using the R package "seraphim"[74].

Estimating transmission advantage

We analyzed 805 SARS-CoV-2 sequences from Gauteng, South Africa, that were uploaded to GISAID with sample collection dates from 1 September - 1 December 2021[16]. We used a multinomial logistic regression model to estimate the growth advantage of Omicron compared to Delta at the time point where the proportion of Omicron reached 50%[75,76]. We fitted the model using the *multinom* function of the *nnet* package and estimated the growth advantage using the package *emmeans* in R.

The difference in the net growth rates (i.e., the growth advantage) between a variant (Omicron) and the wild-type (Delta) can be expressed as follows[77]:

$$\rho = (1 + \tau)\beta(S + \epsilon(1 - S)) - \beta S,$$

where τ is the increase of the intrinsic transmissibility, ϵ is the level of immune evasion, β is the transmission rate of the wild-type, and S is the proportion of the population that is susceptible to the wild-type. This relation can be algebraically solved for τ and ϵ. We further define $R_w = \beta SD$ as the effective reproduction number of the wild-type with D being the generation time. $\Omega = 1 - S$ corresponds to the proportion of the population with protective immunity against infection and subsequent transmission with the wild-type.

We estimated ϵ for different levels of τ and Ω. To propagate the uncertainty, we constructed 95% credible intervals (CIs) of the estimates from 10,000 parameter samples of ρ, D, and R_w. We assumed D to be normally distributed with a mean of 5.2 days and a standard deviation of 0.8 days[78]. We sampled from publicly available estimates of the daily R_w based on confirmed cases during the early growth phase of Omicron in South Africa (1 October - 31 October 2021; range: 0.78-0.85 (https://github.com/covid-19-Re)[48].

Data availability

All SARS-CoV-2 whole genome sequences produced by NGS-SA are deposited in the GISAID sequence database and are publicly available subject to the terms and conditions of the GISAID database. The GISAID accession numbers of sequences used in the phylogenetic analysis, including Omicron and global references, are provided in the **Appendix C - Supplementary Table S3.**

Code availability
All input files (e.g. alignments or XML files), along with all resulting output files and scripts used in the present study will be made available upon request and publicly shared on GitHub at final publication.

Acknowledgements
We thank Linda de Gouveia, Amelia Buys, Carida Fourie, Noluthando Duma, Malusi Ndlovu and other members of the NICD Centre for Respiratory Diseases and Meningitis and Sequencing Core Facility. We thank Nevashan Govender, Genevie Ntshoe, Andronica Moipone Shonhiwa, Darren Muganhiri, Itumeleng Matiea, Eva Mathatha, Fhatuwani Gavhi, Teresa Mashudu Lamola, Matimba Makhubele, Mmaborwa Matjokotja, Simbulele Mdleleni, Masingita Makhubela from the national SARS-CoV-2 NICD surveillance team for NMCSS case data, and Fazil Mckenna, Trevor Graham Bell, Ndivhuwo Munava, Stanford Kwenda, Muzammil Raza Bano and Jimmy Khosa from NICD IT for NMCSS case and test data (in particular, SGTF data). Equally, we thank the global laboratories that generated and made public the SARS-CoV-2 sequences (through GISAID) used as reference dataset in this study (a complete list of individual contributors of sequences is provided in Supplementary Table S3). The research reported in this publication was supported by the Strategic Health Innovation Partnerships Unit of the South African Medical Research Council, with funds received from the South African Department of Science and Innovation. CA received funding from the European Union's Horizon 2020 research and innovation programme - project EpiPose (No 101003688). DPM was funded by the Wellcome Trust (222574/Z/21/Z).

The genomic sequencing in Botswana was supported by the Foundation for Innovative New Diagnostics and Fogarty International Center (5D43TW009610), NIH (5K24AI131924-04; 5K24AI131928-05), as well in kind support from the Botswana government through the Ministry of Health & Wellness. SM was supported in part by the Bill & Melinda Gates Foundation [036530]. Under the grant conditions of the Foundation, a Creative Commons Attribution 4.0 Generic License has already been assigned to the Author Accepted Manuscript version that might arise from this submission

References

1. Tegally, H. *et al.* Sixteen novel lineages of SARS-CoV-2 in South Africa. *Nat. Med.* 27, 440–446 (2021).
2. Tegally, H. *et al.* Detection of a SARS-CoV-2 variant of concern in South Africa. *Nature* 592, 438–443 (2021).
3. Tegally, H. *et al.* Rapid replacement of the Beta variant by the Delta variant in South Africa. *medRxiv* (2021) doi:10.1101/2021.09.23.21264018.
4. Martin, D. P. *et al.* Selection analysis identifies significant mutational changes in Omicron that are likely to influence both antibody neutralization and Spike function (part 1 of 2). https://virological.org/t/selection-analysis-identifies-significant-mutational-changes-in-omicron-that-are-likely-to-influence-both-antibody-neutralization-and-spike-function-part-1-of-2/771 (2021).

5. Faria, N. R. *et al.* Genomics and epidemiology of the P.1 SARS-CoV-2 lineage in Manaus, Brazil. *Science* 372, 815–821 (2021).
6. Dhar, M. S. *et al.* Genomic characterization and epidemiology of an emerging SARS-CoV-2 variant in Delhi, India. *Science* 374, 995–999 (2021).
7. New AY lineages – Pango Network. https://www.pango.network/new-ay-lineages/.
8. Wilkinson, E. *et al.* A year of genomic surveillance reveals how the SARS-CoV-2 pandemic unfolded in Africa. *Science* 374, 423–431 (2021).
9. Scheepers, C. *et al.* The continuous evolution of SARS-CoV-2 in South Africa: a new lineage with rapid accumulation of mutations of concern and global detection. *medRxiv* (2021) doi:10.1101/2021.08.20.21262342.
10. Kleynhans, J. *et al.* SARS-CoV-2 Seroprevalence in a Rural and Urban Household Cohort during First and Second Waves of Infections, South Africa, July 2020-March 2021. *Emerging Infect. Dis.* 27, 3020–3029 (2021).
11. Vermeulen, M. *et al.* Prevalence of anti-SARS-CoV-2 antibodies among blood donors in South Africa during the period January-May 2021. *Res. Sq.* (2021) doi:10.21203/rs.3.rs-690372/v1.
12. Mandolo, J. *et al.* Dynamics of SARS-CoV-2 exposure in Malawian blood donors: a retrospective seroprevalence analysis between January 2020 and February 2021. *medRxiv* (2021) doi:10.1101/2021.08.18.21262207.
13. Fryatt, A. *et al.* Community SARS-CoV-2 seroprevalence before and after the second wave of SARS-CoV-2 infection in Harare, Zimbabwe. *EClinicalMedicine* 41, 101172 (2021).
14. South African COVID-19 Modelling Consortium. *COVID-19 modelling update: Considerations for a potential fourth wave*. 20 https://www.nicd.ac.za/wp-content/uploads/2021/11/SACMC-Fourth-wave-report-17112021-final.pdf (2021).
15. Volz, E. *et al.* Assessing transmissibility of SARS-CoV-2 lineage B.1.1.7 in England. *Nature* (2021) doi:10.1038/s41586-021-03470-x.
16. Shu, Y. & McCauley, J. GISAID: Global initiative on sharing all influenza data - from vision to reality. *Euro Surveill.* 22, 30494 (2017).
17. Elbe, S. & Buckland-Merrett, G. Data, disease and diplomacy: GISAID's innovative contribution to global health. *Global Challenges* 1, 33–46 (2017).
18. Rambaut, A. *et al.* A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nat. Microbiol.* 5, 1403–1407 (2020).
19. World Health Organization. Classification of Omicron (B.1.1.529): SARS-CoV-2 Variant of Concern. https://www.who.int/news-room/statements/26-11-2021-classification-of-omicron-(b.1.1.529)-sars-cov-2-variant-of-concern (2021).
20. Starr, T. N. *et al.* Deep Mutational Scanning of SARS-CoV-2 Receptor Binding Domain Reveals Constraints on Folding and ACE2 Binding. *Cell* 182, 1295-1310.e20 (2020).
21. Greaney, A. J. *et al.* Mapping mutations to the SARS-CoV-2 RBD that escape binding by different classes of antibodies. *Nat. Commun.* 12, 4196 (2021).
22. Greaney, A. J. *et al.* Complete Mapping of Mutations to the SARS-CoV-2 Spike Receptor-Binding Domain that Escape Antibody Recognition. *Cell Host Microbe* 29, 44-57.e9 (2021).
23. Greaney, A. J. *et al.* Comprehensive mapping of mutations in the SARS-CoV-2 receptor-binding domain that affect recognition by polyclonal human plasma antibodies. *Cell Host Microbe* 29, 463-476.e6 (2021).

24. McCallum, M. *et al.* N-terminal domain antigenic mapping reveals a site of vulnerability for SARS-CoV-2. *Cell* 184, 2332-2347.e16 (2021).
25. Cele, S. *et al.* SARS-CoV-2 Omicron has extensive but incomplete escape of Pfizer BNT162b2 elicited neutralization and requires ACE2 for infection. *medRxiv* (2021) doi:10.1101/2021.12.08.21267417.
26. Rössler, A., Riepler, L., Bante, D., Laer, D. von & Kimpel, J. SARS-CoV-2 B.1.1.529 variant (Omicron) evades neutralization by sera from vaccinated and convalescent individuals. *medRxiv* (2021) doi:10.1101/2021.12.08.21267491.
27. Starr, T. N., Greaney, A. J., Dingens, A. S. & Bloom, J. D. Complete map of SARS-CoV-2 RBD mutations that escape the monoclonal antibody LY-CoV555 and its cocktail with LY-CoV016. *Cell Rep. Med.* 2, 100255 (2021).
28. Starr, T. N. *et al.* Prospective mapping of viral mutations that escape antibodies used to treat COVID-19. *Science* 371, 850–854 (2021).
29. Cao, Y. R. *et al.* B.1.1.529 escapes the majority of SARS-CoV-2 neutralizing antibodies of diverse epitopes. *BioRxiv* (2021) doi:10.1101/2021.12.07.470392.
30. Bernasconi, A. *et al.* Report on Omicron Spike mutations on epitopes and immunological/epidemiological/kinetics effects from literature - SARS-CoV-2 coronavirus / nCoV-2019 Genomic Epidemiology - Virological. https://virological.org/t/report-on-omicron-spike-mutations-on-epitopes-and-immunological-epidemiological-kinetics-effects-from-literature/770 (2021).
31. Brown, J. C. *et al.* Increased transmission of SARS-CoV-2 lineage B.1.1.7 (VOC 2020212/01) is not accounted for by a replicative advantage in primary airway cells or antibody escape. *BioRxiv* (2021) doi:10.1101/2021.02.24.432576.
32. Saito, A. *et al.* SARS-CoV-2 spike P681R mutation enhances and accelerates viral fusion. *BioRxiv* (2021) doi:10.1101/2021.06.17.448820.
33. Mlcochova, P. *et al.* SARS-CoV-2 B.1.617.2 Delta variant replication and immune evasion. *Nature* 599, 114–119 (2021).
34. Martin, D. P. *et al.* The emergence and ongoing convergent evolution of the SARS-CoV-2 N501Y lineages. *Cell* 184, 5189-5200.e7 (2021).
35. Wu, H. *et al.* Nucleocapsid mutations R203K/G204R increase the infectivity, fitness, and virulence of SARS-CoV-2. *Cell Host Microbe* (2021) doi:10.1016/j.chom.2021.11.005.
36. Kosakovsky Pond, S. L., Posada, D., Gravenor, M. B., Woelk, C. H. & Frost, S. D. W. GARD: a genetic algorithm for recombination detection. *Bioinformatics* 22, 3096–3098 (2006).
37. Lam, H. M., Ratmann, O. & Boni, M. F. Improved algorithmic complexity for the 3SEQ recombination detection algorithm. *Mol. Biol. Evol.* 35, 247–251 (2018).
38. Martin, D. P. *et al.* RDP5: a computer program for analyzing recombination in, and removing signals of recombination from, nucleotide sequence datasets. *Virus Evol.* 7, veaa087 (2021).
39. van der Walt, E. *et al.* Rapid host adaptation by extensive recombination. *J. Gen. Virol.* 90, 734–746 (2009).
40. Wisotsky, S. R., Kosakovsky Pond, S. L., Shank, S. D. & Muse, S. V. Synonymous Site-to-Site Substitution Rate Variation Dramatically Inflates False Positive Rates of Selection Analyses: Ignore at Your Own Peril. *Mol. Biol. Evol.* 37, 2430–2439 (2020).

41. Smith, M. D. *et al.* Less is more: an adaptive branch-site random effects model for efficient detection of episodic diversifying selection. *Mol. Biol. Evol.* 32, 1342–1353 (2015).
42. Wertheim, J. O., Murrell, B., Smith, M. D., Kosakovsky Pond, S. L. & Scheffler, K. RELAX: detecting relaxed selection in a phylogenetic framework. *Mol. Biol. Evol.* 32, 820–832 (2015).
43. Murrell, B. *et al.* Detecting individual sites subject to episodic diversifying selection. *PLoS Genet.* 8, e1002764 (2012).
44. Kosakovsky Pond, S. L., Poon, A. F. Y., Leigh Brown, A. J. & Frost, S. D. W. A maximum likelihood method for detecting directional evolution in protein sequences and its application to influenza A virus. *Mol. Biol. Evol.* 25, 1809–1824 (2008).
45. Pulliam, J. R. C. *et al.* SARS-CoV-2 reinfection trends in South Africa: analysis of routine surveillance data. *medRxiv* (2021) doi:10.1101/2021.11.11.21266068.
46. Marivate, V. & Combrink, H. M. Use of Available Data To Inform The COVID-19 Outbreak in South Africa: A Case Study. *Data Sci. J.* 19, (2020).
47. Marivate, V. *et al.* Coronavirus disease (COVID-19) case data - South Africa. *Zenodo* (2020) doi:10.5281/zenodo.3819126.
48. Huisman, J. S. *et al.* Estimation and worldwide monitoring of the effective reproductive number of SARS-CoV-2. *medRxiv* (2020) doi:10.1101/2020.11.26.20239368.
49. National Institute for Communicable Diseases. WEEKLY TESTING SUMMARY - NICD. https://www.nicd.ac.za/diseases-a-z-index/disease-index-covid-19/surveillance-reports/weekly-testing-summary/.
50. Wickham, H. ggplot2. *WIREs Comp Stat* 3, 180–185 (2011).
51. Msomi, N., Mlisana, K., de Oliveira, T. & Network for Genomic Surveillance in South Africa writing group. A genomics network established to respond rapidly to public health threats in South Africa. *Lancet Microbe* 1, e229–e230 (2020).
52. SARS-CoV2 genome sequencing protocol (1200bp amplicon "midnight" primer set, using Nanopore Rapid kit). https://dx.doi.org/10.17504/protocols.io.bwyppfvn.
53. Cleemput, S. *et al.* Genome Detective Coronavirus Typing Tool for rapid identification and characterization of novel coronavirus genomes. *Bioinformatics* 36, 3552–3555 (2020).
54. GitHub - epi2me-labs/wf-artic: ARTIC SARS-CoV-2 workflow and reporting. https://github.com/epi2me-labs/wf-artic#readme.
55. Bragg, L. M., Stone, G., Butler, M. K., Hugenholtz, P. & Tyson, G. W. Shining a light on dark sequencing: characterising errors in Ion Torrent PGM data. *PLoS Comput. Biol.* 9, e1003031 (2013).
56. Robinson, J. T. *et al.* Integrative genomics viewer. *Nat. Biotechnol.* 29, 24–26 (2011).
57. Wilm, A. *et al.* LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. *Nucleic Acids Res.* 40, 11189–11201 (2012).
58. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly (Austin)* 6, 80–92 (2012).
59. Cingolani, P. *et al.* Using Drosophila melanogaster as a Model for Genotoxic Chemical Mutational Studies with a New Program, SnpSift. *Front. Genet.* 3, 35 (2012).

60. Hatcher, E. L. *et al.* Virus Variation Resource - improved response to emergent viral outbreaks. *Nucleic Acids Res.* 45, D482–D490 (2017).
61. National Library of Medicine. NCBI Virus: SARS-CoV-2 Data Hub. https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/virus?SeqType_s=Nucleotide&VirusLineage ss=SARS-CoV-2,%20taxid:2697049.
62. covid19-omicron-origins-recombination/aligned 234.shortnames.afa at main · bonilab/covid19-omicron-origins-recombination · GitHub. https://github.com/bonilab/covid19-omicron-origins-recombination/blob/main/4%20GS5%20plus%20Canada%20Outlier%20Lineage/4.2%20aligned_mafft_addfrag_wref/aligned_234.shortnames.afa.
63. Katoh, K., Misawa, K., Kuma, K. & Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* 30, 3059–3066 (2002).
64. Kosakovsky Pond, S. L. *et al.* HyPhy 2.5-A Customizable Platform for Evolutionary Hypothesis Testing Using Phylogenies. *Mol. Biol. Evol.* 37, 295–299 (2020).
65. Benton, D. J. *et al.* Receptor binding and priming of the spike protein of SARS-CoV-2 for membrane fusion. *Nature* 588, 327–330 (2020).
66. Aksamentov, I., Roemer, C., Hodcroft, E. & Neher, R. Nextclade: clade assignment, mutation calling and quality control for viral genomes. *JOSS* 6, 3773 (2021).

Chapter 5: Emergence of SARS-CoV-2 Omicron lineages BA.4 and BA.5 in South Africa

This chapter builds on research work developed in previous chapters and describes the characterization of the Omicron BA.4 and BA.5 sublineages following the continuous genomic monitoring for SARS-CoV-2 in South Africa. The aim of this study was to characterise yet another instance of shifting epidemiology during the pandemic infection cycles in South Africa. This time, the first noticeable changing epidemic parameter was the increase in the proportion of diagnostic tests demonstrating S-gene target failure (SGTF). The results show that the spike proteins of BA.4 and BA.5 are identical, and similar to BA.2 except for the addition of 69–70 deletion (which causes SGTF diagnostic test), L452R (present in the Delta variant), F486V and the wild-type amino acid at Q493. The BA.4 and BA.5 lineages replaced the dominant BA.2 at that time and went on to cause a fifth wave of infections in South Africa. As first author on this study, my role included genomic data generation for the sequences analysed, phylogenetic analysis to characterise them into the two novel lineages, mapping of genomic prevalence estimates onto epidemiological parameters, data visualisation and manuscript writing. The Omicron BA.4 and BA.5 lineages originally described in this study ultimately expanded globally, causing further increases in incidence in many countries. The global dominance of this lineage also meant that the second generation of COVID-19 mRNA vaccines to be licensed were bivalent shots containing spike mRNA from the original strain and BA.4 and BA.5 lineages.

This chapter was published as a peer-reviewed research article in Nature Medicine in June 2022 and can be accessed at the following DOI: 10.1038/s41591-022-01911-2. The chapter is presented in a similar format as the journal article. A full list of authors and affiliations is shown below.

Houriiyah Tegally[1,2], Monika Moir[1], Josie Everatt[3], Marta Giovanetti[4,5,6], Cathrine Scheepers[3,7], Eduan Wilkinson[1], Kathleen Subramoney[8,31], Zinhle Makatini[8,31], Sikhulile Moyo[9,10,11], Daniel G. Amoako[3], Cheryl Baxter[1], Christian L., Althaus[12], Ugochukwu J. Anyaneji[2], Dikeledi Kekana[3], Raquel Viana[13], Jennifer Giandhari[2], Richard J. Lessells[2], Tongai Maponga[14], Dorcas Maruapula[9], Wonderful Choga[9], Mogomotsi Matshaba[11], Mpaphi B. Mbulawa[15], Nokukhanya Msomi[16], NGS-SA consortium[§], Yeshnee Naidoo[1], Sureshnee Pillay[2], Tomasz Janusz Sanko[1], James E. San[2], Lesley Scott[17], Lavanya Singh[2], Nonkululeko A. Magini[2], Pamela Smith-Lawrence[18], Wendy Stevens[17,19], Graeme Dor[19], Derek Tshiabuila[2], Nicole Wolter[3,31], Wolfgang Preiser[14], Florette K. Treurnicht[8,31], Marietjie Venter[20], Georginah Chiloane[20], Caitlyn McIntyre[20], Aine O'Toole[21], Christopher Ruis[22], Thomas P. Peacock[23], Cornelius Roemer[24], Sergei L Kosakovsky Pond[25], Carolyn Williamson[26,27,28,29], Oliver G. Pybus[30], Jinal N. Bhiman[3,7], Allison Glass[13,31], Darren P. Martin[28,29], Ben Jackson[21], Andrew Rambaut[21], Oluwakemi Laguda-Akingba[32,33], Simani Gaseitsiwe[9,10], Anne von Gottberg[3,31,34] & Tulio de Oliveira[1,2,35*]

[1]Centre for Epidemic Response and Innovation (CERI), School of Data Science and Computational Thinking, Stellenbosch University, Stellenbosch, South Africa
[2]KwaZulu-Natal Research Innovation and Sequencing Platform (KRISP), Nelson R. Mandela School of Medicine, University of KwaZulu-Natal, Durban, South Africa
[3]National Institute for Communicable Diseases (NICD) of the National Health Laboratory Service (NHLS), Johannesburg, South Africa
[4]Laboratorio de Flavivirus, Fundacao Oswaldo Cruz, Rio de Janeiro, Brazil
[5]Department of Science and Technology for Humans and the Environment, University of Campus Bio-Medico di Roma, Rome, Italy.
[6]Laboratório de Genética Celular e Molecular, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil
[7]South African Medical Research Council Antibody Immunity Research Unit, School of Pathology, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa
[8]Department of Virology, Charlotte Maxeke Johannesburg Academic Hospital, Johannesburg, South Africa
[9]Botswana Harvard AIDS Institute Partnership, Botswana Harvard HIV Reference Laboratory, Gaborone, Botswana
[10]Harvard T.H. Chan School of Public Health, Boston, MA, USA
[11]Botswana Presidential COVID-19 Taskforce, Gaborone, Botswana
[12]Institute of Social and Preventive Medicine, University of Bern, Bern, Switzerland
[13]Lancet Laboratories, Johannesburg, South Africa
[14]Division of Medical Virology, Faculty of Medicine and Health Sciences, Stellenbosch University, Tygerberg, Cape Town, South Africa
[15]National Health Laboratory, Health Services Management, Ministry of Health and Wellness, Gaborone, Botswana
[16]Discipline of Virology, School of Laboratory Medicine and Medical Sciences and National Health Laboratory Service (NHLS), University of KwaZulu-Natal, Durban, South Africa
[17]Department of Molecular Medicine and Haematology, Faculty of Health Science, School of Pathology, University of the Witwatersrand, Johannesburg, South Africa
[18]Health Services Management, Ministry of Health and Wellness, Gaborone, Botswana
[19]National Priority Program of the National Health Laboratory Service, Johannesburg, South Africa
[20]Zoonotic Arbo and Respiratory Virus Program, Centre for Viral Zoonoses, Department of Medical Virology, University of Pretoria, Pretoria, South Africa
[21]Institute of Evolutionary Biology, University of Edinburgh, Edinburgh, UK
[22]Department of Medicine, University of Cambridge, Cambridge, UK
[23]Department of Infectious Disease, Imperial College London, UK, W2 1PG
[24]Biozentrum, University of Basel
[25]Institute for Genomics and Evolutionary Medicine, Department of Biology, Temple University, Philadelphia, PA 19122, USA
[26]Division of Medical Virology, Faculty of Health Sciences, University of Cape Town, Cape Town, South Africa
[27]Division of Virology, NHLS Groote Schuur Laboratory, Cape Town, South Africa
[28]Wellcome Centre for Infectious Diseases Research in Africa (CIDRI-Africa), Cape Town, South Africa
[29]Institute of Infectious Disease and Molecular Medicine, University of Cape Town, Cape Town, South Africa
[30]Department of Zoology, University of Oxford, Oxford, UK
[31]School of Pathology, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa
[32] NHLS Port Elizabeth Laboratory, Port Elizabeth, South Africa
[33]Faculty of Health Sciences, Walter Sisulu University, Eastern Cape, South Africa
[34]Division of Medical Microbiology, Department of Pathology, Faculty of Health Sciences, University of Cape Town, Cape Town
[35]Department of Global Health, University of Washington, Seattle, WA, USA

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

Abstract

Three lineages (BA.1, BA.2 and BA.3) of the SARS-CoV-2 Omicron variant of concern predominantly drove South Africa's fourth COVID-19 wave. We have now identified two new lineages, BA.4 and BA.5, responsible for a fifth wave of infections. The spike proteins of BA.4

and BA.5 are identical, and comparable to BA.2 except for the addition of 69-70del (present in the Alpha variant and the BA.1 lineage), L452R (present in the Delta variant), F486V and the wild type amino acid at Q493.The two lineages only differ outside of the spike region. The 69-70 deletion in spike allows these lineages to be identified by the proxy marker of S-gene target failure, on the background of variants not possessing this feature . BA.4 and BA.5 have rapidly replaced BA.2, reaching more than 50% of sequenced cases in South Africa by the first week of April 2022. Using a multinomial logistic regression model, we estimate growth advantages for BA.4 and BA.5 of 0.08 (95% CI: 0.08 - 0.09) and 0.10 (95% CI: 0.09 - 0.11) per day respectively over BA.2 in South Africa. The continued discovery of genetically diverse Omicron lineages points to the hypothesis that a discrete reservoir, such as human chronic infections and/or animal hosts, is potentially contributing to further evolution and dispersal of the virus.

Main text

Within days of being discovered in South Africa and Botswana, on November 26, 2021, the Omicron variant of SARS-CoV-2 was designated as a variant of concern (VOC) by the World Health Organization[1]. Initially, Omicron was comprised of three sister lineages, BA.1, BA.2 and BA.3. BA.1 caused most of the infections in South Africa's fourth epidemic wave. However, as that wave receded in mid-January 2022, BA.2 became the dominant South African lineage. Despite being associated with a modest prolongation of the fourth wave, the displacement of BA.1 by BA.2 in South Africa was not associated with a significant resurgence in cases, hospital admissions or deaths. This pattern was not consistent worldwide, however, and in some countries BA.2 was responsible for a greater share of cases, hospitalizations and deaths during the Omicron wave[2–4].

We recently identified two new Omicron lineages that have been designated BA.4 and BA.5 by the Pango Network and pango-designation v1.3, a system of naming and classifying SARS-CoV-2 lineages (Fig. 5.1A)[5,6]. Bayesian phylogenetic methods revealed that BA.4 and BA.5 are distinct from the other Omicron lineages (Molecular clock signal: correlation coefficient = 0.6, $R^2 = 0.4$, Appendix D - Extended Data Fig. 1). BA.4 and BA.5 are estimated to have originated in mid-December 2021 (95% highest posterior density [HPD] 25 November 2021 to 01 January 2022) and early January 2022 (HPD 10 December 2021 to 6 February 2022) respectively (Fig. 5.1A). The most recent common ancestor of BA.4 and BA.5 is estimated to have originated in mid-November 2021 (HPD 29 September 2021 to 6 December 2021) (Fig. 5.1A), coinciding with the emergence of the other lineages, for example BA.2 in early November 2021 (HPD: 9 October 2021 to 29 November 2021). Phylogeographic analysis suggests early dispersal of BA.4 from Limpopo to Gauteng, with later spread to other provinces (Fig. 5.1B); and early dispersal of BA.5 from Gauteng to KwaZulu-Natal, with more limited onward spread to other provinces (Fig. 5.1C).

BA.4 and BA.5 have identical spike proteins, most comparable to BA.2. Relative to BA.2, BA.4 and BA.5 have the additional spike mutations 69-70del, L452R, F486V and wild type amino acid at position Q493 (Fig 5.1D). Outside of spike, BA.4 has additional mutations at ORF7b:L11F and N:P151S and a triple amino acid deletion in NSP1:141-143del whilst BA.5 has the M:D3N mutation. Relative to BA.2, BA.5 has additional reversions at ORF6:D61 and nucleotide positions 26858 and 27259. In addition, BA.4 and BA.5 have a nuc:G12160A synonymous mutation in NSP8 that was present in Epsilon (B.1.429) and has arisen in BA.2 in some locations (Appendix D - Extended Data Fig. 2). BA.4 and BA.5 have identical mutational patterns in the 5' genome

region (from ORF1ab to Envelope) yet exhibit genetic divergence in the 3' region (from M to the 3' genome end). This suggests that BA.4 and BA.5 may be related by a recombination event, with breakpoint between the E and M genes, prior to their emergence into the general population. This scenario is somewhat similar to the relationship between BA.3 and BA.1/BA.2 which also exhibit apparent ancestral recombination[1]. Using the RASCL pipeline[7] (which employs a battery of tests that analyse ratios of synonymous and non-synonymous substitutions both at individual codon sites and I entire protein coding regions) we found no compelling evidence of imbalances between ratios of synonymous and non-synonymous substitutions such as would be indicative of positive selection (i.e. favouring amino acid changes) or negative selection (disfavouring amino acid changes) acting on any of the genes of viruses in either the BA.4 or BA.5 lineages.

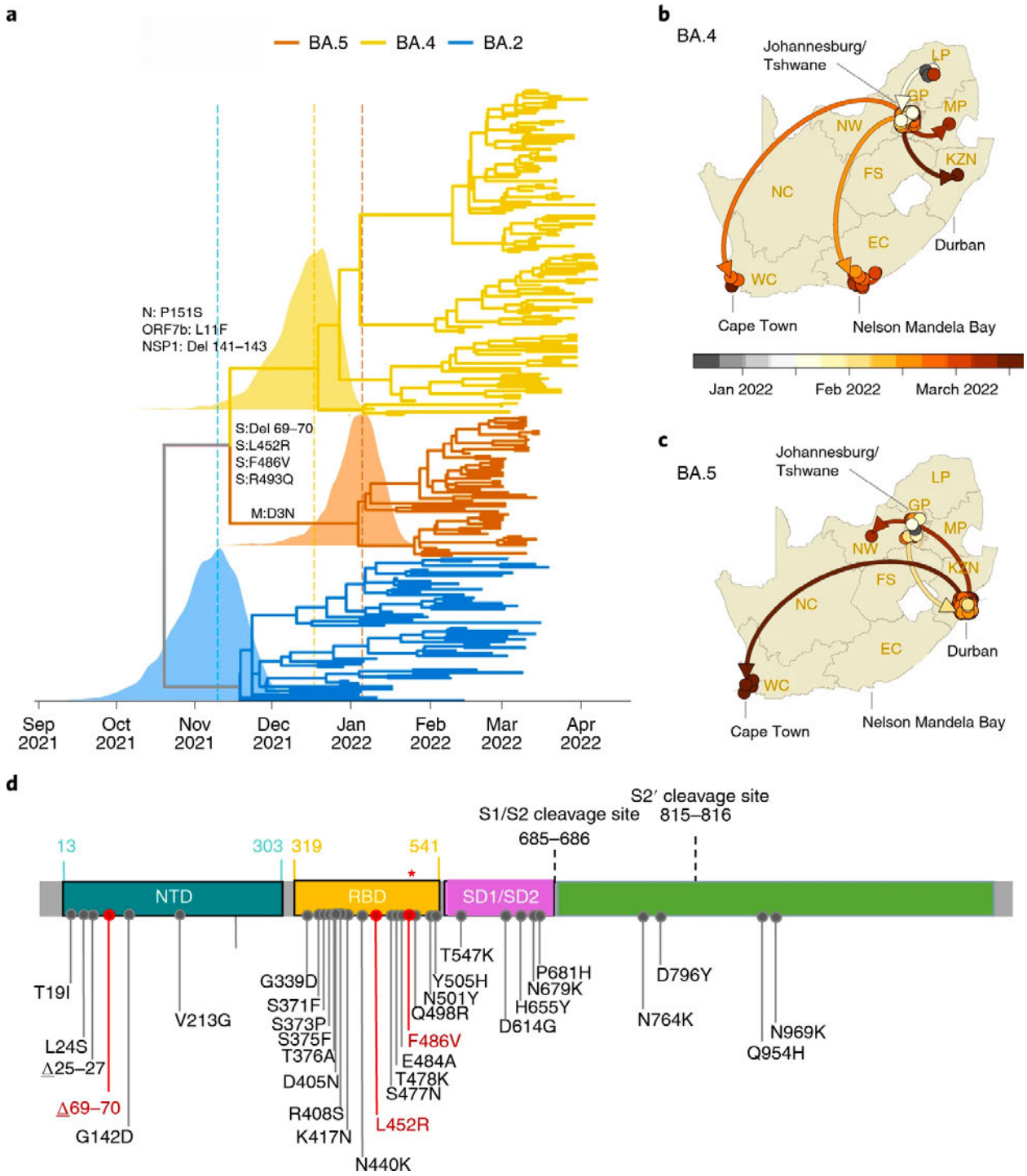


Figure 5.1: A) Time-resolved maximum clade credibility phylogeny of the BA.2, BA.4 and BA.5 lineages (n = 221, sampled between 29 December 2021 and 7 April 2022). Mutations that characterize the lineages are indicated on the branch at which each first emerged. The posterior distribution of the time of the most recent common ancestor (TMRCA) is also shown for BA.2, BA.4 and BA.5. B) Spatiotemporal reconstruction of the spread of the BA.4 lineage in South Africa. C) Spatiotemporal reconstruction of the spread of the BA.5 lineage in South Africa. In B and C, circles represent nodes of the maximum clade credibility phylogeny, coloured according to their inferred time of occurrence (scale shown). EC, Eastern Cape; FS, Free State; GP, Gauteng;

KZN, KwaZulu-Natal; LP, Limpopo; MP, Mpumalanga; NC, Northern Cape; NW, North West; WC, Western Cape. Solid curved lines denote the links between nodes and the directionality of movement is indicated (anti-clockwise along the curve). D) Amino acid mutations in the spike gene of the BA.4 and BA.5 lineages. Mutations that differ from BA.2 are denoted in red, including the wild-type amino acid at position Q493 (denoted by the red *).

It is currently unknown how differences in the mutation profiles of BA.4 and BA.5, relative to BA.2, will impact their phenotypes. Changes at spike amino acids 452, 486 and 493 are likely to influence human angiotensin-converting enzyme-2 (hACE2) and antibody binding. The 452 residue is in immediate proximity to the interaction interface of the hACE2 receptor. The L452R mutation has been associated with an increased affinity for receptor binding with a resultant increased in vitro infectivity[8] . The L452R mutation is also present in the Delta, Kappa and Epsilon variants (and L452Q in Lambda), and mutations at this position have been associated with a reduction in neutralization by monoclonal antibodies (particularly class 2 antibodies) and polyclonal sera[9–11]. Mutations at this position (L452R/M/Q) have also arisen independently in several BA.2 sublineages in different parts of the world, most notably BA.2.12.1 (L452Q) which has become dominant in many parts of the United States. It's therefore unclear whether BA.4/BA.5 will become dominant throughout the world, or whether there will be a period of co-circulation of several different Omicron lineages.

Before the emergence of BA.4 and BA.5, F486V in the receptor binding domain of spike had been observed only in 54 of 10 million publicly available genome sequences in GISAID (https://cov-spectrum.org/explore/World/AllSamples/AllTimes/variants?aaMutations=S%3AF486V&).
Selection analyses focusing on ratios of non-synonymous and synonymous substitution rates at individual codons have indicated that, since December 2020 S:486 has been evolving under strong negative selection favouring the F state at this site (i.e., the amino acid that is found in Wuhan-Hu-1) (Appendix D - Extended Data Fig. 3). Although rare, the F486L mutation has been observed in approximately 500 genomes, most commonly in viruses infecting minks and from human cases linked to mink farms. The F486L mutation has been shown to directly enhance entry into cells expressing mink or ferret ACE2[12]. When binding to hACE2, spike amino acid F486 interacts with hACE2 residues L79, M82, and Y83, which collectively comprise a hotspot for ACE2 differences between mammalian species[13]. Mutations at F486 are associated with a reduction in neutralising activity by class 1 (and some class 2) neutralising antibodies and by polyclonal sera[9–11]. Deep mutational scanning suggests that F486 is a key site for escape of vaccine- and infection-elicited RBD-targeted antibodies, including those still able to neutralize Omicron/BA.1 (https://jbloomlab.github.io/SARS2_RBD_Ab_escape_maps/escape-calc/)[14]. This suggests that BA.4 and BA.5 may be even better at evading neutralizing antibody responses, including those recently elicited by BA.1 infections. Combined with waning population immunity against infection from the initial Omicron/BA.1 wave, this could create the conditions for a significant resurgence in infections.

The S:69-70del means BA.4 and BA.5 can again be presumptively identified (against a background of BA.2 infection) using the proxy marker of S-gene target failure (SGTF) with the TaqPath™ COVID-19 qPCR assay (Thermo Fisher Scientific, Waltham, MA, USA). SGTF was successfully used to track the early spread of BA.1 (which also demonstrates SGTF), later also enabling discrimination between BA.1 and BA.2 infections, since BA.2 viruses generally lack the

S:69-70del[15]. Recent data from public laboratories in South Africa suggest that the proportion of positive PCR tests with SGTF has been increasing since early March, suggesting that BA.4 and BA.5 may be responsible for a growing share of recently confirmed cases (Fig. 5.2A). To assess the validity of SGTF for identifying BA.4/BA.5, we performed qPCR with the TaqPath™ assay on 296 unselected samples submitted for sequencing to KwaZulu-Natal Research Innovation and Sequencing Platform (KRISP) from Gauteng, Eastern Cape and KwaZulu-Natal collected between 6 January and 3 April 2022. Of the 296 samples processed, we had a paired valid qPCR result and sequence for 198. Of the 77 samples with SGTF on qPCR, 66 were BA.4 or BA.5, nine were BA.1, and two were BA.2. No BA.4 and BA.5 genomes were S-gene target positive on qPCR (Extended Data Table 1). These results suggest that SGTF surveillance (where the assay is available) may for now be a reasonable proxy to identify BA.4 and BA.5 for countries with a low prevalence of BA.1.

At the time of writing, we have confirmed BA.4 and/or BA.5 in all nine provinces in South Africa (Eastern Cape, Gauteng, KwaZulu-Natal, Limpopo, Mpumalanga, North West, Northern Cape, Free State and Western Cape) in samples collected between 10 January 2022 and 19 May 2022 (Fig. 5.2B). In the three most populous provinces in South Africa, Gauteng, KwaZulu-Natal and Western Cape, BA.4 and BA.5 have rapidly replaced BA.2, and are responsible for approximately 90% of sequenced cases by the week starting 16 May 2022 (Appendix D - Extended Data Fig. 4). These estimates are based on unselected sampling for genomic surveillance (samples not selected based on SGTF or genotyping). The data suggest geographic heterogeneity in the distribution of these two new lineages, with growth predominantly of BA.4 in Gauteng and BA.5 in KwaZulu-Natal (Appendix D - Extended Data Fig. 4). Internationally, by the end of May 2022, BA.4 and BA.5 had also been detected and was rising in prevalence in several countries: in neighbouring Botswana (estimated prevalence 60%), in Europe (Portugal, Spain, Austria), and in the USA.

We estimated that Omicron BA.4 and BA.5 had a daily growth advantage of 0.08 (95% confidence interval [CI] 0.08 - 0.09) and 0.10 (95% CI 0.09 - 0.11) respectively, relative to BA.2 in South Africa in May 2022 (Fig. 5.2F). These estimates are similar to the estimated daily growth advantage of 0.07 (95% CI: 0.07 – 0.06) of BA.2 over BA.1 in February 2022 (Fig. 5.2C). The BA.4 and BA.5 lineages also show a growth advantage against non-Omicron lineages although these are minimally circulating in the discussed time frame (Extended Data Table 2). The growth advantage of Omicron BA.4 and BA.5 could be mediated by either (i) an increase in its intrinsic transmissibility relative to other variants, (ii) an increase relative to other variants in its capacity to infect, and be transmitted from, previously infected and vaccinated individuals or (iii) both. The estimated time to most recent common ancestor for both BA.4 and BA.5 (mid-November 2021, similar to that for BA.1 and BA.2) argues against the first option because that suggests both lineages would have been circulating throughout the period dominated by BA.1 and then BA.2 without exhibiting a transmission advantage. The observation that both BA.4 and BA.5 (and many lineages within them) have recently started to grow in frequency suggests the growth advantage is recent and uniform across these lineages. It is estimated that almost all of the South African population has some degree of immunity to SARS-CoV-2, provided by a complex mixture of vaccination and prior infections with wild-type, Beta, Delta, and Omicron (particularly BA.1) (Fig 5.2D)[16,17]. Given that the transmission advantage becomes apparent approximately four months from the start of the Omicron wave, it is plausible that waning immunity (particularly that acquired from BA.1 infection) is an important contributory factor. This would also suggest that the effects

of these different Omicron lineages may differ by location, depending on the immune landscape and particularly the patterns of exposure to BA.1 and BA.2.

At the time of writing, a wave of infections caused by the BA.4 and BA.5 lineages was ending in South Africa (Fig. 5.2D). This wave was characterized by a peak in test positivity rate of ~24%, lower than during the Omicron BA.1 wave (~34%), and, because of high population immunity, much lower hospital admissions and deaths than previously recorded during waves of infection in South Africa. It is worthy to note that recorded death metrics were further decoupled from cases and hospitalizations compared to the BA.1 wave. The ability of the BA.4 and BA.5 lineages to drive a new wave of infections can potentially be explained by their ability to evade immunity induced by the BA.1 lineage roughly three months after infection[18]. The fifth wave in South Africa, driven by BA.4 and BA.5, occurred around four months after the fourth wave, driven by BA.1. At the time of writing this report, Botswana was experiencing a rapid rise in cases driven by BA.4 and BA.5, with 19 of 24 health districts experiencing resurgence in cases. To note, Botswana's fourth wave was driven by BA.1, followed by BA.2 lasting about 3.5 months and the country’s fifth wave is occurring approximately two months after the fourth wave.

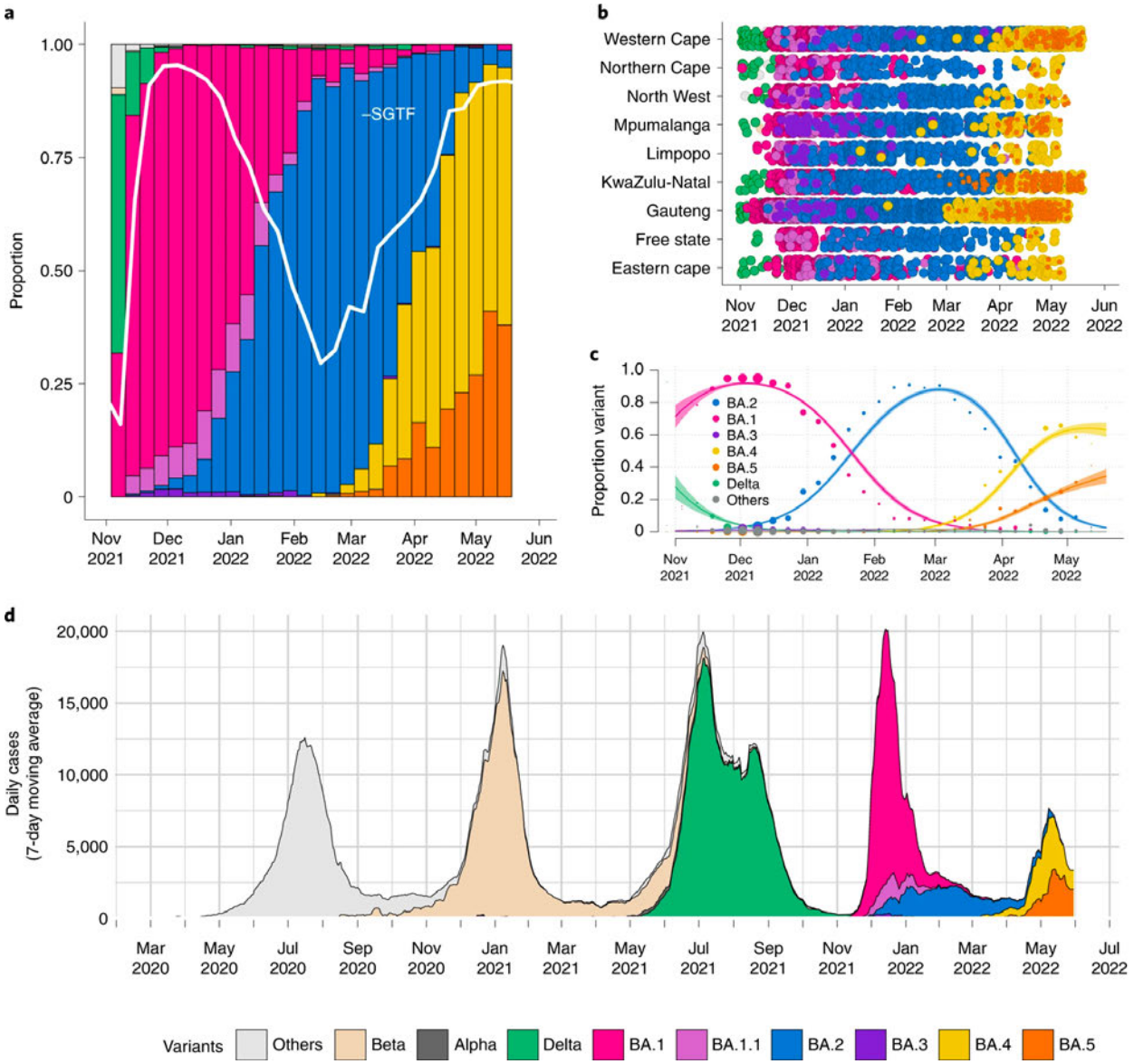


Figure 5.2: A) Changes in the genomic prevalence of Omicron lineages in South Africa from November 2021 (when BA.1 dominated) to May 2022 (when BA.4 and BA.5 were increasing in frequency), superimposed with the proportion of positive TaqPath qPCR tests exhibiting SGTF from November 2021 to May 2022. To note here that estimations of genomic prevalence and SGTF proportions are done from different samples and datasets, and only presented together here for illustrative purposes. B) The count of Omicron lineage genomes per province of South Africa over November 2021 – May 2022. BA.4 and BA.5 have been detected in all nine provinces. C) Modelled linear proportions of the Omicron lineages in South Africa. BA.1 rapidly outcompeted Delta in November 2021 and was then superseded by BA.2 in early 2022. BA.4 and BA.5 appear to be swiftly replacing BA.2 in South Africa. Model fits are based on a multinomial logistic regression and dot size represents the weekly sample size. The shaded areas correspond to the 95% CIs of the model estimates. D) The progression of the 7-day rolling average of daily reported case numbers in South Africa over two years of the epidemic (April 2020 – May 2022). Daily cases are coloured by the inferred proportion of SARS-CoV-2 variants prevalent at a particular period in the epidemic.

There are several limitations to this study. First, the estimated growth advantage of the BA.4 and BA.5 lineages could be biased due to stochastic effects (such as superspreading) in a low incidence setting at the start of a wave, which can lead to overestimates of the growth advantage. Secondly, reliable estimates of the level of population immunity against BA.1 in South Africa are not yet available, making it difficult to precisely estimate transmissibility or immune evasion of the new lineages. There also remains some uncertainty about the origin of the different Omicron lineages and phylogenetic inference is limited by the relatively low sampling coverage in our genomic surveillance (<1% of confirmed cases in South Africa). Furthermore, the lack of sampling on an ancestor of the different Omicron lineages complicates phylogenetic placements. Whilst the Bayesian phylogenetic methods employed here suggest that BA.4 and BA.5 are independent lineages that originated around the same time as BA.1-BA.3, maximum likelihood estimations suggest they could have descended from BA.2. Further sequencing (particularly samples from Gauteng and neighbouring provinces) may help to provide more clarity.

The continued discovery of genetically diverse Omicron lineages shifts the level of support for hypotheses regarding their origin, from an unsampled location to a discrete reservoir, such as human chronic infections (or even a network of chronic human infections) and/or animal reservoirs, potentially contributing to further evolution and dispersal of the virus, although currently the data does not provide any definitive evidence in any direction. We are actively investigating the potential of a yet unidentified animal reservoir in the region. To date, the only reverse zoonoses cases reported from the African region were in African lions and a puma in a private zoo in Johannesburg, South Africa[19]. Although these are unlikely species to play a role in the emergence of new variants, it is a reminder of the susceptibility of certain wildlife species to infections from humans. Following the emergence of Omicron, the World Organisation for Animal Health released a statement calling for enhanced surveillance in animals to identify the origin of new variants[20]. Further genomic sampling and evolutionary investigation will thus be required to explain the origin of Omicron lineages.

In conclusion, we have identified two new Omicron lineages (BA.4 and BA.5), which are associated with a resurgence in infections in South Africa approximately four months on from the start of the Omicron wave. This once again highlights the importance of continued global genomic surveillance and variant analysis to act as an early warning system, giving countries time to prepare and mitigate the public health impact of emerging variants.

Funding Information
This research was supported by the South African Medical Research Council (SAMRC) with funds received from the National Department of Health. 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 (CDC)(U01IP001048; 1 NU51IP000930); 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; the UK Foreign, Commonwealth and Development Office and Wellcome (221003/Z/20/Z); and the UK Department of Health and Social Care and managed by the Fleming Fund and performed under the auspices of the SEQAFRICA project. This research was also supported by The Coronavirus Aid, Relief, and Economic Security Act (CARES ACT) through the 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. Sequencing activities at KRISP and CERI are supported in part by grants from the World Health Organization, the Abbott Pandemic Defense Coalition (APDC), the National Institute of Health USA (U01 AI151698) for the United World Antivirus Research Network (UWARN) and the INFORM Africa project through IHVN (U54 TW012041), and the South African Department of Science and Innovation (SA DSI) and the SAMRC under the BRICS JAF #2020/049. Sequencing at the Botswana Harvard AIDS Institute Partnership was supported by funding from the Bill and Melinda Gates Foundation, Foundation for Innovation in Diagnostics, National Institutes of Health Fogarty International Centre (Grant 3D43TW009610-09S1), HHS/NIH/National Institute of Allergy and Infectious Diseases (NIAID) (5K24AI131928-04; 5K24AI131924-04). The content and findings reported herein are the sole deduction, view and responsibility of the researcher/s and do not reflect the official position and sentiments of the funding agencies.

Acknowledgements
We thank additional members from originating and sequencing laboratories in South Africa, listed as part of the NGS-SA consortium authors, that helped to generate and make public the SARS-CoV-2 sequences (through GISAID) used as a reference dataset in this study (a complete list of individual contributors of sequences is provided in Supplementary Table S1).

Author Contributions

Genomic or Diagnostics data generation: H.T., M.Moir, J.E., C.S., K.S., S.Moyo, D.G.A., U.J.A., D.K., R.V., J.G., T.M., D.M., W.C., M.Matshaba, S.Mayaphi, N.Mbhele, N.B.M., Y.N., S.P., T.J.S., J.E.S, L.Scott, L.Singh, N.A.M., P.S.L., W.S., G.D., D.T., N.W., W.P., F.K.T, O.L.-A.,C.W., J.N.B., N.W., A.vG.

Sample collection and metadata curation: N.Msomi, M.V., K.S., F.K.T., M.D., G.C., A.M., C.M., , N.W., A.vG., Z.M.

Data analysis: H.T., M.Moir, M.G., E.W., J.E., D.G.A., K.S., A.OT., C.R., T.P.P., C.R., O.G.P, D.P.M., A.R., S.L.K.P

Study design and data interpretation: H.T., M.Moir, E.W., C.L.A, R.J.L., C.W., O.G.P, J.B., A.G., D.P.M., B.J., A.R., S.G., J.N.B., A.vG., T.dO

Manuscript writing: H.T., M.Moir, M.G., E.W., C.B., R.J.L., T.dO

Conflict of interest

The authors declare no conflict of interest. Raquel Viana and Allison Glass are employees of Lancet Laboratories.

References:

1. Viana, R. *et al.* Rapid epidemic expansion of the SARS-CoV-2 Omicron variant in southern Africa. *Nature* **603**, 679–686 (2022).
2. Rahimi, F. & Talebi Bezmin Abadi, A. The Omicron subvariant BA.2: Birth of a new challenge during the COVID-19 pandemic. *Int. J. Surg.* **99**, 106261 (2022).
3. Fonager, J. *et al.* Molecular epidemiology of the SARS-CoV-2 variant Omicron BA.2 sub-lineage in Denmark, 29 November 2021 to 2 January 2022. *Euro Surveill.* **27**, (2022).
4. Chen, L.-L. *et al.* Contribution of low population immunity to the severe Omicron BA.2 outbreak in Hong Kong. *Res. Sq.* (2022) doi:10.21203/rs.3.rs-1512533/v1.
5. O'Toole, Á., Pybus, O. G., Abram, M. E., Kelly, E. J. & Rambaut, A. Pango lineage designation and assignment using SARS-CoV-2 spike gene nucleotide sequences. *BMC Genomics* **23**, 121 (2022).
6. Rambaut, A. *et al.* A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nat. Microbiol.* **5**, 1403–1407 (2020).
7. Lucaci, A. G. *et al.* RASCL: Rapid Assessment Of SARS-CoV-2 Clades Through Molecular Sequence Analysis. *BioRxiv* (2022) doi:10.1101/2022.01.15.476448.
8. Motozono, C. *et al.* SARS-CoV-2 spike L452R variant evades cellular immunity and increases infectivity. *Cell Host Microbe* **29**, 1124-1136.e11 (2021).
9. Greaney, A. J. *et al.* Mapping mutations to the SARS-CoV-2 RBD that escape binding by different classes of antibodies. *Nat. Commun.* **12**, 4196 (2021).
10. Greaney, A. J. *et al.* Comprehensive mapping of mutations in the SARS-CoV-2 receptor-binding domain that affect recognition by polyclonal human plasma antibodies. *Cell Host Microbe* **29**, 463-476.e6 (2021).
11. Greaney, A. J. *et al.* Complete Mapping of Mutations to the SARS-CoV-2 Spike Receptor-Binding Domain that Escape Antibody Recognition. *Cell Host Microbe* **29**, 44-57.e9 (2021).
12. Zhou, J. *et al.* Mutations that adapt SARS-CoV-2 to mink or ferret do not increase fitness in the human airway. *Cell Rep.* **38**, 110344 (2022).
13. Lan, J. *et al.* Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. *Nature* **581**, 215–220 (2020).
14. Greaney, A. J., Starr, T. N. & Bloom, J. D. An Antibody-Escape Estimator for Mutations to the SARS-CoV-2 Receptor-Binding Domain. *Virus Evol.* (2022) doi:10.1093/ve/veac021.
15. Scott, L. *et al.* Track Omicron's spread with molecular data. *Science* **374**, 1454–1455 (2021).
16. Sun, K. *et al.* Persistence of SARS-CoV-2 immunity, Omicron's footprints, and projections of epidemic resurgences in South African population cohorts. *medRxiv* (2022) doi:10.1101/2022.02.11.22270854.

17. Madhi, S. A. *et al.* Population Immunity and Covid-19 Severity with Omicron Variant in South Africa. *N. Engl. J. Med.* **386**, 1314–1326 (2022).
18. Khan, K. *et al.* Omicron sub-lineages BA.4/BA.5 escape BA.1 infection elicited neutralizing immunity. *medRxiv* (2022) doi:10.1101/2022.04.29.22274477.
19. Koeppel, K. N. *et al.* SARS-CoV-2 Reverse Zoonoses to Pumas and Lions, South Africa. *Viruses* **14**, (2022).
20. Statement from the Advisory Group on SARS-CoV-2 Evolution in Animals concerning the origins of Omicron variant - OIE - World Organisation for Animal Health. https://www.oie.int/en/document/statement-from-the-advisory-group-on-sars-cov-2-evolution-in-animals-concerning-the-origins-of-omicron-variant/.

Methods:

Epidemiological dynamics

We analysed daily cases of SARS-CoV-2 in South Africa up to 25 April 2022 from publicly released data provided by the National Department of Health and the National Institute for Communicable Diseases. This was accessible through the repository of the Data Science for Social Impact Research Group at the University of Pretoria (https://github.com/dsfsi/covid19za)[21,22]. The National Department of Health releases daily updates on the number of confirmed new cases, deaths and recoveries, with a breakdown by province.

Sampling of SARS-CoV-2

As part of the NGS-SA[23], seven sequencing hubs receive randomly selected samples for sequencing every week according to approved protocols at each site. These samples include remnant nucleic acid extracts or remnant nasopharyngeal and oropharyngeal swab samples from routine diagnostic SARS-CoV-2 PCR testing from public and private laboratories in South Africa. We analysed SARS-CoV-2 genomes generated from samples collected between 1 November 2021 and 20th April 2022.

Ethical statement

The genomic surveillance in South Africa was approved by the University of KwaZulu–Natal Biomedical Research Ethics Committee (BREC/00001510/2020), the University of the Witwatersrand Human Research Ethics Committee (HREC) (M180832), Stellenbosch University HREC (N20/04/008_COVID-19), University of Cape Town HREC (383/2020), University of Pretoria HREC (H101/17) and the University of the Free State Health Sciences Research Ethics Committee (UFS-HSD2020/1860/2710). The genomic sequencing in Botswana was conducted as part of the national vaccine roll-out plan and was approved by the Health Research and Development Committee (Health Research Ethics body, HRDC#00948 and HRDC#00904). Individual participant consent was not required for the genomic surveillance. This requirement was waived by the Research Ethics Committees.

Whole-genome sequencing and genome assembly

RNA was extracted on an automated Chemagic 360 instrument, using the CMG-1049 kit (Perkin Elmer). The RNA was stored at −80 °C before use. Libraries for whole-genome sequencing were prepared using either the Oxford Nanopore Midnight protocol with Rapid Barcoding or the Illumina COVIDseq Assay.

Illumina Miseq/NextSeq

For the Illumina COVIDseq assay, the libraries were prepared according to the manufacturer's protocol. In brief, amplicons were tagmented, followed by indexing using the Nextera UD Indexes Set A. Sequencing libraries were pooled, normalized to 4 nM and denatured with 0.2 N sodium acetate. A 8 pM sample library was spiked with 1% PhiX (PhiX Control v3 adaptor-ligated library used as a control). We sequenced libraries using the 500-cycle v2 MiSeq Reagent Kit on the Illumina MiSeq instrument (Illumina). On the Illumina NextSeq 550 instrument, sequencing was performed using the Illumina COVIDSeq protocol (Illumina), an amplicon-based next-generation sequencing approach. The first-strand synthesis was performed using random hexamers primers from Illumina and the synthesized cDNA underwent two separate multiplex PCR reactions. The pooled PCR amplified products were processed for tagmentation and adapter ligation using IDT for Illumina Nextera UD Indexes. Further enrichment and clean-up was performed according to protocols provided by the manufacturer (Illumina). Pooled samples were quantified using the Qubit 3.0 or 4.0 fluorometer (Invitrogen) and the Qubit dsDNA High Sensitivity assay kit according to the manufacturer's instructions. The fragment sizes were analysed using the TapeStation 4200 (Invitrogen). The pooled libraries were further normalized to 4 nM concentration, and 25 µl of each normalized pool containing unique index adapter sets was combined into a new tube. The final library pool was denatured and neutralized with 0.2 N sodium hydroxide and 200 mM Tris-HCl (pH 7), respectively. Sample library (1.5 pM) was spiked with 2% PhiX. Libraries were loaded onto a 300-cycle NextSeq 500/550 HighOutput Kit v2 and run on the Illumina NextSeq 550 instrument (Illumina).

Midnight protocol

For Oxford Nanopore sequencing, the Midnight primer kit was used as described previously54. cDNA synthesis was performed on the extracted RNA using the LunaScript RT mastermix (New England BioLabs) followed by gene-specific multiplex PCR using the Midnight primer pools, which produce 1,200 bp amplicons that overlap to cover the 30 kb SARS-CoV-2 genome. Amplicons from each pool were pooled and used neat for barcoding with the Oxford Nanopore Rapid Barcoding kit according to the manufacturer's protocol. Barcoded samples were pooled and bead-purified. After the bead clean-up, the library was loaded on a prepared R9.4.1 flow-cell. A GridION X5 or MinION sequencing run was initiated using MinKNOW software with the base-call setting switched off.

Ion Torrent Genexus Integrated Sequencer methodology for rapid whole-genome sequencing of SARS-CoV-2

Viral RNA was extracted using the MagNA Pure 96 DNA and Viral Nucleic Acid kit on the automated MagNA Pure 96 system (Roche Diagnostics) according to the manufacturer's instructions. Extracts were then screened by quantitative PCR to acquire the mean cycle threshold (Ct) values for the SARS-CoV-2 N and ORF1ab genes using the TaqMan 2019-nCoV assay kit v1 (Thermo Fisher Scientific) on the ViiA7 Real-time PCR system (Thermo Fisher Scientific) according to the manufacturer's instructions. Extracts were sorted into batches of n = 8 within a Ct range difference of 5 for a maximum of two batches per run. Extracts with <200 copies were sequenced using the low viral titre protocol. Next-generation sequencing was performed using the Ion AmpliSeq SARS-CoV-2 Research Panel on the Ion Torrent Genexus Integrated Sequencer

(Thermo Fisher Scientific), which combines automated cDNA synthesis, library preparation, templating preparation and sequencing within 24 h. The Ion Ampliseq SARS-CoV-2 Research Panel consists of two primer pools targeting 237 amplicons tiled across the SARS-CoV-2 genome providing >99% coverage of the SARS-CoV-2 genome (~30 kb) and an additional five primer pairs targeting human expression controls. The SARS-CoV-2 amplicons range from 125 bp to 275 bp in length. TRINITY was used for de novo assembly and the Iterative Refinement Meta-Assembler (IRMA) was used for genome assisted assembly as well as FastQC for quality checks.

Genome assembly

We assembled paired-end and Nanopore .fastq reads using Genome Detective v.1.132 (https://www.genomedetective.com), which was updated for the accurate assembly and variant calling of tiled primer amplicon Illumina or Oxford Nanopore reads, and the Coronavirus Typing Tool55. For Illumina assembly, the GATK HaploTypeCaller --min-pruning 0 argument was added to increase mutation calling sensitivity near sequencing gaps. For Nanopore, low-coverage regions with poor alignment quality (<85% variant homogeneity) near sequencing/amplicon ends were masked to be robust against primer drop-out experienced in the spike gene, and the sensitivity for detecting short inserts using a region-local global alignment of reads was increased. We also used the wf_artic (ARTIC SARS-CoV-2) pipeline as built using the Nextflow workflow framework56. In some instances, mutations were confirmed visually with .bam files using Geneious v.2020.1.2 (Biomatters). The reference genome used throughout the assembly process was NC_045512.2 (numbering equivalent to MN908947.3).

Raw reads from the Illumina COVIDSeq protocol were assembled using the Exatype NGS SARS-CoV-2 pipeline v.1.6.1 (https://sars-cov-2.exatype.com/). This pipeline performs quality control on reads and then maps the reads to a reference using Examap. The reference genome used throughout the assembly process was NC_045512.2 (accession number: MN908947.3).

Several of the initial Ion Torrent genomes contained a number of frameshifts, which caused unknown variant calls. Manual inspection revealed that these were probably sequencing errors resulting in mis-assembled regions (probably due to the known error profile of Ion Torrent sequencers). To resolve this, the raw reads from the IonTorrent platform were assembled using the SARSCoV2 RECoVERY (Reconstruction of Coronavirus Genomes & Rapid Analysis) pipeline implemented in the Galaxy instance ARIES (https://aries.iss.it). This pipeline fixed the observed frameshifts, confirming that they were artefacts of mis-assembly; this subsequently resolved the variant calls. The Exatype and RECoVERY pipelines each produce a consensus sequence for each sample. These consensus sequences were manually inspected and polished using Aliview v.1.27 (http://ormbunkar.se/aliview/).

All of the sequences passing internal quality control were deposited in GISAID (https://www.gisaid.org/), and the GISAID accession identifiers are included as part of Extended Data Table 1.

Phylogenetic analysis

We initially analysed genomes from South Africa against the global reference dataset using a custom pipeline based on a local version of NextStrain (https://github.com/nextstrain/ncov)[24]. The pipeline contains several Python scripts that manage the analysis workflow. It performs an

alignment of genomes in NextAlign[25], phylogenetic tree inference in IQ-Tree V1.6.9[26], tree dating and ancestral state construction and annotation (https://github.com/nextstrain/ncov).

The initial phylogenetic analysis enabled us to identify clusters corresponding to the BA.4 (n=120) and BA.5 (n=51) lineages. We extracted these clusters and constructed a preliminary maximum-likelihood tree with a subset of BA.2 sequences (n=52) in IQ-tree. We inspected this maximum-likelihood tree in TempEst v.1.5.3[27] for the presence of a temporal or molecular clock signal. Linear regression of root-to-tip genetic distances against sampling dates indicated that the SARS-CoV-2 sequences evolved in a relatively strong clock-like manner (correlation coefficient = 0.6, $R^2 = 0.4$).

Given that the estimation of tMRCAs and dispersal dynamics of the sampled viruses is best achieved using Bayesian phylogenetic methods, we then estimated time-calibrated phylogenies using the Bayesian software package BEAST v.1.10.4[28]. For this analysis, we used the strict molecular clock model, the HKY+I+G, nucleotide substitution model and the exponential growth coalescent model[29]. We computed Markov chain Monte Carlo (MCMC) in duplicate runs of 20 million states each, sampling every 2,000 steps. Convergence of MCMC chains was checked using Tracer v.1.7.1[30]. Maximum clade credibility trees were summarized from the MCMC samples using TreeAnnotator after discarding 10% as burn-in. The phylogenetic trees were visualized using ggplot and ggtree[31,32].

Phylogeographic analysis

To model phylogenetic diffusion of the new cluster across the country, we used a flexible relaxed random walk diffusion model that accommodates branch-specific variation in rates of dispersal with a Cauchy distribution[33]. For each sequence, latitude and longitude were attributed to the most precise district or provincial information available and linked to the diagnostic sample.

As described in 'Phylogenetic analysis', MCMC chains were run in duplicate for 10 million generations and sampled every 1,000 steps, with convergence assessed using Tracer v.1.7.1. Maximum clade credibility trees were summarized using TreeAnnotator after discarding 10% as burn-in. We used the R package seraphim[34] to extract and map spatiotemporal information embedded in posterior trees.

Lineage classification

We used a previously proposed dynamic lineage classification method[35] from the 'Phylogenetic Assignment of Named Global Outbreak Lineages' (pangolin) software suite v4.0.6 with the --Usher option (https://github.com/cov-lineages/pangolin)[36]. This is aimed at identifying the most epidemiologically important lineages of SARS-CoV-2 at the time of analysis, enabling researchers to monitor the epidemic in a particular geographic region. A lineage is a linear chain of viruses in a phylogenetic tree showing connection from the ancestor to the last descendant. Variant refers to a genetically distinct virus with different mutations to other viruses.

Selection analysis

To identify which (if any) of the observed mutations in the spike protein was most likely to increase viral fitness, we used the natural selection analysis of SARS-CoV-2 pipeline (https://observablehq.com/@spond/revised-sars-cov-2-analytics-page). This pipeline examines

the entire global SARS-CoV-2 nucleotide sequence dataset for evidence of: (i) polymorphisms having arisen in multiple epidemiologically unlinked lineages that have statistical support for non-neutral evolution (mixed effects model of evolution)[37], (ii) sites at which these polymorphisms have support for a greater-than-expected ratio of nonsynonymous-to-synonymous nucleotide substitution rates on internal branches of the phylogenetic tree (fixed-effects likelihood)[38] and (iii) whether these polymorphisms have increased in frequency in the regions of the world in which they have occurred.

Estimating transmission advantage

We analysed 12,528 SARS-CoV-2 sequences from South Africa generated in this study and uploaded to GISAID with sample collection dates from 1 November 2021 to 20 April 2022[39]. We used a multinomial logistic regression model to estimate the growth advantage of Omicron BA.2 lineage compared with BA.1, BA.4 and BA.5 lineages at the time point at which the proportion of Omicron BA.4 and BA.5 collectively reached 50%[40,41]. We fitted the model using the multinom function of the nnet package and estimated the growth advantage using the package emmeans in R[42].

S-Gene Target Failure Monitoring

SGTF monitoring is performed through analysing SARS-CoV-2 laboratory test results from nasopharyngeal specimens received from the public health sector and referred for PCR testing undertaken by the National Health Laboratory Service (NHLS) in South Africa. The NHLS has a single laboratory information system connecting laboratory testing platforms to a corporate data warehouse, where data can be mined in near real-time. The TaqPath™ COVID-19 [Thermo Fisher Scientific, Waltham, MA, USA] assay accounts for around 20% of NHLS PCR tests performed, with around half of those performed in Gauteng. The TaqPath assay targets three gene regions, ORF1ab, N and S, with the lack of probe fluorescence of the latter culminating in S-gene target failure (SGTF). In Fig 5.2A, we analysed and plotted the weekly proportion of positive TaqPath tests with SGTF (defined as samples with non-detectable S gene target and either N or ORF1ab gene positive with CT value <30.

Validation of S-Gene Target status as proxy for BA.4 and BA.5

Using a subset of unselected samples submitted to the KRISP sequencing laboratory, we compared the S-gene target status to the genome lineage assignment. Briefly, RNA was extracted from nasopharyngeal swabs in viral transport media using the CMG-1033-S kit (Chemagen, PerkinElmer, Baesweiler, Germany). 10µl of purified RNA was then amplified using the TaqPath COVID-19 CE-IVD RT-PCR kit (ThermoFisher Scientific, Waltham, MA, USA) and analysed on the Design & Analysis software v2.4. SGTF was denoted by lack of amplification of the S-gene target, with successful amplification of both the remaining ORF1ab and N-gene targets ($Ct \leq 30$).

Statistics

No statistical method was used to predetermine sample size. Data exclusion, randomization and blinding to allocation during experiments and outcome assessment were not applicable to this study.

Data Availability Statement

All of the SARS-CoV-2 genomes generated and presented in this manuscript are publicly accessible through the GISAID platform (https://www.gisaid.org/). The GISAID accession identifiers of the sequences analysed in this study are provided as part of Supplementary Table S1. Other raw data for this study are provided as a supplementary dataset at https://github.com/krisp-kwazulu-natal/SARSCoV2_South_Africa_Omicron_BA4_BA5. The reference SARS-CoV-2 genome (MN908947.3) was downloaded from the NCBI database (https://www.ncbi.nlm.nih.gov/).

Code Availability Statement

All custom scripts to reproduce the analyses and figures presented in this Article are available at https://github.com/krisp-kwazulu-natal/SARSCoV2_South_Africa_Omicron_BA4_BA5.

Supplementary References

21. Marivate, V. & Combrink, H. M. Use of Available Data To Inform The COVID-19 Outbreak in South Africa: A Case Study. *Data Sci. J.* **19**, (2020).
22. Marivate, V. *et al.* Coronavirus disease (COVID-19) case data - South Africa. *Zenodo* (2020) doi:10.5281/zenodo.3819126.
23. Msomi, N., Mlisana, K., de Oliveira, T. & Network for Genomic Surveillance in South Africa writing group. A genomics network established to respond rapidly to public health threats in South Africa. *Lancet Microbe* **1**, e229–e230 (2020).
24. Hadfield, J. *et al.* Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics* **34**, 4121–4123 (2018).
25. GitHub - neherlab/nextalign: Viral genome reference alignment. https://github.com/neherlab/nextalign.
26. Nguyen, L.-T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).
27. Rambaut, A., Lam, T. T., Max Carvalho, L. & Pybus, O. G. Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus Evol.* **2**, vew007 (2016).
28. Suchard, M. A. *et al.* Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evol.* **4**, vey016 (2018).
29. Griffiths, R. C. & Tavaré, S. Sampling theory for neutral alleles in a varying environment. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **344**, 403–410 (1994).
30. Rambaut, A., Drummond, A. J., Xie, D., Baele, G. & Suchard, M. A. Posterior summarization in Bayesian phylogenetics using Tracer 1.7. *Syst. Biol.* **67**, 901–904 (2018).
31. Wickham, H. ggplot2. *WIREs Comp Stat* **3**, 180–185 (2011).
32. Yu, G. Using ggtree to Visualize Data on Tree-Like Structures. *Curr. Protoc. Bioinformatics* **69**, e96 (2020).
33. Lemey, P., Rambaut, A., Welch, J. J. & Suchard, M. A. Phylogeography takes a relaxed random walk in continuous space and time. *Mol. Biol. Evol.* **27**, 1877–1885 (2010).

34. Dellicour, S., Rose, R., Faria, N. R., Lemey, P. & Pybus, O. G. SERAPHIM: studying environmental rasters and phylogenetically informed movements. *Bioinformatics* **32**, 3204–3206 (2016).
35. Rambaut, A. *et al.* A dynamic nomenclature proposal for SARS-CoV-2 to assist genomic epidemiology. *BioRxiv* (2020) doi:10.1101/2020.04.17.046086.
36. O'Toole, Á. *et al.* Assignment of epidemiological lineages in an emerging pandemic using the pangolin tool. *Virus Evol.* **7**, veab064 (2021).
37. Murrell, B. *et al.* Detecting individual sites subject to episodic diversifying selection. *PLoS Genet.* **8**, e1002764 (2012).
38. Kosakovsky Pond, S. L. & Frost, S. D. W. Not so different after all: a comparison of methods for detecting amino acid sites under selection. *Mol. Biol. Evol.* **22**, 1208–1222 (2005).
39. Shu, Y. & McCauley, J. GISAID: Global initiative on sharing all influenza data - from vision to reality. *Euro Surveill.* **22**, 30494 (2017).
40. Davies, N. G. *et al.* Estimated transmissibility and impact of SARS-CoV-2 lineage B.1.1.7 in England. *Science* **372**, (2021).
41. Campbell, F. *et al.* Increased transmissibility and global spread of SARS-CoV-2 variants of concern as at June 2021. *Euro Surveill.* **26**, (2021).
42. Lenth RV. *emmeans: Estimated Marginal Means, aka Least-Squares Means,R package version 1.6.1*. (2021).

NGS-SA consortium author list

Armand (Phillip) Bester[36,37], Mathilda Claassen[14], Deelan Doolabh[26], Innocent Mudau[26], Nokuzola Mbhele[26], Susan Engelbrecht[14], Dominique Goedhals[37,38], Diana Hardie[26,27], Nei-Yuan Hsiao[26,27,28], Arash Iranzadeh[39], Arshad Ismail[3], Rageema Joseph[26], Arisha Maharaj[2], Boitshoko Mahlangu[3], Kamela Mahlakwane[14,40], Ashlyn Davis[8], Gert Marais[26,27], Koleka Mlisana[41,42], Anele Mnguni[3], Thabo Mohale[3], Gerald Motsatsi[3], Peter Mwangi[37,43], Noxolo Ntuli[3], Martin Nyaga[37,43], Luicer Olubayo[28,39], Botshelo Radibe[9], Yajna Ramphal[1], Upasana Ramphal[2], Wilhelmina Strasheim[3], Naume Tebeila[3], Stephanie van Wyk[1], Shannon Wilson[14], Alexander G Lucaci[25], Steven Weaver[25], Akhil Maharaj[2], Yusasha Pillay[2], Michaela Davids[20], Adriano Mendes[20], Simnikiwe Mayaphi[44]

[36]Division of Virology, National Health Laboratory Service, Bloemfontein, South Africa
[37]Division of Virology, University of the Free State, Bloemfontein, South Africa
[38]PathCare Vermaak, Pretoria, South Africa
[39]Division of Computational Biology, Faculty of Health Sciences, University of Cape Town, Cape Town, South Africa.
[40]Division of Medical Virology, Faculty of Medicine and Health Sciences,
Stellenbosch University, Tygerberg, Cape Town, South Africa; NHLS Tygerberg Laboratory, Tygerberg, Cape Town, South Africa
[41]National Health Laboratory Service (NHLS), Johannesburg, South Africa
[42]Centre for the AIDS Programme of Research in South Africa (CAPRISA), Durban, South Africa.
[43]Next Generation Sequencing Unit, Division of Virology, Faculty of Health Sciences, University of the Free State, Bloemfontein, South Africa.
[44]Department of Medical Virology, University of Pretoria, Pretoria, South Africa

Chapter 6: A year of genomic surveillance reveals how the SARS-CoV-2 pandemic unfolded in Africa

This chapter gives in-depth insights into the genomic epidemiology of the SARS-CoV-2 virus in Africa over the first year of the pandemic, as inferred from viral genomes collected in 2021 from several countries across the continent. This was the first paper to present such a continental genomic analysis for SARS-CoV-2. This study was a colossal collaborative effort with contributing authors from more than 150 institutions in Africa for the generation of genomic data. As a co-first author on this study, my contribution to this study included conceptualization, computational method development for genomic data analysis, bioinformatics sequence assembly, data visualisation, writing and editing. The results show that outbreaks during 2020 in each African country were initiated by imported cases, mostly from Europe, and that as the pandemic developed, case numbers in African countries were likely many times higher than reported, and subsequent waves of the pandemic were more influenced by viral importations between African countries, rather than from outside the continent. Finally, the study maps out the evolution and dispersal of high-transmission variants within the continent. This chapter builds upon knowledge from previous chapters to put the genomic epidemiology of SARS-CoV-2 in South Africa in a continental context.

This chapter was published as a peer-reviewed research article in Science in September 2021 and can be accessed at the following DOI: 10.1126/science.abj4336. The chapter is presented in a similar format as the journal article. A full list of authors and affiliations is shown below.

Eduan Wilkinson†[1, 2], Marta Giovanetti†[3, 4], Houriiyah Tegally†[1], James E. San†[1], Richard Zekri[10], Abdoul K. Sangare[11], Abdoul-Salam Ouedraogo[12], Abdul K. Sesay[13], Abechi Priscilla[14], Adedotun-Sulaiman Kemi[14], Adewunmi M. Olubusuyi[15], Adeyemi O.O. Oluwapelumi[16], Adnène Hammami[17], Adrienne A. Amuri[18, 19], Ahmad Sayed[20], Ahmed E.O. Ouma[21], Aida Elargoubi[22, 23], Ajayi N. Anthony[24], Ajogbasile F. Victoria[14], Akano Kazeem[14], Akpede George[25], Alexander J. Trotter[26], Ali A. Yahaya[27], Alpha K. Keita[28, 29], Amadou Diallo[30], Amadou Kone[31], Amal Souissi[32], Amel Chtourou[17], Ana V. Gutierrez[26], Andrew J. Page[26], Anika Vinze[33], Arash Iranzadeh[6, 7], Arnold Lambisia[34], Arshad Ismail[35], Audu Rosemary[36], Augustina Sylverken[37], Ayoade Femi[14], Azeddine Ibrahimi[38], Baba Marycelin[39], Bamidele S. Oderinde[39], Bankole Bolajoko[14], Beatrice Dhaala[40], Belinda L. Herring[27], Berthe-Marie Njanpop-Lafourcade[27], Bronwyn Kleinhans[41], Bronwyn McInnis[10], Bryan Tegomoh[42], Cara Brook[43, 44], Catherine B. Pratt[45], Cathrine Scheepers[34, 46], Chantal G. Akoua-Koffi[47], Charles N. Agoti[34, 48], Christophe Peyrefitte[30], Claudia Daubenberger[49], Collins M. Morang'a[50], D. James Nokes[31, 51], Daniel G. Amoako[35], Daniel L. Bugembe[40], Danny Park[33], David Baker[26], Deelan Doolabh[7], Deogratius Ssemwanga[40, 52], Derek Tshiabuila[1], Diarra Bassirou[30], Dominic S.Y. Amuzu[50], Dominique Goedhals[53], Donwilliams O. Omuoyo[34], Dorcas Maruapula[54], Ebenezer Foster-Nyarko[26], Eddy

K. Lusamaki[18, 19], Edgar Simulundu[55], Edidah M. Ong'era[34], Edith N. Ngabana[18, 19], Edwin Shumba[56], Elmostafa El Fahime[57], Emmanuel Lokilo[18], Enatha Mukantwari[58], Eromon Philomena[14], Essia Belarbi[59], Etienne Simon-Loriere[60], Etilé A. Anoh[47], Fabian Leendertz[59], Faida Ajili[61], Fakayode O. Enoch[62], Fares Wasfi[63], Fatma Abdelmoula[32, 64], Fausta S. Mosha[27], Faustinos T. Takawira[65], Fawzi Derrar[66], Feriel Bouzid[32], Folarin Onikepe[14], Fowotade Adeola[67], Francisca M. Muyembe[18, 19], Frank Tanser[68, 69, 70], Fred A. Dratibi[27], Gabriel K. Mbunsu[19], Gaetan Thilliez[26], Gemma L. Kay[26], George Githinji[34, 71], Gert van Zyl[41, 72], Gordon A. Awandare[50], Grit Schubert[59], Gugu P. Maphalala[73], Hafaliana C. Ranaivoson[44], Hajar Lemriss[74], Happi Anise[14], Haruka Abe[75], Hela H. Karray[17], Hellen Nansumba[76], Hesham A. Elgahzaly[77], Hlanai Gumbo[65], Ibtihel Smeti[32], Ikhlas B. Ayed[32], Ikponmwosa Odia[25], Ilhem Boutiba-Ben Boubaker[78, 79], Imed Gaaloul[22], Inbal Gazy[80], Innocent Mudau[7], Isaac Ssewanyana[76], Iyaloo Konstantinus[81], Jean B. Lekana-Douk[82], Jean-Claude C. Makangara[18, 19], Jean-Jacques M. Tamfum[18, 19], Jean-Michel Heraud[30, 44], Jeffrey G. Shaffer[83], Jennifer Giandhari[1], Jingjing Li[84], Jiro Yasuda[75], Joana Q. Mends[85], Jocelyn Kiconco[52], John M. Morobe[34], John O. Gyapong[85], Johnson C. Okolie[14], John T. Kayiwa[40], Johnathan A. Edwards[68, 86], Jones Gyamfi[85], Jouali Farah[87], Joweria Nakaseegu[52], Joyce M. Ngoi[50], Joyce Namulondo[52], Julia C. Andeko[82], Julius J. Lutwama[40], Justin O'Grady[26], Katherine Siddle[33], Kayode T. Adeyemi[14], Kefentse A. Tumedi[88], Khadija M. Said[34], Kim Hae-Young[89], Kwabena O. Duedu[90], Lahcen Belyamani[38], Lamia Fki-Berrajah[17], Lavanya Singh[1], Leonardo de O. Martins[16], Lynn Tyers[7], Magalutcheemee Ramuth[91], Maha Mastouri[22, 23], Mahjoub Aouni[22], Mahmoud el Hefnawi[92], Maitshwarelo I. Matsheka[88], Malebogo Kebabonye[93], Mamadou Diop[30], Manel Turki[32], Marietou Paye[33], Martin M. Nyaga[94], Mathabo Mareka[95], Matoke-Muhia Damaris[96], Maureen W. Mburu[34], Maximillian Mpina[49, 97, 98], Mba Nwando[99], Michael Owusu[100], Michael R. Wiley[45], Mirabeau T. Youtchou[101], Mitoha O. Ayekaba[97], Mohamed Abouelhoda[102, 103], Mohamed G. Seadawy[104], Mohamed K. Khalifa[20], Mooko Sekhele[95], Mouna Ouadghiri[38], Moussa M. Diagne[30], Mulenga Mwenda[105], Mushal Allam[35], My V.T. Phan[40], Nabil Abid[79, 106], Nadia Touil[107], Nadine Rujeni[108, 109], Najla Kharrat[32], Nalia Ismael[110], Ndongo Dia[30], Nedio Mabunda[110], Nei-yuan Hsiao[7, 111], Nelson B. Silochi[97], Ngoy Nsenga[27], Nicksy Gumede[27], Nicola Mulder[112], Nnaemeka Ndodo[99], Norosoa H Razanajatovo[44], Nosamiefan Iguosadolo[14], Oguzie Judith[14], Ojide C. Kingsley[113], Okogbenin Sylvanus[25], Okokhere Peter[25], Oladiji Femi[114], Olawoye Idowu[14], Olumade Testimony[14], Omoruyi E. Chukwuma[67], Onwe E. Ogah[115], Onwuamah Chika[36], Oshomah Cyril[25], Ousmane Faye[30], Oyewale Tomori[14], Pascale Ondoa[56], Patrice Combe[116], Patrick Semanda[76], Paul E. Oluniyi[14], Paulo Arnaldo[110], Peter K. Quashie[50], Philippe Dussart[44], Phillip A. Bester[53], Placide K. Mbala[18, 19], Reuben Ayivor-Djanie[85], Richard Njouom[117], Richard O. Phillips[118], Richmond Gorman[118], Robert A. Kingsley[26], Rosina A.A. Carr[85], Saâd El Kabbaj[119], Saba Gargouri[17], Saber Masmoudi[32], Safietou Sankhe[30], Salako B. Lawal[36], Samar Kassim[77], Sameh Trabelsi[120], Samar Metha[33], Sami Kammoun[121], Sanaâ Lemriss[122], Sara H.A. Agwa[77], Sébastien Calvignac-Spencer[59], Stephen F. Schaffner[33], Seydou Doumbia[31], Sheila M. Mandanda[18, 19] , Sherihane Aryeetey[123], Shymaa S. Ahmed[123], Siham Elhamoumi[33], Soafy Andriamandimby[44], Sobajo Tope[14], Sonia Lekana-Douki[82], Sophie Prosolek[26], Soumeya Ouangraoua[124, 125], Steve A.

Mundeke[18, 19], Steven Rudder[26], Sumir Panji[112], Sureshnee Pillay[1], Susan Engelbrecht[41, 72], Susan Nabadda[76], Sylvie Behillil[126], Sylvie L. Budiaki[95], Sylvie van der Werf[126], Tapfumanei Mashe[65], Tarik Aanniz[38], Thabo Mohale[35], Thanh Le-Viet[26], Tobias Schindler[49, 97], Ugochukwu J. Anyaneji[1], Ugwu Chinedu[14], Upasana Ramphal[1, 69, 127], Uwanibe Jessica[14], Uwem George[14], Vagner Fonseca[1, 4, 128], Vincent Enouf[126], Vivianne Gorova[129, 130], Wael H. Roshdy[123], William K. Ampofo[50], Wolfgang Preiser[41, 72] , Wonderful T. Choga[54, 131], Yaw Bediako[50], Yeshnee Naidoo[1], Yvan Butera[108, 132, 133], Zaydah R. de Laurent[33], Amadou A. Sall[30], Ahmed Rebai[32], Anne von Gottberg[35], Bourema Kouriba[12], Carolyn Williamson[7, 69, 111], Daniel J. Bridges[105], Ihekweazu Chikwe[99], Jinal Bhiman[35], Madisa Mine[134], Matthew Cotten[40, 135], Sikhulile Moyo[54, 136], Simani Gaseitsiwe[54, 136], Ngonda Saasa[55], Pardis C. Sabeti[33], Pontiano Kaleebu[40], Yenew K. Tebeje[21], Sofonias K. Tessema[21], Happi Christian[14], John Nkengasong[21], Tulio de Oliveira[1, 2, 69, 137]

1. KwaZulu-Natal Research Innovation and Sequencing Platform (KRISP), Nelson R Mandela School of Medicine, University of KwaZulu-Natal; Durban, South Africa.
2. Centre for Epidemic Response and Innovation (CERI), School of Data Science and Computational Thinking, Stellenbosch University; Stellenbosch, South Africa.
3. Laboratorio de Flavivirus, Fundacao Oswaldo Cruz; Rio de Janeiro, Brazil.
4. Laboratório de Genética Celular e Molecular, Universidade Federal de Minas Gerais; Belo Horizonte, Minas Gerais, Brazil.
5. Department of Geography and GIS, University of Cincinnati; Cincinnati, Ohio, USA
6. Institute of Infectious Diseases and Molecular Medicine, Department of Integrative Biomedical Sciences, Computational Biology Division, University of Cape Town; Cape Town, South Africa.
7. Division of Medical Virology, Wellcome Centre for Infectious Diseases in Africa, Institute of Infectious Disease and Molecular Medicine, University of Cape Town; Cape Town, South Africa.
8. Department of Entomology and Plant Pathology, North Carolina State University; Raleigh, North Carolina, United States of America.
9. Bioinformatics Research Center, North Carolina State University; Raleigh, North Carolina, United States of America.
10. Cancer Biology Department, Virology and Immunology Unit, National Cancer Institute, Cairo University, 11796, Egypt
11. Centre d'Infectiologie Charles Mérieux-Mali (CICM-Mali); Bamako, Mali.
12. Bacteriology and Virology Department Souro Sanou University Hospital, Bobo-Dioulasso, Burkina Faso.
13. MRCG at LSHTM Genomics lab; Fajara, Gambia.
14. African Centre of Excellence for Genomics of Infectious Diseases (ACEGID), Redeemer's University; Ede, Osun State, Nigeria.
15. Department of Virology, College of Medicine, University of Ibadan; Ibadan, Nigeria
16. Department of Medical Microbiology and Parasitology. Faculty of Basic Clinical Sciences. College of Health Sciences. University of Ilorin; Ilorin, Kwara State, Nigeria.
17. CHU Habib Bourguiba, Laboratory of Microbiology, Faculty of Medicine of sFax, University of sFax; sFax, Tunisia.
18. Pathogen Sequencing Lab, Institut National de Recherche Biomédicale (INRB); Kinshasa, Democratic Republic of the Congo.
19. Université de Kinshasa (UNIKIN); Kinshasa, the Democratic Republic of the Congo.
20. Genomics Research Program, Children's Cancer Hospital; Cairo, Egypt.

21. Institute of Pathogen Genomics, Africa Centres for Disease Control and Prevention (Africa CDC); Addis Ababa, Ethiopia.
22. Laboratory of Transmissible Diseases and Biological Active Substances (LR99ES27), Faculty of Pharmacy of Monastir; Monastir, Tunisia.
23. Laboratory of Microbiology, University Hospital of Monastir; Monastir, Tunisia.
24. Internal Medicine Department, Alex Ekwueme Federal University Teaching Hospital; Abakaliki, Nigeria.
25. Irrua Specialist Teaching Hospital; Irrua, Nigeria.
26. Quadram Institute Bioscience, Norwich, United Kingdom.
27. World Health Organization, Africa Region; Brazzaville Congo
28. Centre de Recherche et de Formation en Infectiologie de Guinée (CERFIG), Université de Conakry; Conakry, Guinea
29. TransVIHMI, Montpellier University /IRD/INSERM; Montpellier, France
30. Virology Department, Institut Pasteur de Dakar; Dakar, Senegal.
31. Mali-University Clinical Research Center (UCRC); Bamako, Mali.
32. Laboratory of Molecular and Cellular Screening Processes, Centre of Biotechnology of Sfax, University of Sfax; Sfax, Tunisia.
33. Broad Insitute of Harvard and MIT; Cambridge, Massachusetts, United States of America.
34. KEMRI-Wellcome Trust Research Programme/KEMRI-CGMR-C; Kilifi, Kenya.
35. National Institute for Communicable Diseases (NICD) of the National Health Laboratory Service (NHLS); Johannesburg, South Africa.
36. The Nigerian Institute of Medical Research; Yaba, Lagos, Nigeria
37. Institute of Virology, Charité – Universitätsmedizin; Berlin, Germany.
38. Medical Biotechnology Laboratory, Rabat Medical and Pharmacy School, Mohammed The Vth University; Rabat, Morocco.
39. Department of Immunology, University of Maiduguri Teaching Hospital, P.M.B. 1414; Maiduguri, Nigeria
40. MRC/UVRI & LSHTM Uganda Research Unit; Entebbe, Uganda.
41. Division of Medical Virology, Faculty of Medicine and Health Sciences, Stellenbosch University; Tygerberg, Cape Town, South Africa.
42. The Biotechnology Center of the University of Yaoundé I, Cameroon & CDC Foundation; Yaounde, Cameroon.
43. Department of Ecology and Evolution; University of Chicago; Chicago, Illinois, United States of America.
44. Virology Unit, Institut Pasteur de Madagascar; Antananarivo, Madagascar.
45. University of Nebraska Medical Center (UNMC); Omaha, Nebraska, United States of America.
46. Antibody Immunity Research Unit, School of Pathology, University of the Witwatersrand; Johannesburg, South Africa
47. CHU de Bouaké, Laboratoire / Unité de Diagnostic des Virus des Fièvres Hémorragiques et Virus Émergents; Bouaké, Côte d'Ivoire.
48. School of Public Health, Pwani University; Kilifi, Kenya.
49. Swiss Tropical and Public Health Institute; Basel, Switzerland
50. West African Centre for Cell Biology of Infectious Pathogens (WACCBIP), Department of Biochemistry, Cell and Molecular Biology, University of Ghana; Accra, Ghana.
51. School of Life Sciences and Zeeman Institute for Systems Biology and Infectious Disease Epidemiology Research (SBIDER), University of Warwick; Coventry, United Kingdom.
52. Uganda Virus Research Institute; Entebbe, Uganda.
53. Division of Virology, National Health Laboratory Service and University of the Free State; Bloemfontein, South Africa.
54. Botswana Harvard AIDS Institute Partnership & Botswana Harvard HIV Reference Laboratory; Gaborone, Botswana.
55. University of Zambia, School of Veterinary Medicine, Department of Disease Control; Lusaka, Zambia.

56. African Society for Laboratory Medicine; Addis Ababa, Ethiopia
57. Functional genomic Platform / National Centre for Scientific and Technical Research (CNRST); Rabat, Morocco.
58. Rwanda National Reference Laboratory; Kigali, Rwanda.
59. Robert Koch-Institute; Berlin, Germany.
60. G5 Evolutionary Genomics of RNA viruses, Institut Pasteur; Paris, France.
61. Research Unit of Autoimmune Diseases UR17DN02, Military hospital of Tunis , University of Tunis El manar; Tunis, Tunisia.
62. Department of Public Health. Ministry of Health; Ilorin, Kwara State, Nigeria.
63. Laboratory of Clinical Virology, Institut Pasteur de Tunis, Tunis, Tunisia
64. Faculty of pharmacy of Monastir; Monastir, Tunisia.
65. National Microbiology Reference Laboratory; Harare, Zimbabwe.
66. National Influenza Centre, Viral Respiratory Laboratory; Algiers, Algeria.
67. Medical Microbiology and Parasitology Department, College of Medicine, University of Ibadan; Ibadan, Nigeria
68. Lincoln International Institute for Rural Health, University of Lincoln; Lincoln, United Kingdom.
69. Centre for the AIDS Programme of Research in South Africa (CAPRISA); Durban, South Africa.
70. Africa Health Research Institute; KwaZulu-Natal, Durban, South Africa
71. Department of Biochemistry and Biotechnology, Pwani University; Kilifi, Kenya.
72. National Health Laboratory Service (NHLS); Tygerberg, Cape Town, South Africa.
73. Institution and Department: Ministry Of Health, COVID-19 Testing Laboratory; Mbabane, Kingdom of Eswatini.
74. Laboratory of Health Sciences and Technologies, High Institute of Health Sciences, Hassan 1st University; Settat, Morocco.
75. Department of Emerging Infectious Diseases, Institute of Tropical Medicine, Nagasaki University; Nagasaki, Japan.
76. Central Public Health Laboratories (CPHL); Kampala, Uganda.
77. Faculty of Medicine Ain Shams Research institute (MASRI), Ain Shams University; Cairo, Egypt.
78. Charles Nicolle Hospital, Laboratory of Microbiology, National Influenza Center, 1006, Tunis, Tunisia
79. University of Tunis El Manar, Faculty of Medicine of Tunis, LR99ES09, 1007, Tunis, Tunisia
80. Department of Biochemistry and Molecular Biology, The Institute for Medical Research Israel-Canada, Hadassah Medical School, The Hebrew University of Jerusalem; Jerusalem, Israel.
81. Namibia Institute of Pathology; Windhoek, Namibia
82. Centre Interdisciplinaires de Recherches Medicales de Franceville (CIRMF); Franceville, Gabon.
83. Department of Biostatistics and Data Science, School of Public Health and Tropical Medicine, Tulane University; New Orleans, Louisiana, United States of America.
84. Urban Health Collaborative, Dornsife School of Public Health, Drexel University; Philadelphia, United States of America.
85. UHAS COVID-19 Testing and Research Centre, University of Health and Allied Sciences; Ho, Ghana.
86. Rollins School of Public Health, Emory University; Atlanta, Georgia, United States of America.
87. Anoual Laboratory; Casablanca, Morocco.
88. Botswana Institute for Technology Research and Innovation; Gaborone, Botswana.
89. New York University Grossman School of Medicine; New York, United States of America.
90. Centre de Recherches Medicales de Lambarene (CERMEL); Lambarene; Gabon.
91. Virology/Molecular Biology Department, Central Health Laboratory, Ministry of Health and Wellness, Mauritius
92. Center of Scientific Excellence for Influenza Viruses, National Research Centre (NRC); Cairo Egypt.
93. Ministry of health and wellness; Gaborone, Botswana.

94. Next Generation Sequencing Unit and Division of Virology, Faculty of Health Sciences, University of the Free State, Bloemfontein 9300, South Africa
95. National Reference Laboratory Lesotho; Maseru, Lesotho
96. Centre for Biotechnology Research and Development, Kenya Medical Research Institute, Nairobi, Kenya
97. Laboratorio de Investigaciones de Baney; Baney, Equatorial Guinea.
98. Ifakara Health Institute; Dar-es-Salaam, Tanzania.
99. Nigeria Centre for Disease Control; Abuja, Nigeria.
100. Department of Medical Diagnostics, Kumasi Centre for Collaborative Research in Tropical Medicine, Kwame Nkrumah University of Science and Technology; Kumasi, Ghana.
101. Department of Medical Laboratory Science, Niger Delta University; Bayelsa State, Nigeria
102. Systems and Biomedical Engineering Department, Faculty of Engineering, Cairo University, Cairo 12613, Egypt.
103. King Faisal Specialist Hospital and Research Center, Riyadh, Kingdom of Saudi Arabia.
104. Biological Prevention Department, Main Chemical Laboratories ,Egypt Army; Cairo, Egypt.
105. PATH; Lusaka, Zambia.
106. Department of Biotechnology, High Institute of Biotechnology of Sidi Thabet; University of Manouba; BP-66, 2020 Ariana-Tunis, Tunisia.
107. Genomic Center for Human Pathologies (GENOPATH), Faculty of Medicine and Pharmacy, University Mohammed V; Rabat, Morocco.
108. Rwanda National Joint Task Force COVID-19, Rwanda Biomedical Centre, Ministry of Health; Kigali, Rwanda
109. School of Health Sciences, College of Medicine and Health Sciences, University of Rwanda; Kigali, Rwanda
110. Instituto Nacional de Saude (INS); Maputo, Mozambique.
111. National Health Laboratory Service (NHLS); Cape Town, South Africa.
112. Computational Biology Division, Department of Integrative Biomedical Sciences, IDM, CIDRI Africa Wellcome Trust Centre, University of Cape Town; Cape Town, South Africa.
113. Virology Laboratory, Alex Ekwueme Federal University Teaching Hospital; Abakaliki, Nigeria.
114. Department of Epidemiology and Community Health, Faculty of Clinical Sciences. College of Health Sciences. University of Ilorin; Ilorin, Kwara State, Nigeria.
115. Alex Ekwueme Federal University Teaching Hospital; Abakaliki, Nigeria.
116. Mayotte Hospital Center; Mayotte, France.
117. Virology Service, Centre Pasteur of Cameroun; Yaounde, Cameroun.
118. Kumasi Centre for Collaborative Research in Tropical Medicine, Kwame Nkrumah University of Science and Technology; Kumasi, Ghana
119. Laboratoire de Recherche et d'Analyses Médicales de la Gendarmerie Royale; Rabat, Morocco.
120. Clinical and Experimental Pharmacology Lab, LR16SP02, National Center of Pharmacovigilance, University of Tunis El Manar; Tunis, Tunisia.
121. CHU Hedi Chaker Sfax, Service de Pneumologie; Tunis, Tunisia.
122. Laboratoire de Recherche et d'Analyses Médicales de la Gendarmerie Royale; Rabat, Morocco.
123. Central Public Health Laboratories (CPHL); Cairo, Egypt.
124. Centre MURAZ; Ouagadougou, Burkina Faso
125. National Institute of Public Health of Burkina Faso (INSP/BF); Ouagadougou, Burkina Faso
126. National Reference Center for Respiratory Viruses, Molecular Genetics of RNA Viruses, UMR 3569 CNRS, University of Paris, Institut Pasteur; Paris, France.
127. Sub-Saharan African Network For TB/HIV Research Excellence (SANTHE); Durban, South Africa
128. Coordenação Geral de Laboratórios de Saúde Pública/Secretaria de Vigilância em Saúde, Ministério da Saúde, Brasília, Distrito Federal, Brazil.
129. World Health Organization, WHO Lesotho; Maseru, Lesotho
130. Med24 Medical Centre; Ruwa, Zimbabwe.

131. Division of Human Genetics, Department of Pathology, University of Cape Town; Cape Town, South Africa.
132. Center for Human Genetics, College of Medicine and Health Sciences, University of Rwanda; Kigali, Rwanda
133. Laboratory of Human Genetics, GIGA Research Institute; Liège, Belgium
134. National Health laboratory; Gaborone, Botswana.
135. MRC-University of Glasgow Centre for Virus Research; Glasgow, United Kingdom.
136. Harvard T.H. Chan School of Public Health; Boston, Massachusetts, United States of America.
137. Department of Global Health, University of Washington; Seattle, USA.

† These authors contributed equally

Abstract

The progression of the SARS-CoV-2 pandemic in Africa has so far been heterogeneous and the full impact is not yet well understood. Here, we describe the genomic epidemiology using a dataset of 8746 genomes from 33 African countries and two overseas territories. We show that the epidemics in most countries were initiated by importations predominantly from Europe, which diminished following the early introduction of international travel restrictions. As the pandemic progressed, ongoing transmission in many countries and increasing mobility led to the emergence and spread within the continent of many variants of concern and interest, such as B.1.351, B.1.525, A.23.1 and C.1.1. Although distorted by low sampling numbers and blind-spots, the findings highlight that Africa must not be left behind in the global pandemic response, otherwise it could become a breeding ground for new variants.

Main Text

Severe acute respiratory syndrome-related coronavirus 2 (SARS-CoV-2) emerged in late 2019 in Wuhan, China(*1*, *2*). Since then, the virus has spread to all corners of the world, causing almost 150 million cases of coronavirus disease 2019 (COVID-19) and over three million deaths by the end of April 2021. Throughout the pandemic, it has been noted that Africa accounts for a relatively low proportion of reported cases and deaths – by the end of April 2021, there had been ~4.5 million cases and ~120000 deaths on the continent, corresponding to less than 4% of the global burden. However, emerging data from seroprevalence surveys and autopsy studies in some African countries suggests that the true number of infections and deaths may be several fold higher than reported(*3*, *4*). In addition, a recent analysis has shown that the second wave of the pandemic was more severe than the first wave in many African countries (*5*).

The first cases of COVID-19 on the African continent were reported in Nigeria, Egypt and South Africa between mid-February and early March 2020, and most countries had reported cases by the end of March 2020 (*6*, *7* ,*8*). These early cases were concentrated amongst air travellers returning from regions of the world with high levels of community transmission. Many African countries introduced early public health and social measures (PHSM), including international travel controls,

quarantine for returning travellers, and internal lockdown measures to limit the spread of the virus and give health services time to prepare (*9*, *5*). The initial phase of the epidemic was then heterogeneous with relatively high case numbers reported in North Africa and Southern Africa, and fewer cases reported in other regions.

From the onset of the pandemic, genomic surveillance has been at the forefront of the COVID-19 response in Africa (*10*). Rapid implementation of SARS-CoV-2 sequencing by various laboratories in Africa enabled genomic data to be generated and shared from the early imported cases. In Nigeria, the first genome sequence was released just three days after the announcement of the first case (*6*). Similarly, in Uganda, a sequencing programme was set up rapidly to facilitate virus tracing, and the collection of samples for sequencing began immediately upon confirmation of the first case (*11*). In South Africa, the network for genomic surveillance in South Africa (NGS-SA) was established in March 2020 and within weeks genomic analysis was helping to characterize outbreaks and community transmission (*12*).

Genomic surveillance has also been critical for monitoring ongoing SARS-CoV-2 evolution and detection of new SARS-CoV-2 variants in Africa. Intensified sampling by NGS-SA in the Eastern Cape Province of South Africa in November 2020, in response to a rapid resurgence of cases, led to the detection of B.1.351 (501Y.V2)(*13*). This variant was subsequently designated a variant of concern (VOC) by the World Health Organization (WHO), due to evidence of increased transmissibility(*14*) and resistance to neutralizing antibodies elicited by natural infection and vaccines (*15–17*).

Here, we perform phylogenetic and phylogeographic analysis of SARS-CoV-2 genomic data from 33 African countries and two overseas territories to help characterize the dynamics of the pandemic in Africa. We show that the early introductions were predominantly from Europe, but that as the pandemic progressed there was increasing spread between African countries. We also describe the emergence and spread of a number of key SARS-CoV-2 variants in Africa, and highlight how the spread of B.1.351 (501Y.V2) and other variants contributed to the more severe second wave of the pandemic in many countries.

Results

SARS-CoV-2 genomic data

By 5 May 2021, 14504 SARS-CoV-2 genomes had been submitted to the GISAID database (*18*) from 38 African countries and two overseas territories (Mayotte and Réunion) (Fig. 6.1A). Overall, this corresponds to approximately one sequence per ~300 reported cases. Almost half of the sequences were from South Africa (n=5362), consistent with it being responsible for almost half of the reported cases in Africa. Overall, the number of sequences correlates closely with the number of reported cases per country (Fig. 6.1B). The countries/territories with the highest

coverage of sequencing (defined as genomes per reported case) are Kenya (n=856, one sequence per ~203 cases), Mayotte (n=721; one sequence per ~21 cases), and Nigeria (n=660, one sequence per ~250 cases). Although genomic surveillance started early in many countries, few have evidence of consistent sampling across the whole year. Half of all African genomes were deposited in the first ten weeks of 2021, suggesting intensified surveillance in the second wave following the detection of B.1.351/501Y.V2 and other variants (Fig. 6.1C and 6.1D).

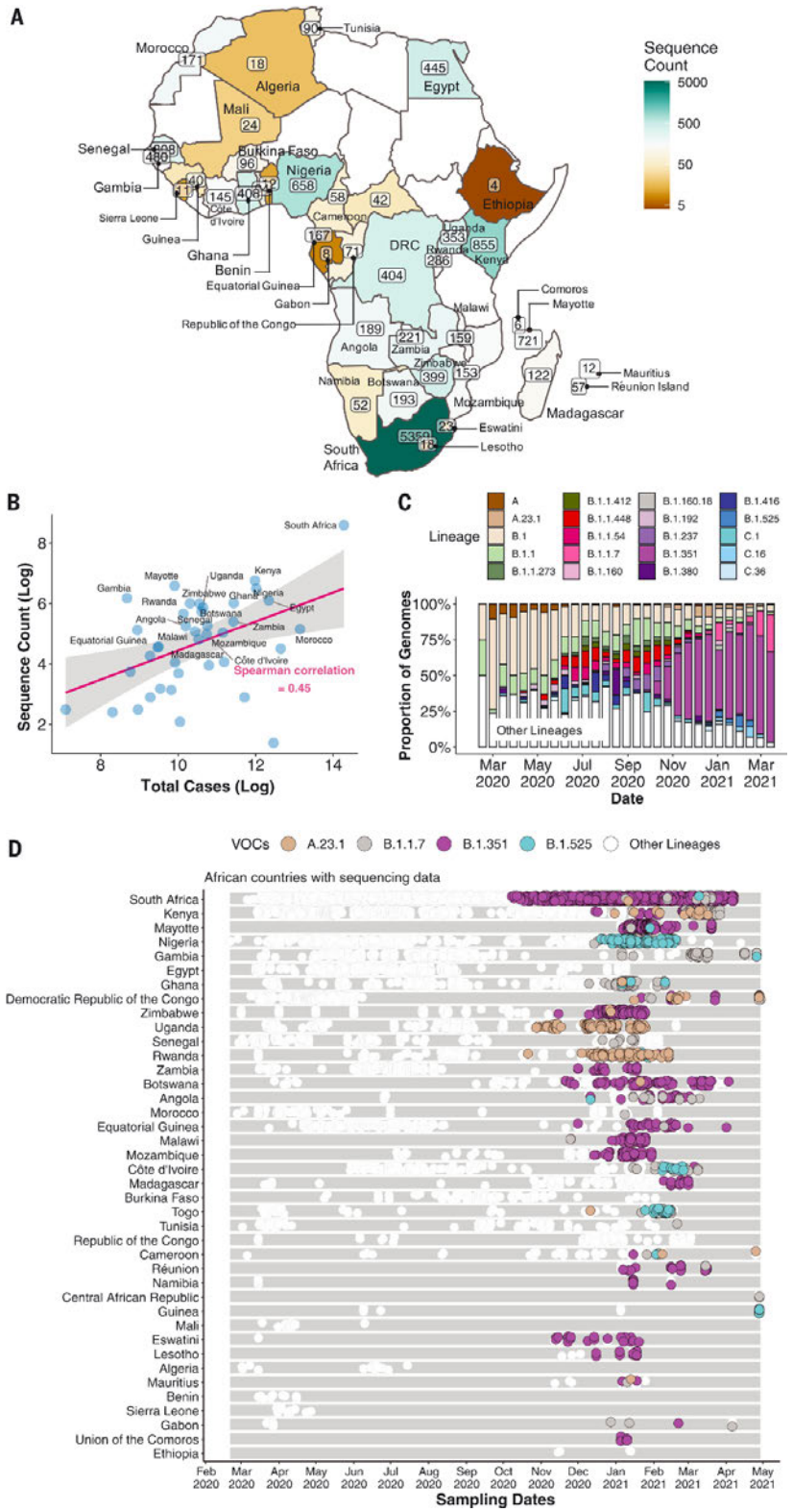


Figure 6.1: *SARS-CoV-2 sequences in Africa. (A) Map of the African continent with the number of SARS-CoV-2 sequences reflected in GISAID as of 5 May 2021. (B) Regression plot of the number of viral sequences vs. the number of reported COVID-19 cases in various African*

countries as of 5 May 2021. Countries with >500 sequences are labelled. (C) Progressive distribution of the top 20 PANGO lineages on the African continent. (D) Temporal sampling of SARS-CoV-2 sequences in African countries (ordered by total number of sequences) through time with VOCs of note highlighted and annotated according to their PANGO lineage assignment.

Genetic diversity and lineage dynamics in Africa

Of the 10326 genomes retrieved from GISAID by the end of March 2021, 8,746 genomes passed quality control (QC) and met the minimum metadata requirements. These genomes from Africa were compared in a phylogenetic framework with 11891 representative genomes from around the world. Ancestral location state reconstruction of the dated phylogeny (hereafter referred to as discrete phylogeographic reconstruction) allowed us to infer the number of viral imports and exports between Africa and the rest of the world, and between individual African countries. African genomes in this study spanned the whole global genetic diversity of SARS-CoV-2, a pattern that largely reflects multiple introductions over time from the rest of the world (Fig. 6.2A).

In total, we detected at least 757 (95% CI: 728 - 786) viral introductions into African countries between the start of 2020 and February 2021, over half of which occurred before the end of May 2020. Whilst the early phase of the pandemic was dominated by importations from outside Africa, predominantly from Europe, there was then a shift in the dynamics, with an increasing number of importations from other African countries as the pandemic progressed (Fig. 6.2B and 2C). A rarefaction analysis in which we systematically subsampled genomes shows that vastly more introductions would have likely been identified with increased sampling in Africa or globally, suggesting that the introductions we identified are really just the “ears of the hippo”, or tip of the iceberg (Appendix E - Fig. S1).

South Africa, Kenya and Nigeria appear as major sources of importations into other African countries (Fig. 6.2D), although this is likely to be influenced by these three countries having the greatest number of deposited sequences. Particularly striking is the southern African region, where South Africa is the source for a large proportion (~80%) of the importations to other countries in the region. The North African region demonstrates a different pattern to the rest of the continent, with more viral introductions from Europe and Asia (particularly the Middle East) than from other African countries (Appendix E - Fig. S2).

Africa has also contributed to the international spread of the virus with at least 324 (95% CI: 728 - 786) exportation events from Africa to the rest of the world detected in this dataset. Consistent with the source of importations, most exports were to Europe (41%), Asia (26%) and North America (14%). As with the number of importations exports were relatively evenly distributed over the one year period (Appendix E - Fig. S3). However, an increase in the number of exportation events occurred between December 2020 and March 2021, which coincided with the second wave of infections in Africa and with some relaxations of travel restrictions around the world.

The early phase of the pandemic was characterized by the predominance of lineage B.1. This was introduced multiple times to African countries and has been detected in all but one of the countries included in this analysis. After its emergence in South Africa, B.1.351 became the most frequently detected SARS-CoV-2 lineage found in Africa (n=1,769; ~20%) (Fig. 6.1C). It was first sampled on 8 October 2020 in South Africa (*13*) and has since spread to 20 other African countries.

As air travel came to an almost complete halt in March/April 2020, the number(s) of detectable viral imports into Africa decreased and the pandemic entered a phase that was characterized in sub-Saharan Africa by sustained low levels of within-country movements and occasional international viral movements between neighbouring countries; presumably via road and rail links between these. Though some border posts between countries were closed during the initial lockdown period (Table S1), others remained open to allow trade to continue. Regional trade in southern Africa was only slightly impacted by lockdown restrictions and quickly rebounded to pre-pandemic levels (Appendix E - Fig. S4) following the relaxation of restrictions between June 2020 and December 2020.

Although lineage A viruses were imported into several African countries, they only account for 1.3% of genomes sampled in Africa. Despite lineage A viruses initially causing many localized clustered outbreaks, each the result of independent introductions to several countries (e.g. Burkina Faso, Cote d'Ivoire and Nigeria), they were later largely replaced by lineage B viruses as the pandemic evolved. This is possibly due to the increased transmissibility of B lineage viruses by virtue of the D614G mutation in spike (*19*, *20*). However, there is evidence of an increasing prevalence of lineage A viruses in some African countries (*11*). In particular, A.23.1 emerged in East Africa and appears to be increasing rapidly in prevalence in Uganda and Rwanda (*11*). Furthermore, a highly divergent variant from lineage A was recently identified in Angola from individuals arriving from Tanzania (*21*).

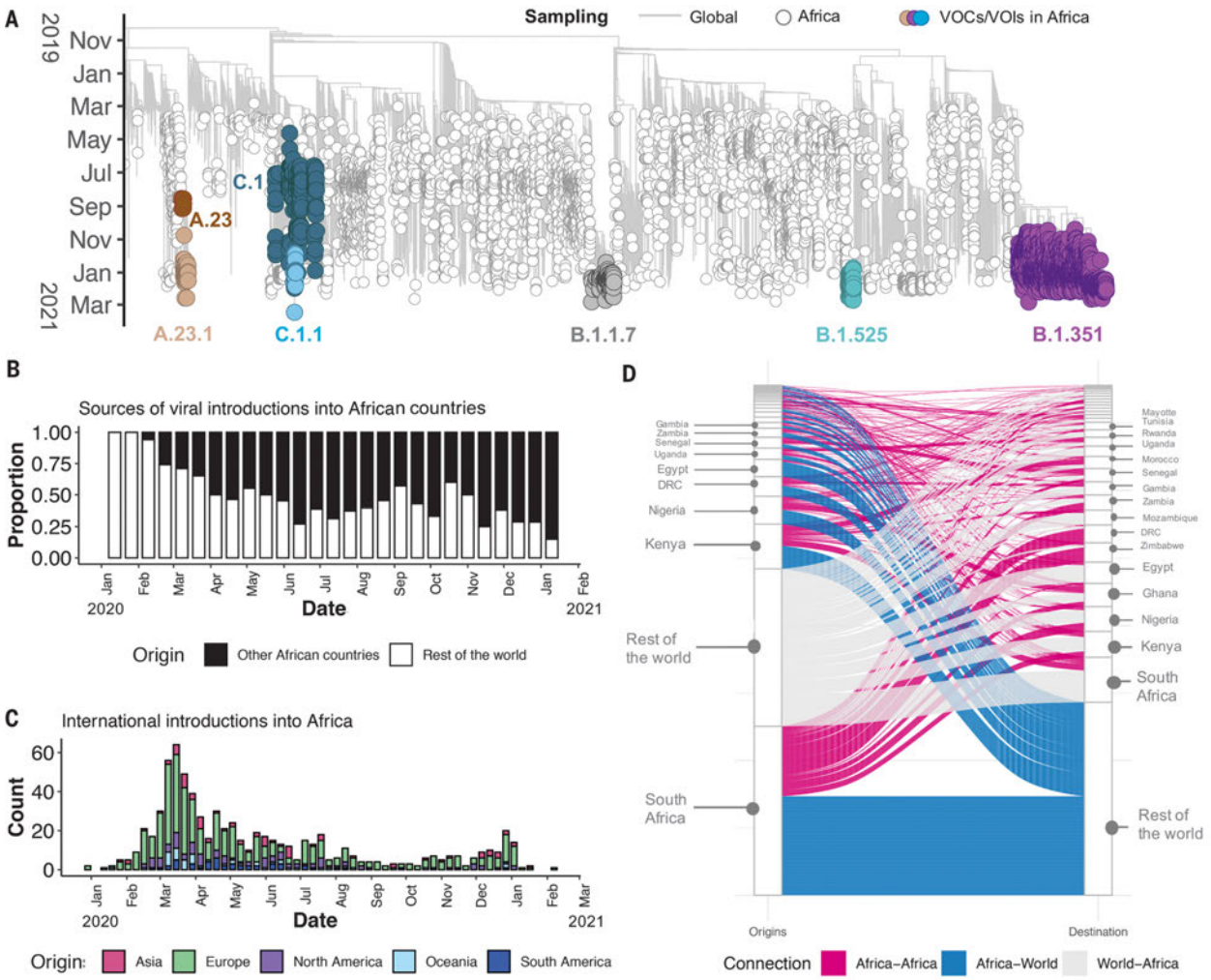


Figure 6.2: *(2A) Time resolved Maximum Likelihood tree containing 8,746 high quality African SARS-CoV-2 near-full-genome sequences analyzed against a backdrop of global reference sequences. Variants of interest (VOI) and concern (VOC) are highlighted on the phylogeny. (2B) Sources of viral introductions into African countries characterized as external introductions from the rest of the world vs internal introductions from other African countries. (2C) Total external viral introductions over time into Africa. (2D) The number of viral imports and exports into and out of various African countries depicted as internal (between African countries in pink) or external (between African and non-African countries in blue and grey).*

Emergence and spread of new SARS-CoV-2 variants

In order to determine how some of the key SARS-CoV-2 variants are spreading within Africa, we performed phylogeographic analyses on the VOC B.1.351, the variant of interest (VOI) B.1.525, and on two additional variants that emerged and that we designated as VOIs for this analysis (A.23.1 and C.1.1). These African VOCs and VOIs have multiple mutations on Spike glycoprotein and molecular clock analysis of these four datasets provided strong evidence that these four lineages are evolving in a clocklike manner (Fig. 6.3A, 6.3B).

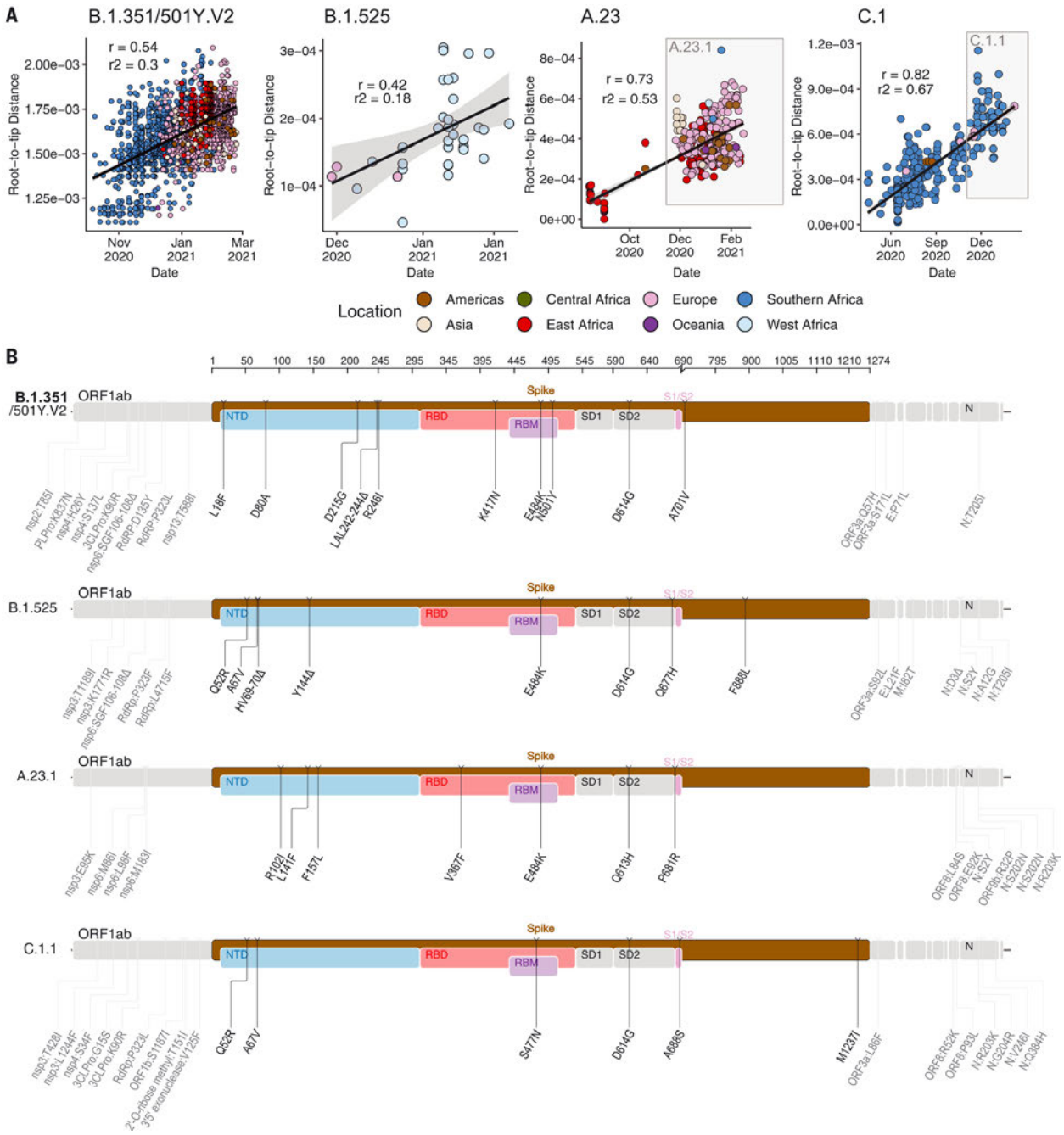


Figure 6.3: *(A) Root-to-tip regression plots for four lineages of interest. C.1 and A.23 show continued evolution into VOIs C.1.1 and A.23.1 respectively. (B) Genome maps of four VOCs/VOIs where the spike region is shown in detail and in color and the rest of the genome is shown in grey. ORF: open reading frame; NTD: N-terminal domain; RBD: receptor binding domain; RBM: receptor binding motif; SD1: subdomain 1 and SD2: subdomain 2.*

B.1.351 was first sampled in South Africa in October 2020, but phylogeographic analysis suggests that it emerged earlier, around August 2020. It is defined by ten mutations in the spike protein, including K417N, E484K and N501Y in the receptor-binding domain (Fig. 6.3B). Following its emergence in the Eastern Cape, it spread extensively within South Africa (Fig. 6.4A). By November 2020, the variant had spread into neighbouring Botswana and Mozambique and by December 2020 it had reached Zambia and Mayotte. Within the first three months of 2021, further exports from South Africa into Botswana, Zimbabwe, Mozambique and Zambia occurred. By March 2021, B.1.351 had become the dominant lineage within most Southern African countries as well as the overseas territories of Mayotte and Réunion (Appendix E - Fig. S5). Our phylogeographic reconstruction also demonstrates movement of B.1.351 into East and Central Africa directly from southern Africa. Our discrete phylogeographic analysis of a wider sample of B.1.351 isolates demonstrate the spread of the lineage into West Africa. This patient from West Africa had a known travel history to Europe so it possible the patient acquired the infection while in Europe or in transit and not from other African sources (Fig. S6).

B.1.525 is a VOI defined by six substitutions in the spike protein (Q52R, A67V, E484K, D614G, O677H and F888L), and two deletions in the N-terminal domain (HV69-70Δ and Y144Δ). This was first sampled in the United Kingdom in mid-December 2020, but our phylogeographic reconstruction suggests that the variant originated in Nigeria in November 2020 (95% highest posterior density (HPD) 2020-11-01 to 2020-12-03) (Fig. 6.4B). Since then it has spread throughout much of Nigeria and neighbouring Ghana. Given sparse sampling from other neighbouring countries within West and Central Africa (Fig. 6.1A & 6.1C), the extent of the spread of this VOI in the region is not clear. Beyond Africa, this VOI has spread to Europe and the US (Appendix E - Fig. S6).

We designated A.23.1 and C.1.1 as VOIs for the purposes of this analysis, as they present good examples of the continued evolution of the virus within Africa(*11*, *13*). Lineage A.23, characterized by three spike mutations (F157L, V367F and Q613H), was first detected in a Ugandan prison in Amuru in July 2020 (95% HPD: 2020-07-15 to 2020-08-02). From there, the lineage was transmitted to Kitgum prison, possibly facilitated by the transfer of prisoners. Subsequently, the A.23 lineage spilled into the general population and spread to Kampala, adding other spike mutations (R102I, L141F, E484K, P681R) along with additional mutations in nsp3, nsp6, ORF8 and ORF9, prompting a new lineage classification, A.23.1 (Fig. 6.3A & 6.3B). Since the emergence of A.23.1 in September 2020 (95% HPD: 2020-09-02 to 2020-09-28), it has spread regionally into neighbouring Rwanda and Kenya and has now also reached South Africa and Botswana in the south and Ghana in the west (Fig. 6.4C). However, our phylogeographic reconstruction of A.23.1 suggests that the introduction into Ghana may have occurred via Europe (Appendix E - Fig. S6), whereas the introductions into southern Africa likely occurred directly from East Africa. This is consistent with epidemiological data suggesting that the case detected in South Africa was a contact of an individual who had recently travelled to Kenya.

Lineage C.1 emerged in South Africa in March 2020 (95% HPD: 2020-03-13 to 2020-04-17) during a cluster outbreak prior to the first wave of the epidemic(*13*). C.1.1 is defined by the spike mutations S477N, A688S, M1237I and also contains the Q52R and A67V mutations similar to B.1.525 (Fig. 6.3B). A continuous trait phylogeographic reconstruction of the movement dynamics of these lineages suggests that C.1 emerged in the city of Johannesburg and spread within South Africa during the first wave (Fig. 6.4D). Independent exports of C.1 from South Africa led to regional spread to Zambia (June-July, 2020) and Mozambique (July-August 2020), and the evolution to C.1.1 seems to have occurred in Mozambique around mid-September 2020 (95% HPD: 2020-09-07 to 2020-10-05). In depth analysis of SARS-CoV-2 genotypes from Mozambique suggest that the C.1.1. lineage was the most prevalent in the country until the introduction of B.1.351, which has dominated the epidemic since (Appendix E - Fig. S5).

The VOC B.1.1.7, which was first sampled in Kent, England in September 2020(*22*), has also increased in prevalence in several African countries (Appendix E - Fig. S5) To date, this VOC has been detected in eleven African countries, as well as the Indian Ocean islands of Mauritius and Mayotte (Appendix E - Fig. S7). The time-resolved phylogeny suggests that this lineage was introduced into Africa on at least 16 occasions between November 2020 and February 2021 with evidence of local transmission in Nigeria and Ghana.

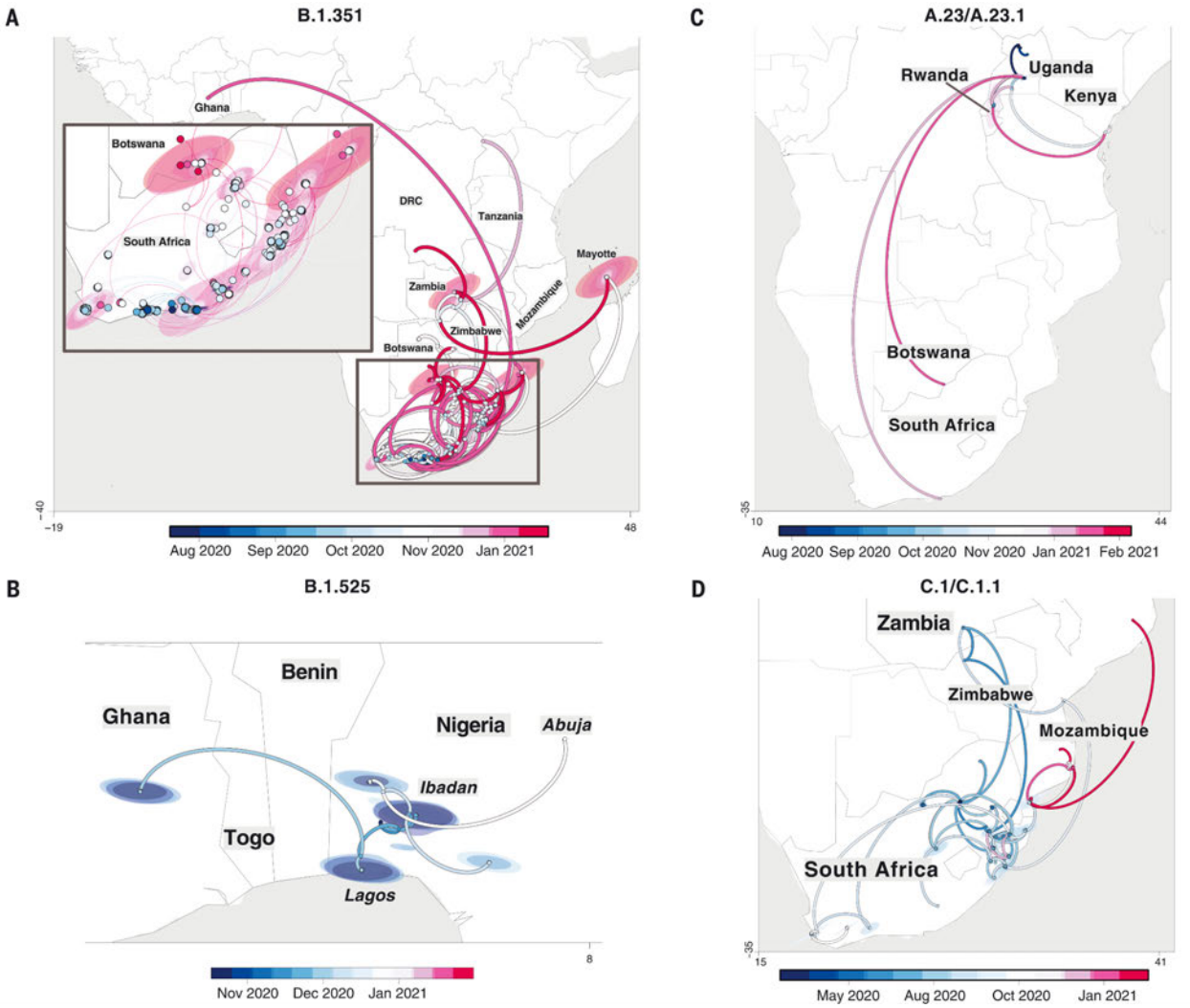


Figure 6.4: *Phylogeographic reconstruction of the spread of four VOCs/VOIs across the African continent using sequences showing strict continuous transmission across geographical regions. Curved lines denote the direction of transmission in the anti-clockwise direction. Solid lines show transmission paths as inferred by phylogeographic reconstruction and colored by date, whereas dashed lines show known travel history of the particular case considered.*

Conclusions

Our phylogeographic reconstruction of past viral dissemination patterns suggests a strong epidemiological linkage between Europe and Africa, with 64% of detectable viral imports into Africa originating in Europe and 41% of detectable viral exports from Africa landing in Europe (Fig. 6.1C). This phylogeographic analysis also suggests a changing pattern of viral diffusion into and within Africa over the course of 2020. In almost all instances the earliest introductions of SARS-CoV-2 into individual African countries were from countries outside Africa.

High rates of COVID-19 testing and consistent genomic surveillance in the south of the continent have led to the early identification of VOCs such as B.1.351 and VOIs such as C.1.1 (*13*). Since

the discovery of these southern African variants, several other SARS-CoV-2 VOIs have emerged in different parts of the world, including elsewhere on the African continent, such as B.1.525 in West Africa and A.23.1 in East Africa). There is strong evidence that both of these VOIs are rising in frequency in the regions where they have been detected, which suggests that they may possess higher fitness than other variants in these regions. Although more focused research on the biological properties of these VOIs is needed to confirm whether they should be considered VOCs, it would be prudent to assume the worst and focus on limiting their spread. It will be important to investigate how these different variants compete against one another if they occupy the same region.

Our focused phylogenetic analysis of the B.1.351 lineage revealed that in the final months of 2020 this variant spread from South Africa into neighbouring countries, reaching as far north as the DRC by February 2021. This spread may have been facilitated through rail and road networks that form major transport arteries linking South Africa's ocean ports to commercial and industrial centres in Botswana, Zimbabwe, Zambia and the southern parts of the DRC. The rapid, apparently unimpeded spread of B.1.351 into these countries suggests that current land-border controls that are intended to curb the international spread of the virus are ineffective. Perhaps targeted testing of cross-border travellers, genotyping of positive cases and the focused tracking of frequent cross-border travellers such as long distance truckers, would more effectively contain the spread of future VOCs and VOIs that emerge within this region.

The dominance of VOIs and VOCs in Africa has important implications for vaccine rollouts on the continent. For one, slow rollout of vaccines in most African countries creates an environment in which the virus can replicate and evolve: this will almost certainly produce additional VOCs, any of which could derail the global fight against COVID-19. On the other hand, with the already widespread presence of known variants, difficult decisions balancing reduced efficacy and availability of vaccines have to be made. This also highlights how crucial it is that trials are done. From a public health perspective, genomic surveillance is only one item in the toolkit of pandemic preparedness. It is important that such work is closely followed by genotype to phenotype research to determine the actual significance of continued evolution of SARS-CoV-2 and other emerging pathogens.

The rollout of vaccines across Africa has been painfully slow (Appendix E - Fig. S8 and S9). There have, however, been notable successes that suggest the situation is not hopeless. The small island nation of the Seychelles had vaccinated 70% of its population by May 2021. Morocco has kept pace with many developed nations and by mid-March had vaccinated ~16% of its population. Rwanda, one of Africa's most resource constrained countries, had, within three weeks of obtaining its first vaccine doses in early March, managed to provide first doses to ~2.5% of its population. For all other African countries, at the time of writing, vaccine coverage (first dose) was <1.0% of the general population.

The effectiveness of molecular surveillance as a tool for monitoring pandemics is largely dependent on continuous and consistent sampling through time, rapid virus genome sequencing and rapid reporting. When this is achieved, molecular surveillance can ensure the early detection of changing pandemic characteristics. Further, when such changes are discovered, molecular surveillance data can also guide public health responses. In this regard, the molecular surveillance data that are being gathered by most African countries are less useful than they could be. For example, the time-lag between when virus samples are taken and when sequences for these samples are deposited in sequence repositories is so great in some cases that the primary utility of genomic surveillance data is lost (Appendix E - Fig. S10). This lag is driven by several factors depending on the laboratory or country in question: (*i*) lack of reagents due to disruptions in global supply chains; (*ii*) lack of equipment and infrastructure within the originating country; (*iii*) scarcity of technical skills in laboratory methods or bioinformatic support; and (*iv*) hesitancy by some health officials to release data. More recent sampling and prompt reporting is crucial to reveal the genetic characteristics of currently circulating viruses in these countries.

The patchiness of African genomic surveillance data is therefore the main weakness of our study. However, there is evidence that the situation is improving, with ~50% of African SARS-CoV-2 genome sequences having been submitted to the GISAID database within the first 10-weeks of 2021. While the precise factors underlying this surge in sequencing effort are unclear, important drivers are almost certainly both increased global interest in genomic surveillance following the discovery of multiple VOCs and VOIs since December 2020. We cannot reject that the observed increase in exports from Africa may be due to intensified sequencing activity following the detection of variants around the world. It is important to note here that phylogeographic reconstruction of viral spread is highly dependent on sampling where there is the caveat that the exact routes of viral movements between countries cannot be inferred if there is no sampling in connecting countries. Furthermore, our efforts to reconstruct the movement dynamics of SARS-CoV-2 across the continent are almost certainly biased by uneven sampling between different African countries. It is not a coincidence that we identified South Africa, Kenya and Nigeria, which have sampled and sequenced the most SARS-CoV-2 genomes, as major sources of viral transmissions between sub-Saharan African countries. However, these countries had also the highest number of infections, which may decrease the sampling biases (Fig. 6.1A).

The reliability of genomic surveillance as a tool to prevent the emergence and spread of dangerous variants is dependent on the intensity with which it is embraced by national public health programs. As with most other parts of the world, the success of genomic surveillance in Africa requires more samples being tested for COVID-19, higher proportions of positive samples being sequenced within days of sampling, and persistent analyses of these sequences for concerning signals such as (*i*) the presence of novel non-synonymous mutations at genomic sites associated with pathogenicity and immunogenicity, (*ii*) evidence of positive selection at codon sites where non-

synonymous mutations are observed, and (*iii*) evidence of lineage expansions. In spite of limited sampling, Africa has identified many of the VOCs and VOIs that are being transmitted across the world. Detailed characterization of the variants and their impact on vaccine induced immunity is of extreme importance. If the pandemic is not controlled in Africa, we may see the production of vaccine escape variants that may profoundly affect the population in Africa and across the world.

Acknowledgments

We gratefully acknowledge the authors from the originating laboratories and the submitting laboratories, who generated and shared via GISAID genetic sequence data on which this research is based (Table S4). We also wish to acknowledge the contribution of Kruger Maria from the National Genomics Surveillance of South Africa (NGS-SA) platform for their contribution towards the sequencing effort in Cape Town; South Africa. Similarly, we wish to thank Aya M Elsaame, Shimaa M, Elsayed and Reham M. Darwish from the Faculty of Medicine Ain Shams Research institute (MASRI), for their efforts towards sequencing in Egypt. Sidy Bane, Moumine Sanogo, Dramane Diallo, Antieme Combo Georges Togo and Aminatou Coulibaly from the University Clinical Research Centre (UCRC), at the University of Sciences, Techniques and Technologies of Bamako we wish to extend our thanks for the contribution they have made towards sequencing efforts in Mali. Finally we wish to acknowledge the contribution of Dr Matshidiso Moeti and Dr Abdou Salam Gueye from the World Health Organization for their contribution towards combating SARS-CoV-2 on the African continent.

Funding:

The University of Ghana (WACCBIP) team was funded by a Wellcome/African Academy of Sciences Developing Excellence in Leadership Training and Science (DELTAS) grant (DEL-15-007 and 107755/Z/15/Z: Awandare); National Institute of Health Research (NIHR) (17.63.91) grants using UK aid from the UK Government for a global health research group for Genomic surveillance of malaria in West Africa (Wellcome Sanger Institute, UK) and global research unit for Tackling Infections to Benefit Africa (TIBA partnership, University of Edinburgh); and the World Bank African Centres of Excellent grant (WACCBIP-NCDs: Awandare).

Project ADAGE PRFCOV19-GP2 (2020-2022), which includes 40 researchers from the Center of Biotechnology of Sfax, the University of Sfax, the University of Monastir, the University Hospital Hédi Chaker of Sfax, the Military Hospital of Tunis, and Dacima Consulting. Ministry of Higher Education and Scientific Research and Ministry of Health of the Republic of Tunisia.

The Uganda contributions were funded by the UK Medical Research Council (MRC/UKRI) and the UK Department for International Development (DFID) under the MRC/DFID Concordat agreement (grant agreement number NC_PC_19060) and by the Wellcome, DFID - Wellcome Epidemic Preparedness – Coronavirus (grant agreement number 220977/Z/20/Z) awarded to MC.

Work from Quadram Institute Bioscience was funded by The Biotechnology and Biological Sciences Research Council Institute Strategic Programme Microbes in the Food Chain BB/R012504/1 and its constituent projects BBS/E/F/000PR10348, BBS/E/F/000PR10349, BBS/E/F/000PR10351, and BBS/E/F/000PR10352 and by the Quadram Institute Bioscience BBSRC funded Core Capability Grant (project number BB/CCG1860/1).

The Africa Pathogen Genomics Initiative (Africa PGI) at the Africa CDC is supported by the Bill and Melinda Gates Foundation (INV018978 and INV018278), Illumina Inc, Center for Disease Control and Prevention (CDC), and Oxford Nanopore Technologies. Sequences generated in Zambia through PATH were funded by the Bill & Melinda Gates Foundation. The findings and conclusions contained within are those of the authors and do not necessarily reflect positions or policies of the Bill & Melinda Gates Foundation. Funding for sequencing in Côte d'Ivoire, Burkina Faso and part of the sequencing in the Democratic Republic of the Congo was granted by the German Federal Ministry of Education and Research (BMBF).

Sequencing efforts from Morocco have been supported by Academie Hassan II of Science and Technology, Morocco. Funding for surveillance, sampling and testing in Madagasar: World Health Organization (WHO), the US Centers for Disease Control and Prevention (US CDC: Grant#U5/IP000812-05), the United States Agency for International Development (USAID: Cooperation Agreement 72068719CA00001), the Office of the Assistant Secretary for Preparedness and Response in the U.S. Department of Health and Human Services (DHHS: grant number IDSEP190051-01-0200). Finding for sequencing: Bill & Melinda Gates Foundation (GCE/ID OPP1211841), Chan Zuckerberg Biohub, and the Innovative Genomics Institute at UC Berkeley.

Botswana Harvard AIDS Institute was supported by the following funding: H3ABioNet through funding from the National Institutes of Health Common Fund [U41HG006941]. H3ABioNet is an initiative of the Human Health and Heredity in Africa Consortium (H3Africa) programme of the African Academy of Science (AAS). HHS/NIH/National Institute of Allergy and Infectious Diseases (NIAID) (5K24AI131928-04; 5K24AI131924-04); Sub-Saharan African Network for TB/HIV Research Excellence (SANTHE), a DELTAS Africa Initiative [grant # DEL-15-006]. The DELTAS Africa Initiative is an independent funding scheme of the African Academy of Sciences (AAS)'s Alliance for Accelerating Excellence in Science in Africa (AESA) and supported by the New Partnership for Africa's Development Planning and Coordinating Agency (NEPADAgency)

with funding from the Wellcome Trust [grant #107752/Z/15/Z] and the United Kingdom (UK) government.

South African Medical Research Council (SAMRC) and the Department of Technology and Innovation as part of the Network for Genomic Surveillance in South Africa (NGS-SA) and the Stellenbosch University Faculty of Medicine & Health Sciences, Strategic Equipment Fund". Darren P. Martin is funded by the Wellcome Trust (Wellcome Trust grant number 222574/Z/21/Z). Cathrine Scheepers at the NICD is supported by the National Institute of Allergy and Infectious Diseases of the National Institutes of Health under the Award Number U01AI136677. Furthermore, pandemic surveillance in South Africa and Senegal was supported in part through National Institutes of Health USA grant U01 AI151698 for the United World Antiviral Research Network (UWARN).

Sequencing efforts in the Democratic Republic of the Congo were funded by the Bill & Melinda Gates Foundation under grant INV-018030 awarded to CBP and further supported by funding from the Africa CDC through the ASLM (African Society of Laboratory Medicine) for Accelerating SARS-CoV-2 Genomic Surveillance in Africa.

Sequencing efforts in Rwanda were commissioned by the National Institute of Health Research (NIHR) Global Health Research programme (16/136/33) using UK aid from the UK government(funding to EM and NR through TIBA partnership) and additional funds from the Government of Rwanda through RBC/National Reference Laboratory in collaboration with the Belgian Development Agency (ENABEL) for additional genomic sequencing at GIGA Research Institute- Liege/Belgium.

The sequencing effort in Equatorial Guinea was supported by a public–private partnership, the Bioko Island Malaria Elimination Project, composed of the government of Equatorial Guinea Ministries of Mines and Hydrocarbons, and Health and Social Welfare, Marathon EG Production Limited, Noble Energy, Atlantic Methanol Production Company, and EG LNG.

Samples collection and typing in Mali were supported by Fondation Merieux-France and Sequence efforts has been supported by the by the Enable and Enhance Initiative of the German Federal Government’s Security Cooperation against Biological Threats in the G5 Sahel Region.

The Nigeria work was made possible by support from Flu Lab and a cohort of generous donors through TED’s Audacious Project, including the ELMA Foundation, MacKenzie Scott, the Skoll Foundation, and Open Philanthropy. Further Nigeria funding were supported by grants from the National Institute of Allergy and Infectious Diseases (https://www.niaid.nih.gov), NIH-H3Africa (https://h3africa.org) (U01HG007480 and U54HG007480), the World Bank grant (worldbank.org) (ACE IMPACT project) to Christian Happi.

Author contributions:

Conceptualization: EW, HT†, JGS, JO, JOG, KOD, RAD, RAK, RL, SKT, SM±, TdO

Methodology: ANZ, AR, CA, DPM, DR, EW, HT†, JAE, JG†, JGS, KHY, KOD, LdOM, MAB, MC, MG, MMN, MVTP, PA, TdO, VF

Investigation: AE, AI†, ANZ, AR, AS†, AvG, CAK, CW, DC, DN, DPM, DR, DS, EM, EN, EW, FL, GG, HT†, JAE, JES, JG†, JGS, JJT, JL, JN‡, JQM, KHY, M-MD, MAB, MC, MG, MM†, MM±, MS, MVTP, NA, NHR, NK, PK, RAAC, RAD, RG, SAM, SFA, SM±, SO, TLV, VF, WP

Sampling: A-SK, AA, AAA, AAS, AD, AE, AEOO, AF, AG, AH, AI, AK†, AKS, AKS†, AL, AMO, AO, AP, AR†, AS‡, AV, AVG, AvG, BB, BH, BK, BK†, BM, BN, BO, BT, CA, CD, CP, CS, CW, DB, DC, DD, DG, DGA, DN, DS, EKL, EM, EMO, EN, EP, ES, FA, FA‡, FE, FM, FM†, FO, FT, FT†, GAA, GG, GM, GPM, GT, GvZ, HC, HE, HN, IC, IG, IG†, IK, IM, IO, IS, JA, JCM, JD, JES, JG, JG†, JJL, JK, JL, JMH, JMM, JN, JN†, JTK, KA, KMS, LB, M-MD, MA†, MC, MD, MeH, MGS, MKK, MM§, MM†, MM±, MMD, MN, MO†, MOA, MR, MS, MWM, MY, NA, NG, NH, NHR, NI, NK, NM, NN, NS, NS†, OC†, OEC, OF, OF†, OI, OJ, OK, OO, OP, OS, OT†, P, PB, PC, PCS, PD, PK, PKQ, PO, PS, RAD, RG, RN, SA, SA†, SB, SBL, SD, SE, SeK, SFA, SG†, SK, SL, SLD, SM, SM†, SMM, SN, SP†, SS, ST, TLV, TM, TS, UC, UG, UJ, UR, VG, WA, WC, WP, WR, YB, YKT, YN, ZRD

Sequencing: A-SK, AA, AAA, AAS, AC, AD, AEOO, AF, AI, AI‡, AK, AK†, AKK, AKS, AKS†, AL, ANZ, AP, AS, AS†, AS‡, ASO, AT, AV, AVG, AvG, AY, BB, BD, BH, BK, BK†, BM†, BN, BT, CA, CB, CBP, CD, CMM, CP, CS, DB, DD, DG, DGA, DJB, DLB, DM, DOO, DP, DSYA, DT, EF, EFN, EKL, EL, EMO, EP, ES, ES†, ESL, FA, FA†, FAD, FD, FM, FM†, FO, FT†, FW, GAA, GG, GPM, GT, GvZ, HA, HA†, HC, HCR, HE, HG†, HK, HN, IB, IC, IG, IG†, IK, IM, IS, JA, JB, JCM, JD, JF, JG, JG†, JJL, JK, JMH, JMM, JMN, JN, JN†, JQM, JTK, JY, KA, KMS, KOD, KS, KT, LB, LF, LS, LT, M-MD, MA†, MAB, MC, MD, MeH, MGS, MIM, MKK, MM, MM§, MM‡, MMD, MMN, MO, MO†, MOA, MVTP, MWM, MY, ND, NG, NH, NI, NI†, NM, NN, NN†, NS, NS†, NT, OC, OC†, OEC, OF, OI, OJ, OT†, PA, PB, PCS, PD, PEO, PK, PKQ, PM, PO, PS, RAAC, RG, RN, ROP, SA, SA†, SB, SBL, SCS, SD, SE, SE, SeK, SG, SG†, SHA, SK†, SL, SLD, SM, SM†, SM±, SMM, SN, SP†, SP‡, SR, SS, ST, ST†, SvdW, TA, TM, TM†, TS, UC, UG, UJ, UJA, UR, VG, WA, WC, WR, YB, YB†, YKT, YN, ZRD

Visualization: AC, AI‡, AK, AKK, AS, AS†, AY, BT, CB, CMM, DB†, DOO, DP, DR, DSYA, EA, EB, ESL, EW, FAD, FB, FD, FW, GS, HA, HA†, HG†, HL, HT†, IA, IS, JAE, JB, JF, JG†, JMN, JY, KHY, KS, LF, LS, LT, MA, MA†, MG, MT, MVTP, MW, ND, NI†, NK, NN†, NT, OC, OT, PA, PCS, PEO, RAAC, SB, SFS, SHA, SK†, SM±, TA, TM†, VE, YB†

Funding acquisition: AJP, AR, AvG, BK, CA, CAK, CBP, CW, DC, DJB, DN, FL, GAA, GG, GPM, HC, JES, JJT, JL, JMH, JN‡, JO, KOD, M-MD, MC, MIM, MM±, MVTP, NA, PCS, PK, PM, RAK, SAM, SE, SM†, SvdW, TdO, WP

Project administration: AJP, AR, AV†, AvG, BK, CW, DJB, DN, EW, FA†, FT, GAA, GPM, GS, GT, HC, JCO, JJT, JMH, JO, JOG, JY, KOD, MC, MK, MM†, MP, MVTP, MW, NR, OT, PCS, PK, PM, RAK, SAM, SE, SFS, SG†, SM†, TdO

Supervision: AJP, AR, BK, CW, DN, EN, EW, FT, GAA, GK, HC, JB, JMH, JN‡, JO, JOG, KOD, MA†, MC, MIM, MM†, MMN, MS, NM†, NR, PCS, PK, PM, RAK, SE, SeK, SG†, SM, SM†, SP, TdO

Writing – original draft: AKS, ANZ, BK, DPM, EW, FT, GK, HT†, JB, JCM, MA†, MAB, MC, MG, MM, NM†, RL

Writing – review & editing: ANZ, BK, CMM, DN, DPM, DR, DSYA, DT, EKL, EL, ESL, EW, HT†, JES, JGS, LdOM, MAB, MC, MeH, PKQ, PM, RL, SKT, TdO, UJA

*Author contribution listed alphabetically. A full list of author abbreviations included on the GitHub repo (https://github.com/krisp-kwazulu-natal/africa-covid19-genomics).

Competing interests: Dr. Paradis Sabeti is a founder and shareholder of Sherlock biosciences, and is both on the Board and serves as shareholder of the Danaher Corporation. We the authors of the manuscript, to the best of our knowledge, declare no other conflicts of interest.

Data and materials availability: All sequences that were used in the present study are listed in Supplementary Table S4 (accessible on the GitHub repository) along with their GISAID sequence IDs, dates of sampling, the originating and submitting laboratories and main authors. All input files (e.g. alignments or XML files), all resulting output files and scripts used in the study are shared publicly on GitHub (https://github.com/krisp-kwazulu-natal/africa-covid19-genomics) (*23*).

References

1. C. Wang, P. W. Horby, F. G. Hayden, G. F. Gao, A novel coronavirus outbreak of global health concern. *Lancet*. **395**, 470–473 (2020).
2. Q. Li, X. Guan, P. Wu, X. Wang, L. Zhou, Y. Tong, R. Ren, K. S. M. Leung, E. H. Y. Lau, J. Y. Wong, X. Xing, N. Xiang, Y. Wu, C. Li, Q. Chen, D. Li, T. Liu, J. Zhao, M. Liu, W. Tu, Z. Feng, Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus-Infected Pneumonia. *N. Engl. J. Med.* **382**, 1199–1207 (2020).
3. S. Uyoga, I. M. O. Adetifa, H. K. Karanja, J. Nyagwange, J. Tuju, P. Wanjiku, R. Aman, M. Mwangangi, P. Amoth, K. Kasera, W. Ng'ang'a, C. Rombo, C. Yegon, K. Kithi, E. Odhiambo, T. Rotich, I. Orgut, S. Kihara, M. Otiende, C. Bottomley, G. M.

Warimwe, Seroprevalence of anti-SARS-CoV-2 IgG antibodies in Kenyan blood donors. *Science*. **371**, 79–82 (2021).

4. L. Mwananyanda, C. J. Gill, W. MacLeod, G. Kwenda, R. Pieciak, Z. Mupila, R. Lapidot, F. Mupeta, L. Forman, L. Ziko, L. Etter, D. Thea, Covid-19 deaths in Africa: prospective systematic postmortem surveillance study. *BMJ*. **372**, n334 (2021).

5. S. J. Salyer, J. Maeda, S. Sembuche, Y. Kebede, A. Tshangela, M. Moussif, C. Ihekweazu, N. Mayet, E. Abate, A. O. Ouma, J. Nkengasong, The first and second waves of the COVID-19 pandemic in Africa: a cross-sectional study. *Lancet*. **397**, 1265–1275 (2021).

6. P. Oluniyi, First African SARS-CoV-2 genome sequence from Nigerian COVID-19 case. *Virological* (2020).

7. M. A. Medhat, M. El Kassas, COVID-19 in Egypt: Uncovered figures or a different situation? *J. Glob. Health*. **10**, 010368 (2020).

8. M. Allam, A. Ismail, Z. T. H. Khumalo, S. Kwenda, P. van Heusden, R. Cloete, C. K. Wibmer, P. S. Mtshali, F. Mnyameni, T. Mohale, K. Subramoney, S. Walaza, W. Ngubane, N. Govender, N. V. Motaze, J. N. Bhiman, SA-COVID-19 response team, Genome Sequencing of a Severe Acute Respiratory Syndrome Coronavirus 2 Isolate Obtained from a South African Patient with Coronavirus Disease 2019. *Microbiol. Resour. Announc.* **9** (2020), doi:10.1128/MRA.00572-20.

9. N. Haider, A. Y. Osman, A. Gadzekpo, G. O. Akipede, D. Asogun, R. Ansumana, R. J. Lessells, P. Khan, M. M. A. Hamid, D. Yeboah-Manu, L. Mboera, E. H. Shayo, B. T. Mmbaga, M. Urassa, D. Musoke, N. Kapata, R. A. Ferrand, P.-C. Kapata, F. Stigler, T. Czypionka, D. McCoy, Lockdown measures in response to COVID-19 in nine sub-Saharan African countries. *BMJ Glob Health*. **5** (2020), doi:10.1136/bmjgh-2020-003319.

10. S. C. Inzaule, S. K. Tessema, Y. Kebede, A. E. Ogwell Ouma, J. N. Nkengasong, Genomic-informed pathogen surveillance in Africa: opportunities and challenges. *Lancet Infect. Dis.* (2021), doi:10.1016/S1473-3099(20)30939-7.

11. D. Lule Bugembe, M. V. T. Phan, I. Ssewanyana, P. Semanda, H. Nansumba, B. Dhaala, S. Nabadda, A. O'Toole, A. Rambaut, P. Kaleebu, M. Cotten, A SARS-CoV-2 lineage A variant (A.23.1) with altered spike has emerged and is dominating the current Uganda epidemic. *medRxiv* (2021), doi:10.1101/2021.02.08.21251393.

12. J. Giandhari, S. Pillay, E. Wilkinson, H. Tegally, I. Sinayskiy, M. Schuld, J. Lourenço, B. Chimukangara, R. Lessells, Y. Moosa, I. Gazy, M. Fish, L. Singh, K. Sedwell Khanyile, V. Fonseca, M. Giovanetti, L. Carlos Junior Alcantara, F. Petruccione, T. de Oliveira, Early transmission of SARS-CoV-2 in South Africa: An epidemiological and phylogenetic report. *Int. J. Infect. Dis.* **103**, 234–241 (2021).

13. H. Tegally, E. Wilkinson, R. J. Lessells, J. Giandhari, S. Pillay, N. Msomi, K. Mlisana, J. N. Bhiman, A. von Gottberg, S. Walaza, V. Fonseca, M. Allam, A. Ismail, A. J. Glass, S. Engelbrecht, G. Van Zyl, W. Preiser, C. Williamson, F. Petruccione, A. Sigal, T. de Oliveira, Sixteen novel lineages of SARS-CoV-2 in South Africa. *Nat. Med.* **27**, 440–446 (2021).

14. C. A. Pearson, T. W. Russell, N. Davies, A. J. Kucharski, Estimates of severity and transmissibility of novel SARS-CoV-2 variant 501Y.V2 in South Africa | CMMID Repository (2021), (available at https://cmmid.github.io/topics/covid19/sa-novel-variant.html).

15. S. Cele, I. Gazy, L. Jackson, S.-H. Hwa, H. Tegally, G. Lustig, J. Giandhari, S. Pillay, E. Wilkinson, Y. Naidoo, F. Karim, Y. Ganga, K. Khan, A. B. Balazs, B. I. Gosnell, W. Hanekom, M.-Y. S. Moosa, R. J. Lessells, T. de Oliveira, A. Sigal, Escape of SARS-CoV-2 501Y.V2 variants from neutralization by convalescent plasma. *medRxiv* (2021), doi:10.1101/2021.01.26.21250224.

16. S. A. Madhi, V. Baillie, C. L. Cutland, M. Voysey, A. L. Koen, L. Fairlie, S. D. Padayachee, K. Dheda, S. L. Barnabas, Q. E. Bhorat, C. Briner, G. Kwatra, K. Ahmed, P. Aley, S. Bhikha, J. N. Bhiman, A. E. Bhorat, J. du Plessis, A. Esmail, M. Groenewald, Wits-VIDA COVID Group, Efficacy of the ChAdOx1 nCoV-19 Covid-19 Vaccine against the B.1.351 Variant. *N. Engl. J. Med.* **384**, 1885–1898 (2021).

17. C. K. Wibmer, F. Ayres, T. Hermanus, M. Madzivhandila, P. Kgagudi, B. Oosthuysen, B. E. Lambson, T. de Oliveira, M. Vermeulen, K. van der Berg, T. Rossouw, M. Boswell, V. Ueckermann, S. Meiring, A. von Gottberg, C. Cohen, L. Morris, J. N. Bhiman, P. L. Moore, SARS-CoV-2 501Y.V2 escapes neutralization by South African COVID-19 donor plasma. *Nat. Med.* **27**, 622–625 (2021).

18. Y. Shu, J. McCauley, GISAID: Global initiative on sharing all influenza data - from vision to reality. *Euro Surveill.* **22**, 30494 (2017).

19. E. Volz, V. Hill, J. T. McCrone, A. Price, D. Jorgensen, Á. O'Toole, J. Southgate, R. Johnson, B. Jackson, F. F. Nascimento, S. M. Rey, S. M. Nicholls, R. M. Colquhoun, A. da Silva Filipe, J. Shepherd, D. J. Pascall, R. Shah, N. Jesudason, K. Li, R. Jarrett, T. R. Connor, Evaluating the Effects of SARS-CoV-2 Spike Mutation D614G on Transmissibility and Pathogenicity. *Cell*. **184**, 64-75.e11 (2021).

20. B. Korber, W. M. Fischer, S. Gnanakaran, H. Yoon, J. Theiler, W. Abfalterer, N. Hengartner, E. E. Giorgi, T. Bhattacharya, B. Foley, K. M. Hastie, M. D. Parker, D. G. Partridge, C. M. Evans, T. M. Freeman, T. I. de Silva, Sheffield COVID-19 Genomics Group, C. McDanal, L. G. Perez, H. Tang, D. C. Montefiori, Tracking Changes in SARS-CoV-2 Spike: Evidence that D614G Increases Infectivity of the COVID-19 Virus. *Cell*. **182**, 812-827.e19 (2020).

21. T. de Oliveira, S. Lutucuta, J. Nkengasong, J. Morais, J. Paula Paixao, Z. Neto, P. Afonso, J. Miranda, K. David, L. Ingles, A. P. A. P. R. Carralero, H. R. Freitas, F. Mufinda, S. K. Tessema, H. Tegally, E. J. San, E. Wilkinson, J. Giandhari, S. Pillay, M. Giovanetti, R. J. Lessells, A novel variant of interest of SARS-CoV-2 with multiple spike mutations is identified from travel surveillance in Africa. *medRxiv* (2021), doi:10.1101/2021.03.30.21254323.

22. S. A. Kemp, R. P. Datir, D. A. Collier, I. Ferreira, A. Carabelli, W. Harvey, D. L. Robertson, R. K. Gupta, Recurrent emergence and transmission of a SARS-CoV-2 Spike deletion ΔH69/ΔV70. *BioRxiv* (2020), doi:10.1101/2020.12.14.422555.

23. S. E. James, HouriiyahT, krisp-kwazulu-natal/africa-covid19-genomics: A year of genomic surveillance reveals how the SARS-CoV-2 pandemic unfolded in Africa - Code and Scripts. *Zenodo* (2021), doi:10.5281/zenodo.5386379.

Methods

Ethics statement

This project relied on sequence data and associated metadata publicly shared by the GISAID data repository and adhere to the term and conditions laid out by GISAID. The African samples processed in this study were obtained anonymously from material exceeding the routine diagnosis of SARS-CoV-2 in African public health laboratories that belong to the public network within the Africa CDC. Individual institutional review board (IRB) references or material transfer agreements (MTAs) for countries are list below.

Angola - (MTA - CON8260), Botswana - Genomic surveillance in Botswana was approved by the Health Research and Development Committee (Protocol HPDME 13/18/1), Nigeria – (NHREC/01/01/2007), Mali - study of the sequence of SARS-CoV-2 isolates in Mali - Letter of Ethical Committee (N0-2020 /201/CE/FMPOS/FAPH of 09/17/2020), Mozambique - (MTA - CON7800), Malawi - (MTA - CON8265), South Africa - The use of South African samples for sequencing and genomic surveillance were approved by University of KwaZulu-Natal Biomedical Research Ethics Committee (ref. BREC/00001510/2020); the University of the Witwatersrand Human Research Ethics Committee (HREC) (ref. M180832); Stellenbosch University HREC (ref. N20/04/008_COVID-19); and the University of Cape Town HREC (ref. 383/2020), Tunisia - For sequences derived from sampling in Tunisia, all patients provided their informed consent to use their samples for sequencing of the viral genomes. The ethical agreement was provided to the research project ADAGE (PRFCOVID19GP2) by the Committee of protection of persons (Tunisian Ministry of Health) under the reference (CPP SUD N 0265/2020), Uganda - The use of samples and sequences from Uganda were approved by the Uganda Virus Research Institute - Research and Ethics Committee UVRI-REC Federalwide Assurance [FWA] FWA No. 00001354, study reference - GC/127/20/04/771 and by the Uganda National Council for Science and Technology, reference number - HS936ES) and Zimbabwe (MTA - CON8271).

Data quality control

10326 African complete and near-complete genome sequences were retrieved from GISAID on 16 March 2021 (2pm SAST). Sampling strategies in various participating countries are outlined in Supplementary Table S3. Prior to phylogenetic reconstruction we removed low quality sequences, which included those identified as being of low quality by NextClade (n=18; https://clades.nextstrain.org), those with missing sampling dates (n = 189), those with <90% coverage (n = 1017), those with > 40 SNPs (n = 39), those with >10 ambiguous base-calls per genome (n = 128), and those with clustered SNPs (n = 189).

High quality African near-complete genome sequences (n=8,746) were aligned against an extensive reference dataset of 11891 SARS-CoV-2 sequences from around the world that included sequences sampled since the start of the outbreak, including all those sampled up until the end of February 2020.

Phylogenetic reconstruction

The African sequences were aligned against the reference panel using MAFFT v7.471(*24*). The first 100 and last 50 bases as well as positions 13402, 24389 and 24390 relative to the reference strain Wuhan-Hu-1 (*18,(25)* were masked as these three sites are known for primer contamination resulting in ambiguity. The subsequent alignment was used to infer a maximum likelihood (ML) phylogenetic tree in IQTREE v1.6.9(*26*). The tree was inferred with the general time reversible (GTR) model of nucleotide substitution and a proportion of invariable sites (+I). To infer some confidence measures of branches in the phylogeny and for subsequent downstream analyses we performed 100 bootstrap replicates using Booster(*27*).

The raw ML tree topology was used to estimate the number of viral transmission events between various Africa countries and the rest of the world. TreeTime(*28*) was used to transform this ML tree topology into a dated tree using a constant rate of 8.0×10^{-4} nucleotide substitutions per site per year , after the exclusion of outlier sequences. A migration model was fitted to the resulting time-scaled phylogenetic tree in TreeTime, mapping country and regional locations to tips and internal nodes. Using the resulting annotated tree topology we could count the number of transitions between Africa and the rest of the world.

Lineage classification

We used the dynamic lineage classification method called Phylogenetic Assignment of Named Global Outbreak LINeages (PANGOLIN)(*29*). This was aimed at identifying the most epidemiologically important lineages of SARS-CoV-2 circulating within the African continent and to identify the lineage dynamics within African regions and across the continent. For the purpose of clarity, we define a lineage as a linear chain of viruses in a phylogenetic tree showing connection from the ancestor to the most recent descendant. A unique variant refers to a genetically distinct

virus with different mutations to other viruses of the same lineage. Variants of concern (VOC) and variants of interest (VOI) were designated based on the World Health Organization framework as of 13 April 2021. We included two other lineages, namely A.23.1 and C.1.1, and designated them as VOI for the purposes of this analysis. We included these two as they demonstrated continued evolution of African lineages into potentially more transmissible variants with the acquisition of mutations in the spike glycoprotein.

Phylogeographic reconstruction

VOCs and VOIs that emerged on the African continent (B.1.351, B.1.525, A.23.1 and C.1.1) were marked on the time-resolved phylogenetic tree constructed above. Genome sequences from these four lineages were extracted for phylogeographic reconstruction. First, we investigated the dynamics of SARS-CoV-2 infection and virus lineage movements over longer distances (through Europe or East to West Africa) using a sampled set of time-scaled phylogenies and the sampling location of each geo-referenced SARS-CoV-2 sequence. We discretized sequence sampling locations by considering distinct geographic areas and/or regions (in and outside Africa) as shown in Appendix E - Supplementary Figure S6.

Initially, discrete phylogeographic reconstructions were conducted for all VOC and VOI using the asymmetric discrete trait model implemented in BEASTv1.10.4(*30*). From those estimates we then modelled the phylogenetic diffusion and spread of the lineages on the African continent by analysing localized transmission (between neighbouring countries) using a flexible relaxed random walk (RRW) diffusion model(*31*) that accommodates branch-specific variation in rates of dispersal with a Cauchy distribution. For each sequence, latitude and longitude coordinates were attributed to the lowest administrative level locator in GISAID.

Multiple sequence alignments were performed for each lineage with MAFFT v7.471. Maximum likelihood trees for each of the alignments were inferred in IQTREE v1.6.9 (GTR+I). Prior to phylogeographic reconstruction each cluster/lineage was assessed for molecular clock signal in TempEst v1.5.3(*32*) following the removal of potential outliers that may violate the molecular clock assumption. Markov Chain Monte Carlo (MCMC) analyses were set up in BEAST v1.10.4 in duplicate for 100 million interactions and sampling every 10000 steps in the chain. Convergence for each run was assessed in Tracer v1.7.1 (ESS for all relevant model parameters >200). Maximum clade credibility trees for each run were summarized using TreeAnnotator after discarding the initial 10% as burn-in. We used the R package “seraphim”(*33*) to extract and map spatiotemporal information embedded in the posterior trees. Note that a transmission link on the phylogeographic map can denote one or more transmission events depending on the phylogeographic inference.

Sensitivity of introduction analysis to sampling biases

Three sensitivity analyses were performed to examine how robust the main results of our introduction analysis were to known biases in sampling across space and time. For our first analysis, we randomly selected 10 of the bootstrap tree topologies that was inferred using Booster for discrete state ancestral state reconstruction as described earlier. The average number of imports and exports between Africa and the rest of the world per week were then plotted overtime along with the standard error for each discrete time point.

In the second, we performed a rarefaction analysis to determine how the number of introductions into Africa varies depending on the extent of sampling in African (internal) and non-African (external) countries. Rarefaction was performed by starting with the full set of samples and subsampling a random subset of samples from the full set at sampling fractions varying from 0.1 to 1.0. Subsampling was performed 10 times at each sampling fraction to create replicate datasets, which were used to place confidence internals on the number of introductions identified at each subsampling fraction.

Because it would have been too computationally intensive to reconstruct phylogenies *de novo* from each subsampled dataset, we adopted a subsample-then-prune approach(*34*). For each subsampled dataset, samples not included in the subsampled set were pruned from the full ML phylogeny using the *extract_tree_with_taxa* function in DendroPy version 4.5.1(*35*). Ancestral locations were then reconstructed for internal nodes in each subsampled or pruned tree using maximum parsimony(*36*). The total number of introductions into Africa was then computed based on the number of branches in the tree in which the parent node was reconstructed to be external and the child node was reconstructed to be in Africa.

The second analysis was performed to determine how sensitive the temporal distribution of introduction events was to uneven sampling through time. Perhaps most importantly, we sought to determine if the increasing proportion of introductions estimated to be from other African countries through time was an artefact of increased sampling effort during late 2020 and early 2021. To obtain a more uniform temporal distribution of sampling times, we capped the number of samples from Africa each month at a maximum threshold (n=400) and then randomly down-sampled to this threshold count in months that exceeded the threshold. As in the rarefaction analysis, samples excluded after subsampling were pruned from the ML tree after which ancestral states were reconstructed by maximum parsimony.

Epidemiological modelling

Data on regional trade of all imported and exported goods between South Africa and other Eastern and Southern African countries during 2020 was extracted from the United Nations Comtrade Database(*37*), which records trade statistics for more than 5,000 commodity groups by the Harmonized System. Data for cumulative COVID-19 cases and related deaths, vaccinated people,

and cumulative numbers of COVID-19 tests performed by March 30, 2021 were obtained from the Johns Hopkins University database(*38*). Country level maps of each variable were created using ArcGIS® by ESRI version 10.5 (http://www.esri.com).

Supplementary References

24. K. Katoh, D. M. Standley, MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).

25. F. Wu, S. Zhao, B. Yu, Y.-M. Chen, W. Wang, Z.-G. Song, Y. Hu, Z.-W. Tao, J.-H. Tian, Y.-Y. Pei, M.-L. Yuan, Y.-L. Zhang, F.-H. Dai, Y. Liu, Q.-M. Wang, J.-J. Zheng, L. Xu, E. C. Holmes, Y.-Z. Zhang, A new coronavirus associated with human respiratory disease in China. *Nature*. **579**, 265–269 (2020).

26. L.-T. Nguyen, H. A. Schmidt, A. von Haeseler, B. Q. Minh, IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).

27. F. Lemoine, J. B. Domelevo Entfellner, E. Wilkinson, D. Correia, M. Dávila Felipe, T. De Oliveira, O. Gascuel, Renewing Felsenstein's phylogenetic bootstrap in the era of big data. *Nature*. **556**, 452–456 (2018).

28. P. Sagulenko, V. Puller, R. A. Neher, TreeTime: Maximum-likelihood phylodynamic analysis. *Virus Evol.* **4**, vex042 (2018).

29. A. Rambaut, E. C. Holmes, V. Hill, A. OToole, J. McCrone, C. Ruis, L. du Plessis, O. Pybus, A dynamic nomenclature proposal for SARS-CoV-2 to assist genomic epidemiology. *BioRxiv* (2020), doi:10.1101/2020.04.17.046086.

30. A. J. Drummond, M. A. Suchard, D. Xie, A. Rambaut, Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol. Biol. Evol.* **29**, 1969–1973 (2012).

31. P. Lemey, A. Rambaut, J. J. Welch, M. A. Suchard, Phylogeography takes a relaxed random walk in continuous space and time. *Mol. Biol. Evol.* **27**, 1877–1885 (2010).

32. A. Rambaut, T. T. Lam, L. Max Carvalho, O. G. Pybus, Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus Evol.* **2**, vew007 (2016).

33. S. Dellicour, R. Rose, N. R. Faria, P. Lemey, O. G. Pybus, SERAPHIM: studying environmental rasters and phylogenetically informed movements. *Bioinformatics*. **32**, 3204–3206 (2016).

34. E. Wilkinson, D. Rasmussen, O. Ratmann, T. Stadler, S. Engelbrecht, T. de Oliveira, Origin, imports and exports of HIV-1 subtype C in South Africa: A historical perspective. *Infect. Genet. Evol.* **46**, 200–208 (2016).

35. J. Sukumaran, M. T. Holder, DendroPy: a Python library for phylogenetic computing. *Bioinformatics*. **26**, 1569–1571 (2010).

36. D. Sankoff, Minimal mutation trees of sequences. *SIAM J. Appl. Math.* **28**, 35–42 (1975).

37. UN Comtrade | International Trade Statistics Database, (available at https://comtrade.un.org/).

38. GitHub - CSSEGISandData/COVID-19: Novel Coronavirus (COVID-19) Cases, provided by JHU CSSE, (available at https://github.com/CSSEGISandData/COVID-19).

Chapter 7: The evolving SARS-CoV-2 epidemic in Africa: Insights from rapidly expanding genomic surveillance

Two years into the SARS-CoV-2 pandemic, the number of genomes sequenced for African countries had increased 10-fold, compared to the first year described in the Chapter 6. Chapter 7, a study where I am the first-author, shows how the sequencing and surveillance capacities in Africa have expanded during the pandemic, and how the coordinated efforts of many African institutions and public health actors have in a short time made great contributions to pandemic surveillance and data gathering. My contributions to this paper began from curating the genomic and epidemiological datasets for analysing the epidemic in Africa in context of the world, followed by annotation of genomic data by sequencing locations and calculating sequencing turn-around times for all genomes. The results show that sequencing locally produces genomes with a much lower turn-around time compared to sending specimens abroad for sequencing. This is a crucial take-away as capacity is built for genomic surveillance of SARS-CoV-2 and other pathogens across the continent for maximum public health benefit. I was also a key contributor to the phylogeographic methods employed in this study, which allows the spatiotemporal mapping of important viral lineages and shows for the first time how the dispersal of Delta and Omicron in Africa were distinct to Alpha and Beta before them. As with previous chapters, I led the data visualisation for this project and provided major contributions to the manuscript writing.

This chapter was published as a peer-reviewed research article in Science in September 2022 and can be accessed at the following DOI: 10.1126/science.abq5358. The chapter is presented in a similar format as the journal article. A full list of authors and affiliations is shown below.

Houriiyah Tegally[1,2]†, James E. San[1,2], Matthew Cotten[3,4], Monika Moir[1], Bryan Tegomoh[5,6], Gerald Mboowa[7], Darren P. Martin[8,9], Cheryl Baxter[1,10], Arnold W. Lambisia[11], Amadou Diallo[12], Daniel G. Amoako[13,14], Moussa M. Diagne[12], Abay Sisay[15,16], Abdel-Rahman N. Zekri[17], Abdou Salam Gueye[18], Abdoul K. Sangare[19], Abdoul-Salam Ouedraogo[20], Abdourahmane Sow[21], Abdualmoniem O. Musa[22,23,24], Abdul K. Sesay[25], Abe G. Abias[26], Adam I. Elzagheid[27], Adamou Lagare[28], Adedotun-Sulaiman Kemi[29], Aden Elmi Abar[30], Adeniji A. Johnson[31,32], Adeola Fowotade[33,34], Adeyemi O. Oluwapelumi[35,36], Adrienne A. Amuri[37,38], Agnes Juru[39], Ahmed Kandeil[40], Ahmed Mostafa[40], Ahmed Rebai[41], Ahmed Sayed[42], Akano Kazeem[43,44], Aladje Balde[45,46], Alan Christoffels[7,47], Alexander J. Trotter[48], Allan Campbell[49], Alpha K. Keita[50,51], Amadou Kone[52], Amal Bouzid[41,53], Amal Souissi[41], Ambrose Agweyu[11], Amel Naguib[54], Ana V. Gutierrez[48], Anatole Nkeshimana[55], Andrew J. Page[48], Anges Yadouleton[56], Anika Vinze[57], Anise N. Happi[43], Anissa Chouikha[58,59], Arash Iranzadeh[8,9], Arisha Maharaj[1], Armel L. Batchi-Bouyou[60,61], Arshad Ismail[13], Augustina A. Sylverken[62,63], Augustine

Goba[64,65], Ayoade Femi[43,44], Ayotunde E. Sijuwola[43], Baba Marycelin[66,67], Babatunde L. Salako[29,32], Bamidele S. Oderinde[66], Bankole Bolajoko[43], Bassirou Diarra[52], Belinda L. Herring[18], Benjamin Tsofa[11], Bernard Lekana-Douki[68,69], Bernard Mvula[70], Berthe-Marie Njanpop-Lafourcade[18], Blessing T. Marondera[71], Bouh Abdi Khaireh[72,73], Bourema Kouriba[19], Bright Adu[74], Brigitte Pool[75], Bronwyn McInnis[17], Cara Brook[76,77], Carolyn Williamson[9,10,78], Cassien Nduwimana[55], Catherine Anscombe[79,80], Catherine B. Pratt[81], Cathrine Scheepers[13,82], Chantal G. Akoua-Koffi[83,84], Charles N. Agoti[11,85], Chastel M. Mapanguy[60,86], Cheikh Loucoubar[12], Chika K. Onwuamah[87], Chikwe Ihekweazu [88], Christian N. Malaka[89], Christophe Peyrefitte[12], Chukwa Grace[43,44], Chukwuma E. Omoruyi[33,34], Clotaire D. Rafaï[90], Collins M. Morang'a[91], Cyril Erameh [92], Daniel B. Lule[3], Daniel J. Bridges[93], Daniel Mukadi-Bamuleka[37], Danny Park[57], David A. Rasmussen[94,95], David Baker[48], David J. Nokes[11,96], Deogratius Ssemwanga[3,97], Derek Tshiabuila[2], Dominic S.Y. Amuzu[91], Dominique Goedhals[98], Donald S. Grant[64,65,99], Donwilliams O. Omuoyo[11], Dorcas Maruapula[100], Dorcas W. Wanjohi[7], Ebenezer Foster-Nyarko[48], Eddy K. Lusamaki[37,38,51], Edgar Simulundu[101], Edidah M. Ong'era[11], Edith N. Ngabana[37,38], Edward O. Abworo[102], Edward Otieno[11], Edwin Shumba[71], Edwine Barasa[11], El Bara Ahmed[103,104], Elhadi A. Ahmed[23], Emmanuel Lokilo[37], Enatha Mukantwari[105], Eromon Philomena[43], Essia Belarbi[106], Etienne Simon-Loriere[107], Etilé A. Anoh[83], Eusebio Manuel[108], Fabian Leendertz[106], Fahn M. Taweh[109], Fares Wasfi[58], Fatma Abdelmoula[41,110], Faustinos T. Takawira[39], Fawzi Derrar[111], Fehintola V Ajogbasile[43], Florette Treurnicht[112,113], Folarin Onikepe[43,44], Francine Ntoumi[60,114], Francisca M. Muyembe[37,38], Frank E.Z. Ragomzingba[115], Fred A. Dratibi[116,117], Fred-Akintunwa Iyanu[43], Gabriel K. Mbunsu[38], Gaetan Thilliez[48], Gemma L. Kay[48], George O. Akpede[92], Gert U. van Zyl[118,119], Gordon A. Awandare[91], Grace S. Kpeli[120,121], Grit Schubert[106], Gugu P. Maphalala[122], Hafaliana C. Ranaivoson[77], Hannah E Omunakwe[123], Harris Onywera[7], Haruka Abe[124], Hela Karray[125], Hellen Nansumba[126], Henda Triki[58], Herve Albéric Adje Kadjo[127], Hesham Elgahzaly[128], Hlanai Gumbo[39], Hota Mathieu[129], Hugo Kavunga-Membo[37], Ibtihel Smeti[41], Idowu B. Olawoye[43], Ifedayo M.O. Adetifa[88,130], Ikponmwosa Odia[92], Ilhem Boutiba-Ben Boubaker[131,132], Iluoreh Ahmed Mohammad[43], Isaac Ssewanyana[126], Isatta Wurie[133], Iyaloo S. Konstantinus[134], Jacqueline Wemboo Afiwa Halatoko[135], James Ayei[26], Janaki Sonoo[136], Jean-Claude C. Makangara[37,38], Jean-Jacques M. Tamfum[37,38], Jean-Michel Heraud[12,77], Jeffrey G. Shaffer[137], Jennifer Giandhari[2], Jennifer Musyoki[11], Jerome Nkurunziza[138], Jessica N. Uwanibe[43], Jinal N. Bhiman[13,113], Jiro Yasuda[124], Joana Morais[139,140], Jocelyn Kiconco[97], John D. Sandi[64,65], John Huddleston[141], John K. Odoom [74], John M. Morobe[11], John O. Gyapong[120], John T. Kayiwa[3], Johnson C. Okolie[43], Joicymara S. Xavier[1,142,143], Jones Gyamfi[120], Joseph F. Wamala[144], Joseph H.K. Bonney[74], Joseph Nyandwi[55,145], Josie Everatt[13], Joweria Nakaseegu[97], Joyce M. Ngoi[91], Joyce Namulondo[97], Judith U. Oguzie[43,44], Julia C. Andeko[68], Julius J. Lutwama[3], Juma J.H. Mogga[144], Justin O'Grady[48], Katherine J. Siddle[57], Kathleen Victoir[146], Kayode T. Adeyemi[43,44], Kefentse A. Tumedi[147], Kevin S. Carvalho[148], Khadija Said Mohammed[11], Koussay Dellagi[146], Kunda G. Musonda[149], Kwabena O. Duedu[120,121], Lamia Fki-Berrajah[125], Lavanya Singh[2], Lenora M. Kepler[94,95], Leon Biscornet[75], Leonardo de Oliveira Martins[48], Lucious Chabuka[150], Luicer

Olubayo[8], Lul Deng Ojok[26], Lul Lojok Deng[26], Lynette I. Ochola-Oyier[11], Lynn Tyers[9], Madisa Mine[151], Magalutcheemee Ramuth[136], Maha Mastouri[152,153], Mahmoud ElHefnawi[154], Maimouna Mbanne[12], Maitshwarelo I. Matsheka[147], Malebogo Kebabonye[155], Mamadou Diop[12], Mambu Momoh[64,65,156], Maria da Luz Lima Mendonça[148], Marietjie Venter[157], Marietou F Paye[57], Martin Faye[12], Martin M. Nyaga[158], Mathabo Mareka[159], Matoke-Muhia Damaris[160], Maureen W. Mburu[11], Maximillian G. Mpina[161,162,163], Michael Owusu[164], Michael R. Wiley[81,165], Mirabeau Y. Tatfeng[166], Mitoha Ondo'o Ayekaba[162], Mohamed Abouelhoda[167,168], Mohamed Amine Beloufa[111], Mohamed G. Seadawy[169,170], Mohamed K. Khalifa[171], Mooko Marethabile Matobo[159], Mouhamed Kane[12], Mounerou Salou[172], Mphaphi B. Mbulawa[155], Mulenga Mwenda[93], Mushal Allam[173], My V.T. Phan[3], Nabil Abid[152,174], Nadine Rujeni[175,176], Nadir Abuzaid[177], Nalia Ismael[178], Nancy Elguindy[54], Ndeye Marieme Top[12], Ndongo Dia[12], Nédio Mabunda[178], Nei-yuan Hsiao[9,78], Nelson Boricó Silochi[162], Ngiambudulu M. Francisco[139], Ngonda Saasa[179], Nicholas Bbosa[3], Nickson Murunga[11], Nicksy Gumede[18], Nicole Wolter[13,113], Nikita Sitharam[1], Nnaemeka Ndodo[88], Nnennaya A. Ajayi[180], Noël Tordo[181], Nokuzola Mbhele[9], Norosoa H. Razanajatovo[77], Nosamiefan Iguosadolo[43], Nwando Mba[88], Ojide C. Kingsley[182], Okogbenin Sylvanus[92], Oladiji Femi[183], Olubusuyi M. Adewumi[31,32], Olumade Testimony[43,44], Olusola A. Ogunsanya[43], Oluwatosin Fakayode[184], Onwe E. Ogah[185], Ope-Ewe Oludayo[43], Ousmane Faye[12], Pamela Smith-Lawrence[155], Pascale Ondoa[71], Patrice Combe[186], Patricia Nabisubi [187,188], Patrick Semanda[126], Paul E. Oluniyi[43], Paulo Arnaldo[178], Peter Kojo Quashie[91], Peter O. Okokhere[92,189], Philip Bejon[11], Philippe Dussart[77], Phillip A. Bester[190], Placide K. Mbala[37,38], Pontiano Kaleebu[3,97], Priscilla Abechi[43,44], Rabeh El-Shesheny[40,191], Rageema Joseph[9], Ramy Karam Aziz[192,193], René G. Essomba[194,195], Reuben Ayivor-Djanie[91,120,121], Richard Njouom[196], Richard O. Phillips[63], Richmond Gorman[63], Robert A. Kingsley[48], Rosa Maria D.E.S.A. Neto Rodrigues[197,198], Rosemary A. Audu[29], Rosina A.A. Carr[120,121], Saba Gargouri[125], Saber Masmoudi[41], Sacha Bootsma[144], Safietou Sankhe[12], Sahra Isse Mohamed[199], Saibu Femi[43], Salma Mhalla[132,200], Salome Hosch[161,201], Samar Kamal Kassim[128], Samar Metha[57], Sameh Trabelsi[202], Sara Hassan Agwa[128], Sarah Wambui Mwangi[7], Seydou Doumbia[52], Sheila Makiala-Mandanda[37,38], Sherihane Aryeetey[63], Shymaa S. Ahmed[54], Side Mohamed Ahmed[103], Siham Elhamoumi[57], Sikhulile Moyo[100,203], Silvia Lutucuta[139], Simani Gaseitsiwe[100,203], Simbirie Jalloh[64,65], Soa Fy Andriamandimby[77], Sobajo Oguntope [43], Solène Grayo[181], Sonia Lekana-Douki[68], Sophie Prosolek[48], Soumeya Ouangraoua[204,205], Stephanie van Wyk[1], Stephen F. Schaffner[57], Stephen Kanyerezi[187,188], Steve Ahuka-Mundeke[37,38], Steven Rudder[48], Sureshnee Pillay[2], Susan Nabadda[126], Sylvie Behillil[206], Sylvie L. Budiaki[159], Sylvie van der Werf[206], Tapfumanei Mashe[39,207], Thabo Mohale[13], Thanh Le-Viet[48], Thirumalaisamy P. Velavan[114,208], Tobias Schindler[161,162,201], Tongai G. Maponga[118], Trevor Bedford[141,209], Ugochukwu J. Anyaneji[2], Ugwu Chinedu[43,44], Upasana Ramphal[2,10,210], Uwem E. George[43], Vincent Enouf[206], Vishvanath Nene[102], Vivianne Gorova[211,212], Wael H. Roshdy[54], Wasim Abdul Karim[1], William K. Ampofo[213], Wolfgang Preiser[118,119], Wonderful T. Choga[100,214], Yahaya Ali Ahmed[18], Yajna Ramphal[1], Yaw Bediako[91,215], Yeshnee Naidoo[2], Yvan Butera[175,216,217], Zaydah R. de Laurent[11], Africa Pathogen Genomics Initiative (PGI) Consortium, Ahmed E.O. Ouma[7], Anne von

Gottberg[13,113], George Githinji[11,218], Matshidiso Moeti[18], Oyewale Tomori[43], Pardis C. Sabeti[57], Amadou A. Sall[12], Samuel O. Oyola[102], Yenew K. Tebeje[7], Sofonias K. Tessema[7], Tulio de Oliveira[1,2,10,219]*, Christian Happi[43,44], Richard Lessells[2], John Nkengasong[7], Eduan Wilkinson[1,2]†*

[1]Centre for Epidemic Response and Innovation (CERI), School of Data Science and Computational Thinking, Stellenbosch University; Stellenbosch, South Africa.
[2]KwaZulu-Natal Research Innovation and Sequencing Platform (KRISP), Nelson R Mandela School of Medicine, University of KwaZulu-Natal; Durban, South Africa.
[3]MRC/UVRI & LSHTM Uganda Research Unit; Entebbe, Uganda.
[4]MRC-University of Glasgow Centre for Virus Research; Glasgow, United Kingdom.
[5]The Biotechnology Centre of the University of Yaoundé I; Yaoundé, Cameroon.
[6]CDC Foundation; Atlanta, Georgia, Nebraska Department of Health and Human Services; Lincoln, Nebraska
[7]Institute of Pathogen Genomics, Africa Centres for Disease Control and Prevention (Africa CDC); Addis Ababa, Ethiopia.
[8]Institute of Infectious Diseases and Molecular Medicine, Department of Integrative Biomedical Sciences, Computational Biology Division, University of Cape Town; Cape Town, South Africa.
[9]Division of Medical Virology, Wellcome Centre for Infectious Diseases in Africa, Institute of Infectious Disease and Molecular Medicine, University of Cape Town; Cape Town, South Africa.
[10]Centre for the AIDS Programme of Research in South Africa (CAPRISA); Durban, South Africa.
[11]KEMRI-Wellcome Trust Research Programme; Kilifi, Kenya
[12]Virology Department, Institut Pasteur de Dakar; Dakar, Senegal.
[13]National Institute for Communicable Diseases (NICD) of the National Health Laboratory Service (NHLS); Johannesburg, South Africa.
[14]School of Health Sciences, College of Health Sciences, University of KwaZulu-Natal; Durban, KwaZulu-Natal, South Africa.
[15]Department of Medical Laboratory Sciences, College of Health Sciences, Addis Ababa University; Addis Ababa, Ethiopia.
[16]Department of Microbial, Cellular and Molecular Biology, College of Natural and Computational Sciences, Addis Ababa University; Addis Ababa, Ethiopia.
[17]Cancer Biology Department, Virology and Immunology Unit, National Cancer Institute, Cairo University; Cairo, Egypt.
[18]World Health Organization, Africa Region; Brazzaville, Republic of the Congo.
[19]Centre d'Infectiologie Charles Mérieux-Mali (CICM-Mali); Bamako, Mali.
[20]Bacteriology and Virology Department Souro Sanou University Hospital; Bobo-Dioulasso, Burkina Faso.
[21]West African Health Organisation
[22]Faculty of Medicine and Health Sciences. Kassala University; Kassala City, Sudan.
[23]Department of Microbiology, Faculty of Medical Laboratory Sciences, University of Gezira; Gezira, Sudan.
[24]General Administration of Laboratories and Blood Banks, Ministry of Health, Kassala State; Sudan
[25]MRC Unit The Gambia at LSHTM; Fajara, Gambia.
[26]National Public Health Laboratory, Ministry of Health; Juba, Republic of South Sudan.
[27]Libyan Biotechnology Research Center; Tripoli, Libya.
[28]Center for Medical and Sanitary Research (CERMES)
[29]The Nigerian Institute of Medical Research; Yaba, Lagos, Nigeria.
[30]Laboratoire de la Caisse Nationale de Sécurité Sociale; Djibouti, Republic of Djibouti.
[31]Department of Virology, College of Medicine, University of Ibadan; Ibadan, Nigeria.
[32]Infectious Disease Institute, College of Medicine, University of Ibadan; Ibadan, Nigeria
[33]Medical Microbiology and Parasitology Department, College of Medicine, University of Ibadan; Ibadan, Nigeria.

[34]Biorepository Clinical Virology Laboratory, College of Medicine, University of Ibadan; Ibadan, Nigeria.
[35]Department of Medical Microbiology and Parasitology. Faculty of Basic Clinical Sciences. College of Health Sciences. University of Ilorin; Ilorin, Kwara State, Nigeria.
[36]The Pirbright Institute; Woking, United Kingdom.
[37]Pathogen Sequencing Lab, Institut National de Recherche Biomédicale (INRB); Kinshasa, the Democratic Republic of the Congo.
[38]Université de Kinshasa (UNIKIN); Kinshasa, the Democratic Republic of the Congo.
[39]National Microbiology Reference Laboratory; Harare, Zimbabwe.
[40]Center of Scientific Excellence for Influenza Viruses, National Research Centre (NRC); Cairo, Egypt.
[41]Laboratory of Molecular and Cellular Screening Processes, Centre of Biotechnology of Sfax, University of Sfax; Sfax, Tunisia.
[42]Genomics and Epigenomics program, Research dept.,CCHE57357
[43]African Centre of Excellence for Genomics of Infectious Diseases (ACEGID), Redeemer's University; Ede, Osun State, Nigeria.
[44]Department of Biological Sciences, Faculty of Natural Sciences, Redeemer's University; Ede, Osun State, Nigeria.
[45]Laboratório de Biologia Molecular Jean Piaget
[46]University Jean Piaget in Guinea-Bissau
[47]SAMRC bioinformatics unit, SA Bioinformatics Institute, University of the Western Cape; Cape Town, South Africa.
[48]Quadram Institute Bioscience; Norwich, United Kingdom.
[49]Central Public Health Reference Laboratories; Freetown, Sierra Leone.
[50]Centre de Recherche et de Formation en Infectiologie de Guinée (CERFIG), Université de Conakry; Conakry, Guinea.
[51]TransVIHMI, Institut de Recherche pour le Développement, Institut National de la Santé et de la Recherche Médicale (INSERM), Montpellier University, 34090; Montpellier, France.
[52]University Clinical Research Center (UCRC), University of Sciences, Techniques and Technology of Bamako; Bamako, Mali.
[53]Sharjah Institute for Medical Research, College of Medicine, University of Sharjah; Sharjah, United Arab Emirates.
[54]Central Public Health Laboratories (CPHL); Cairo, Egypt.
[55]National Institute of Public Health, Burundi
[56]Laboratoire des Fièvres Hémorragiques Virales du Benin
[57]Broad Institute of Harvard and MIT; Cambridge, MA, United States.
[58]Laboratory of Clinical Virology, WHO Reference Laboratory for Poliomyelitis and Measles in the Eastern Mediterranean Region, Pasteur Institute of Tunis, University Tunis El Manar (UTM); Tunis 1002, Tunisia.
[59]Research Laboratory "Virus, Vectors and Hosts: One Health Apporach and Technological Innovation for a Better Health", LR20IPT02, Pasteur Institute; Tunis 1002, Tunisia.
[60]Fondation Congolaise pour la Recherche Médicale; Brazzaville, Republic of the Congo.
[61]Marien Ngouabi, Brazzaville, Republic of Congo
[62]Kwame Nkrumah University of Science and Technology, Department of Theoretical and Applied Biology; Kumasi, Ghana.
[63]Kumasi Centre for Collaborative Research in Tropical Medicine, Kwame Nkrumah University of Science and Technology; Kumasi, Ghana
[64]Viral Haemorrhagic Fever Laboratory, Kenema Government Hospital; Kenema, Sierra Leone.
[65]Ministry of Health and Sanitation, Freetown, Sierra Leone. Eastern Province, Sierra Leone
[66]Department of Immunology, University of Maiduguri Teaching Hospital, P.M.B. 1414; Maiduguri, Nigeria.
[67]Department of Medical Laboratory Science, College of Medical Sciences, University of Maiduguri, P.M.B. 1069; Maiduguri, Borno State, Nigeria.
[68]Centre Interdisciplinaires de Recherches Medicales de Franceville (CIRMF); Franceville, Gabon.

[69]Département de Parasitologie-Mycologie Université des Sciences de la Santé (USS); Libreville, Gabon.
[70]National HIV Reference Laboratory, Community Health Sciences Unit, Ministry of Health; Lilongwe, Malawi.
[71]African Society for Laboratory Medicine; Addis Ababa, Ethiopia
[72]National Medical and Molecular Biology Laboratory, Ministry of Health; Djibouti, Republic of Djibouti.
[73]Africa CDC, Rapid Responder, Team Djibouti
[74]Noguchi Memorial Institute for Medical Research, University of Ghana; Legon, Ghana
[75]Seychelles Public Health Laboratory, Public Health Authority, Ministry of Health Seychelles
[76]Department of Ecology and Evolution; University of Chicago; Chicago, Illinois, United States of America.
[77]Virology Unit, Institut Pasteur de Madagascar; Antananarivo, Madagascar.
[78]National Health Laboratory Service (NHLS); Cape Town, South Africa.
[79]Malawi-Liverpool-Wellcome Trust Clinical Research Programme; Malawi
[80]Liverpool School of Tropical Medicine; Liverpool, United Kingdom.
[81]University of Nebraska Medical Center (UNMC); Omaha, NE, United States.
[82]SAMRC Antibody Immunity Research Unit, School of Pathology, University of the Witwatersrand; Johannesburg, South Africa.
[83]CHU de Bouaké, Laboratoire / Unité de Diagnostic des Virus des Fièvres Hémorragiques et Virus Émergents; Bouaké, Côte d'Ivoire.
[84]UNIVERSITE ALASSANE OUATTARA
[85]School of Public Health, Pwani University; Kilifi, Kenya.
[86]Faculty of Science and Techniques, University Marien Ngouabi; Brazzaville, Republic of the Congo.
[87]Centre for Human Virology and Genomics, Nigerian Institute of Medical Research; Yaba, Lagos, Nigeria.
[88]Nigeria Centre for Disease Control & Prevention; Abuja, Nigeria.
[89]Laboratoire des Arbovirus, Fièvres Hémorragiques virales, Virus Emergents et Zoonoses, Institut Pasteur de Bangui; Bangui, Central African Republic.
[90]Le Laboratoire National de Biologie Clinique et de Santé Publique (LNBCSP); Bangui, Central African Republic.
[91]West African Centre for Cell Biology of Infectious Pathogens (WACCBIP), College of Basic and Applied Sciences, University of Ghana; Accra, Ghana.
[92]Institute of Lassa Fever Research and Control, Irrua Specialist Teaching Hospital; Irrua, Nigeria.
[93]PATH; Lusaka, Zambia.
[94]Department of Entomology and Plant Pathology, North Carolina State University; Raleigh, NC, United States.
[95]Bioinformatics Research Center, North Carolina State University; Raleigh, NC, United States.
[96]School of Life Sciences and Zeeman Institute for Systems Biology and Infectious Disease Epidemiology Research (SBIDER), University of Warwick; Coventry, United Kingdom.
[97]Uganda Virus Research Institute; Entebbe, Uganda.
[98]PathCare Vermaak, Pretoria, South Africa and Division of Virology, University of the Free State; Bloemfontein, South Africa.
[99]College of Medicine and Allied Health Sciences, University of Sierra Leone
[100]Botswana Harvard AIDS Institute Partnership & Botswana Harvard HIV Reference Laboratory; Gaborone, Botswana.
[101]Macha Research Trust; Choma, Zambia.
[102]International Livestock Research Institute (ILRI); Nairobi, Kenya.
[103]INRSP; Nouakchott, Mauritania.
[104]Faculté DE MÉDECINE DE Nouakchott
[105]Rwanda National Reference Laboratory; Kigali, Rwanda.
[106]Robert Koch-Institute; Berlin, Germany.
[107]G5 Evolutionary Genomics of RNA viruses, Institut Pasteur; Paris, France.
[108]Direcção Nacional da Saúde Pública, Ministério da Saúde; Luanda, Angola.
[109]National Public Health Reference Laboratory-National Public Health Institute of Liberia
[110]Faculty of pharmacy of Monastir; Monastir, Tunisia.

[111]National Influenza Centre, Institut Pasteur d'Algérie; Algiers, Algeria.
[112]Department of Virology, National Health Laboratory Service (NHLS), Charlotte Maxeke Johannesburg Academic Hospital; Johannesburg, South Africa.
[113]School of Pathology, Faculty of Health Science, University of the Witwatersrand, Johannesburg, South Africa.
[114]Institute of Tropical Medicine, Universitätsklinikum Tübingen; Tübingen, Germany.
[115]Ministère de Santé Publique et de la Solidarité Nationale, Chad
[116]WHO Int Comoros
[117]World Health Organization, Africa Region; Brazzaville, Republic of the Congo.
[118]Division of Medical Virology, Faculty of Medicine and Health Sciences, Stellenbosch University; Tygerberg, Cape Town, South Africa.
[119]National Health Laboratory Service (NHLS); Tygerberg, Cape Town, South Africa.
[120]UHAS COVID-19 Testing and Research Centre, University of Health and Allied Sciences; Ho, Ghana.
[121]Department of Biomedical Sciences, University of Health and Allied Sciences, PMB 31, Ho. Ghana
[122]Institution and Department: Ministry Of Health, COVID-19 Testing Laboratory; Mbabane, Kingdom of Eswatini.
[123]Satellite Molecular Laboratory, Rivers State University Teaching Hospital; Port Harcourt, Nigeria.
[124]Department of Emerging Infectious Diseases, Institute of Tropical Medicine, Nagasaki University; Nagasaki, Japan.
[125]CHU Habib Bourguiba, Laboratory of Microbiology, Faculty of Medicine of Sfax, University of Sfax; Sfax, Tunisia.
[126]Central Public Health Laboratories (CPHL); Kampala, Uganda.
[127]Institut Pasteur de Cote d'Ivoire, Departement des Virus Epidemiques; Abidjan, Cote d'Ivoire.
[128]Faculty of Medicine Ain Shams Research institute (MASRI), Ain Shams University; Cairo, Egypt.
[129]Doctoral School of Technical and Environmental Sciences, Department of Biology and Human Health; N'Djamena, Chad.
[130]Department of Infectious Diseases Epidemiology, London School of Hygiene & Tropical Medicine, United Kingdom
[131]Charles Nicolle Hospital, Laboratory of Microbiology, National Influenza Center, 1006; Tunis, Tunisia.
[132]University of Tunis El Manar, Faculty of Medicine of Tunis, Research Laboratory LR99ES09; Tunis, Tunisia.
[133]College of Medicine and Allied Health Science, University of Sierra Leone
[134]Namibia Institute of Pathology; Windhoek, Namibia.
[135]National Institute of Hygiene; Lomé, Togo.
[136]Virology/Molecular Biology Department, Central Health Laboratory, Victoria Hospital, Ministry of Health and Wellness; Port Louis, Mauritius.
[137]Department of Biostatistics and Data Science, School of Public Health and Tropical Medicine, Tulane University; New Orleans, LA, United States of America.
[138]WHO Burundi; Gitega, Burundi.
[139]Grupo de Investigação Microbiana e Imunológica, Instituto Nacional de Investigação em Saúde (National Institute for Health Research); Luanda, 3635, Angola.
[140]Departamento de Bioquímica, Faculdade de Medicina, Universidade Agostinho Neto
[141]Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Center; Seattle, WA, United States.
[142]Universidade Federal de Minas Gerais; Belo Horizonte, Brazil.
[143]Institute of Agricultural Sciences, Universidade Federal dos Vales do Jequitinhonha e Mucuri; Unaí, Brazil.
[144]WHO South Sudan; Juba, South Sudan.
[145]Faculty of medicine, University of Burundi
[146]Pasteur Network, Institut Pasteur, 25-26 rue du Docteur Roux, 75015; Paris, France
[147]Botswana Institute for Technology Research and Innovation; Gaborone, Botswana.
[148]Instituto Nacional de Saúde Pública, Cape Verde
[149]Zambia National Public Health Institute; Lusaka, Zambia.
[150]Public Health Institute of Malawi

[151]National Health Laboratory; Gaborone, Botswana.
[152]Laboratory of Transmissible Diseases and Biologically Active Substances (LR99ES27), Faculty of Pharmacy, University of Monastir; Monastir, Tunisia.
[153]Laboratory of Microbiology, University Hospital of Monastir; Monastir, Tunisia.
[154]Biomedical Informatics and Chemoinformatics Group, Informatics and Systems Department, National Research Centre; Cairo, Egypt.
[155]Ministry of Health and Wellness; Gaborone, Botswana.
[156]Eastern Technical University of Sierra Leone; Kenema, Sierra Leone.
[157]Zoonotic Arbo and Respiratory Virus Program, Centre for Viral Zoonoses, Department of Medical Virology, University of Pretoria; Pretoria, South Africa.
[158]Next Generation Sequencing Unit and Division of Virology, Faculty of Health Sciences, University of the Free State, Bloemfontein, South Africa.
[159]National Reference Laboratory Lesotho; Maseru, Lesotho.
[160]Centre for Biotechnology Research and Development, Kenya Medical Research Institute; Nairobi, Kenya.
[161]Swiss Tropical and Public Health Institute; Basel, Switzerland.
[162]Laboratorio de Investigaciones de Baney; Baney, Equatorial Guinea.
[163]Ifakara Health Insitute; Ifakara, Tanzania.
[164]Department of Medical Diagnostics, Kumasi Centre for Collaborative Research in Tropical Medicine, Kwame Nkrumah University of Science and Technology; Kumasi, Ghana.
[165]PraesensBio; Lincoln, NE, United States.
[166]Department of Medical Laboratory Science, Niger Delta University; Bayelsa State, Nigeria
[167]Systems and Biomedical Engineering Department, Faculty of Engineering, Cairo University, 12613; Cairo, Egypt.
[168]King Faisal Specialist Hospital and Research Center; Riyadh, Kingdom of Saudi Arabia.
[169]Biological Prevention Department, Ministry of Defence; Cairo, Egypt.
[170]Faculty of Science, Fayoum University; Fayoum, Egypt.
[171]Molecular Pathology Lab Childeren Cancer Hospital, 57357; Cairo, Egypt.
[172]Laboratoire Biolim FSS/Université de Lomé; Lomé, Togo.
[173]Department of Genetics and Genomics, College of Medicine and health sciences, United Arab Emirates University; Abu Dhabi, United Arab Emirates.
[174]High Institute of Biotechnology of Monastir; University of Monastir; Rue Taher Haddad 5000, Monastir, Tunisia.
[175]Rwanda National Joint Task Force COVID-19, Rwanda Biomedical Centre, Ministry of Health; Kigali, Rwanda
[176]School of Health Sciences, College of Medicine and Health Sciences, University of Rwanda; Kigali, Rwanda
[177]Department of Microbiology, Faculty of Medical laboratory Sciences, Omdurman Islamic University, Sudan.
[178]Instituto Nacional de Saúde (INS); Marracuene, Mozambique.
[179]University of Zambia, School of Veterinary Medicine, Department of Disease Control; Lusaka, Zambia.
[180]Internal Medicine Department, Alex Ekwueme Federal University Teaching Hospital; Abakaliki, Nigeria.
[181]Institut Pasteur de Guinée; Conarky, Guinea.
[182]Virology Laboratory, Alex Ekwueme Federal University Teaching Hospital; Abakaliki, Nigeria.
[183]Department of Epidemiology and Community Health, Faculty of Clinical Sciences. College of Health Sciences. University of Ilorin; Ilorin, Kwara State, Nigeria.
[184]Department of Public Health, Ministry of Health; Ilorin, Kwara State, Nigeria.
[185]Alex Ekwueme Federal University Teaching Hospital; Abakaliki, Nigeria.
[186]Mayotte Hospital Center; Mayotte, France.
[187]The African Center of Excellence in Bioinformatics and Data-Intensive Sciences, The Infectious Diseases Institute; Kampala, Uganda
[188]Immunology and Molecular Biology, Makerere University; Kampala, Uganda.
[189]Department of Medicine, Faculty of Clinical Sciences, College of Medicine, Ambrose Alli University; Ekpoma, Edo State, Nigeria.

[190]Division of Virology, National Health Laboratory Service and University of the Free State; Bloemfontein, South Africa.
[191]Infectious Hazards Preparedness, World Health Organization, Eastern Mediterranean Regional Office; Cairo, Egypt.
[192]Department of Microbiology and Immunology, Faculty of Pharmacy, Cairo University, 11562; Cairo, Egypt.
[193]Microbiology and Immunology Research Program, Children's Cancer Hospital Egypt 57357, 11617; Cairo, Egypt.
[194]National Public Health Laboratory, Ministry of Public Health of Cameroon
[195]Faculty of Medicine and Biomedical Sciences, University of Yaoundé; Yaoundé, Cameroon.
[196]Virology Service, Centre Pasteur of Cameroun; Yaounde, Cameroun.
[197]Coordenadora da rede do Diagnóstico Tuberculose/ HIV/ COVID-19 na Instituição - Laboratório Nacional de Referência da Tuberculose em São Tomé e Príncipe; São Tomé, São Tomé and Principe.
[198]Ponto focal para Melhoria da qualidade dos Laboratórios (SLIPTA) ao nível de São Tomé e Príncipe; São Tomé, São Tomé and Principe.
[199]National Public Health Reference Laboratory (NPHRL), Somalia
[200]Faculty of Medicine of Monastir, University of Monastir; Monastir, Tunisia.
[201]University of Basel
[202]Clinical and Experimental Pharmacology Lab, LR16SP02, National Center of Pharmacovigilance, University of Tunis El Manar; Tunis, Tunisia.
[203]Harvard T.H. Chan School of Public Health; Boston, MA, United States.
[204]Centre MURAZ; Ouagadougou, Burkina Faso
[205]National Institute of Public Health of Burkina Faso (INSP/BF); Ouagadougou, Burkina Faso
[206]National Reference Center for Respiratory Viruses, Molecular Genetics of RNA Viruses, UMR 3569 CNRS, Université Paris Cité, Institut Pasteur; Paris, France
[207]World Health Organization; Harare, Zimbabwe.
[208]Vietnamese-German Center for Medical Research; Hanoi, Vietnam.
[209]Howard Hughes Medical Institute; Seattle, WA, United States
[210]Sub-Saharan African Network For TB/HIV Research Excellence (SANTHE); Durban, South Africa
[211]World Health Organization, WHO Lesotho; Maseru, Lesotho.
[212]Med24 Medical Centre; Ruwa, Zimbabwe.
[213]Department of Virology, Noguchi Memorial Institute for Medical Research, University of Ghana; Legon, Ghana.
[214]Division of Human Genetics, Department of Pathology, University of Cape Town; Cape Town, South Africa.
[215]Yemaachi Biotech; Accra, Ghana.
[216]Center for Human Genetics, College of Medicine and Health Sciences, University of Rwanda; Kigali, Rwanda.
[217]Laboratory of Human Genetics, GIGA Research Institute; Liège, Belgium.
[218]Department of Biochemistry and Biotechnology, Pwani University; Kilifi, Kenya.
[219]Department of Global Health, University of Washington; Seattle, WA, United States.

† These authors contributed equally

Structured Abstract:

Introduction: Investment in Africa over the past year in regards to SARS-CoV-2 sequencing has led to a massive increase in the number of sequences, exceeding 100 000 sequences generated to date to track the pandemic on the continent. These sequences have had a profound impact on how public health officials in Africa have navigated the COVID-19 pandemic. **Rationale:** Here we

demonstrate how the first 100 000 SARS-CoV-2 sequences from Africa have helped monitor the epidemic on the continent, how genomic surveillance in Africa expanded over the course of the pandemic, and how we adapted our sequencing methods to deal with an evolving virus. Finally, we also examine how viral lineages have spread across the continent in a phylogeographic framework to gain insights into the underlying temporal and spatial transmission dynamics for several Variants of Concern (VOCs). **Results:** Our results indicate a growing number of countries in Africa that can sequence the virus within their own borders, coupled with a shorter turnaround time from the time of sampling to sequence submission when sequencing is performed locally. Ongoing SARS-CoV-2 evolution necessitated the continual updating of primer sets, as a result, eight primer sets were designed in tandem with the viral evolution and employed to ensure effective sequencing throughout the course of the pandemic. The pandemic unfolded through multiple waves of infection each driven by distinct genetic lineages, with B.1-like ancestral strains associated with the first pandemic wave of infections in 2020. Successive waves on the continent were fueled by different Variants of Concern (VOCs) with Alpha and Beta co-circulating in distinct spatial patterns during the second wave, while Delta and Omicron during the third and fourth wave, respectively, affected the continent as a whole. Phylogeographic reconstruction points towards distinct differences in viral importation and exportation patterns associated with the Alpha, Beta, Delta, and Omicron variants and sub-variants, both when considering Africa versus the rest of the world and viral dissemination within the continent. Our epidemiological and phylogenetic inferences therefore underscore the heterogeneous nature of the pandemic on the continent and highlight key insights and challenges; for instance, recognizing limitations of low testing proportions. We also highlight the early warning capacity that genomic surveillance in Africa has had for the rest of the world with the detection of new lineages and variants, the most recent being the characterization of various Omicron sub-variants. **Conclusion:** Sustained investment for diagnostics and genomic surveillance in Africa is needed as the virus continues to evolve. This is important not only to help combat SARS-CoV-2 on the continent, but can also be used as a platform to help address the many emerging and re-emerging infectious disease threats in Africa (e.g. Rift valley fever, ebola or poliomyelitis). In particular, capacity building for local sequencing within countries or within the continent should be prioritized, as this is generally associated with shorter turnaround times, providing the most benefit to local public health authorities tasked with pandemic response and mitigation, and allowing for the fastest reaction to localized outbreaks. These investments are crucial for pandemic preparedness and response and will serve the health of the continent well into the 21st century.

Print Abstract:

Investment in SARS-CoV-2 sequencing in Africa over the past year has led to a major increase in the number of sequences generated, now exceeding 100 000 genomes, used to track the pandemic on the continent. Our results show an increase in the number of African countries able to sequence domestically, and highlight that local sequencing enables faster turnaround time and more regular

routine surveillance. Despite limitations of low testing proportions, findings from this genomic surveillance study underscores the heterogeneous nature of the pandemic and shed light on the distinct dispersal dynamics of Variants of Concern, particularly Alpha, Beta, Delta, and Omicron, on the continent. Sustained investment for diagnostics and genomic surveillance in Africa is needed as the virus continues to evolve, while the continent faces many emerging and re-emerging infectious disease threats. These investments are crucial for pandemic preparedness and response and will serve the health of the continent well into the 21st century.

One-Sentence Summary: Expanding Africa SARS-CoV-2 sequencing capacity in a rapidly evolving pandemic.

Introduction

What originally started as a small cluster of pneumonia cases in Wuhan, China over two years ago (*1*), quickly turned into a global pandemic. Coronavirus Disease 2019 (COVID-19) is the clinical manifestation of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) infection; and by March 2022 there had been over 437 million reported cases and over 5.9 million reported deaths (*2*). Though Africa accounts for the lowest number of reported cases and deaths thus far, with ~11.3 million reported cases and 245 000 reported deaths as of February 2022, the continent has played an important role in shaping the scientific response to the pandemic with the implementation of genomic surveillance and the identification of two of the five variants of concerns (VOCs) (*3*, *4*).

Since it emerged in 2019, SARS-CoV-2 has continued to evolve and adapt (*5*). This has led to the emergence of several viral lineages that carry mutations that confer some viral adaptive advantages that increase transmission and infection (*6*, *7*), or counter the effect of neutralizing antibodies from vaccination (*8*) or previous infections (*9–11*). The World Health Organization (WHO) classifies certain viral lineages as variants of concert (VOCs) or variants of interest (VOIs) based on the potential impact they may have on the pandemic, with VOCs regarded as the highest risk. To date, five VOCs have been classified by the WHO, two of which were first detected on the African continent (Beta and Omicron) (*3*, *4*, *12*), while two more (Alpha and Delta) (*12*, *13*) have spread extensively on the continent in successive waves. The remaining VOC, Gamma (*14*), originated in Brazil and had a limited influence in Africa with only four recorded sequenced cases.

For genomic surveillance to be useful for public health responses, sampling for sequencing needs to be both spatially and temporally representative. In the case of SARS-CoV-2 in Africa, this means extending the geographic coverage of sequencing capacity to capture the dynamic genomic epidemiology in as many locations as possible. In a meta-analysis of the first 10 000 SARS-CoV-2 sequences generated in 2020 from Africa (*15*) several blindspots were identified with regards to genomic surveillance on the continent. Since then, much investment has been devoted to building

capacity for genomic surveillance in Africa, coordinated mostly by the Africa Centers for Disease Control (Africa CDC) and the regional office of the WHO in Africa (or WHO AFRO), but also provided by several national and international partners resulting in an additional 90 000 sequences shared over the past year (April 2021 - March 2022). This makes the sequencing effort for SARS-CoV-2 a phenomenal milestone. In comparison, only 12 000 whole genome influenza sequences (*16*) and only ~3 700 whole genome HIV sequences (*17*) from Africa have been shared publicly even though HIV has plagued the continent for decades.

Here we describe how the first 100 000 SARS-CoV-2 sequences from Africa have helped describe the pandemic on the continent, how this genomic surveillance in Africa has expanded, and how we adapted our sequencing methods to deal with an evolving virus. We also highlight the impact that genomic sequencing in Africa has had on the global public health response, particularly through the identification and early analysis of new variants. Finally, we also describe here for the first time how the Delta and Omicron variants have spread across the continent, and how their transmission dynamics were distinct from the Alpha and Beta variants that preceded them.

Results

Epidemic waves driven by variant dynamics and geography

Scaling up sequencing in Africa has provided a wealth of information on how the pandemic unfolded on the continent. The epidemic has largely been spatially heterogeneous across Africa, but most countries have experienced multiple waves of infection (*18–29*), with significant local and regional diversity in the first and to a lesser extent the second waves, followed by successive sweeps of the continent with Delta and Omicron (**Figure 7.1A**). In all regions of the continent, different lineages and VOIs evolved and co-circulated with VOCs and in some cases, contributed considerably to epidemic waves.

In North Africa (**Figure 7.1B, Appendix F - Supp Fig S1A**), B.1 lineages and Alpha dominated in the first and second wave of the pandemic and were replaced by Delta and Omicron in the third and fourth waves, respectively. Interestingly, the C.36 and C.36.3 sub-lineage dominated the epidemic in Egypt (~40% of reported infections) before July 2021 when it was replaced by Delta (*30*). Similarly, in Tunisia the first and second waves were associated with the B.1.160 lineage and were replaced by Delta during the country's third wave of infections. In southern Africa (**Figure 7.1C, Appendix F - Supp Fig S1C**), we see a similar pandemic profile with B.1 dominating the first wave, but instead of Alpha, Beta was responsible for the second wave, followed by Delta and Omicron. Another lineage that was flagged for close monitoring in the region was C.1.2, due to its mutational profile and predicted capacity for immune escape (*31*). However, the C.1.2 lineage did not cause many infections in the region as it was circulating at a time when Delta was dominant. In West Africa (**Figure 7.1D, Appendix F - Supp Fig S1B**), the B.1.525 lineage caused a large

proportion of infections in the second and third waves where it shared the pandemic landscape with the Alpha variant. As with other regions on the continent, these variants were later replaced by the Delta and then Omicron VOCs in successive waves. In Central Africa (**Figure 7.1E, Appendix F - Supp Fig S1D**), the B.1.620 lineage caused most of the infections between January and June 2021 (*32*) before systematically being replaced by Delta and then Omicron. Lastly, in East Africa (**Figure 7.1F, Appendix F - Supp Fig S1E**) the A.23.1 lineage dominated the second wave of infections in Uganda (*33*) and much of East Africa. In all of these regions, minor lineages such as B.1.525, C.36 and A.23.1 were eventually replaced by VOCs that emerged in later waves.

Finally, we directly compared the official recorded cases in Africa with the ongoing SARS-CoV-2 genomic surveillance (GISAID date of access 2022-03-31) for a crude estimation of variants' contribution to cases. We observe that Delta was responsible for an epidemic wave between May and October 2021 (**Figure 7.1A**) and had the greatest impact on the continent with almost 34.2% of overall infections in Africa possibly attributed to it. Beta was responsible for an epidemic wave at the end of 2020 and beginning of 2021 (**Figure 7.1A**), with 13.3% of infections overall attributed to it. Notably, Alpha, despite being predominant in other parts of the world at the beginning of 2021, had only minimal significance in Africa, accounting for just 4.3% of infections. At the time of writing, the Omicron VOC had contributed to 21.6% of overall sequenced infections. At this time the Omicron wave was still unfolding globally and in Africa with the expansion of several sub-lineages (*34*), such that its full impact is yet to be determined. However, due to increased population immunity (*35*), from SARS-CoV-2 infection and vaccination (**Appendix F - Supp Fig S2**), the impact of Omicron on mortality has been less in comparison to the other VOCs, as can be observed by the relatively low death rate in South Africa during the Omicron wave (*36*). The findings from mapping epidemiological numbers onto genomic surveillance data are reliable as far as the proportional scaling of genomic sampling across Africa with the size and timing of epidemic waves (**Appendix F - Supp Fig S3**, b = 0.011, SE = 0.001, $p < 2 \times 10^{-16}$). The findings from mapping epidemiological numbers onto genomic surveillance data are reliable as far as the proportional scaling of genomic sampling across Africa with the size and timing of epidemic waves (**Appendix F - Supp Fig S3,** $\chi^2 = 20964.02$, *p-value* $< 2.2 \times 10^{-16}$).

This comes with the obvious caveats that testing and reporting practices have varied widely across the continent, along with genomic surveillance volumes throughout the pandemic. Countries in Africa with reported data have tested in proportions from as little as 0.1 daily tests per million population to more than 1 000 tests per million (**Appendix F - Supp Fig S4**). Some countries have consistently tested at high proportions, for example South Africa, Botswana, Morocco and Tunisia. Incidentally, these countries have also generally reported more cases per million, providing an indication that recorded low incidence in other parts of the continent has been an underestimate due to low testing rates. However, even for these countries, epidemic numbers are certainly underrepresented and underdetected, given that in several timeframes, test positivity rates were still on the higher end, approaching or exceeding 20% (**Appendix F - Supp Fig S4**), and as

concluded by seroprevalence surveys and estimates of true infection burdens in Africa (*37*, *38*). Findings of attributing case numbers of variants must therefore be interpreted in context of this limitation but can nevertheless provide a qualitative overview of the spatial and temporal dynamics of VOCs in relation to epidemic progression in Africa.

The African regional- (**Appendix F - Supp Table S1**) and country-specific (**Appendix F - Supp Table S2**) Nextstrain builds also clearly support the changing nature of the pandemic over time. From these builds we observe a strong association of B.1-like viruses circulating on the continent during the first wave. These "ancestral" lineages were subsequently replaced by the Alpha and Beta variants which dominated the pandemic landscape during the second wave, and were later replaced by the Delta and Omicron variants during the third and fourth waves.

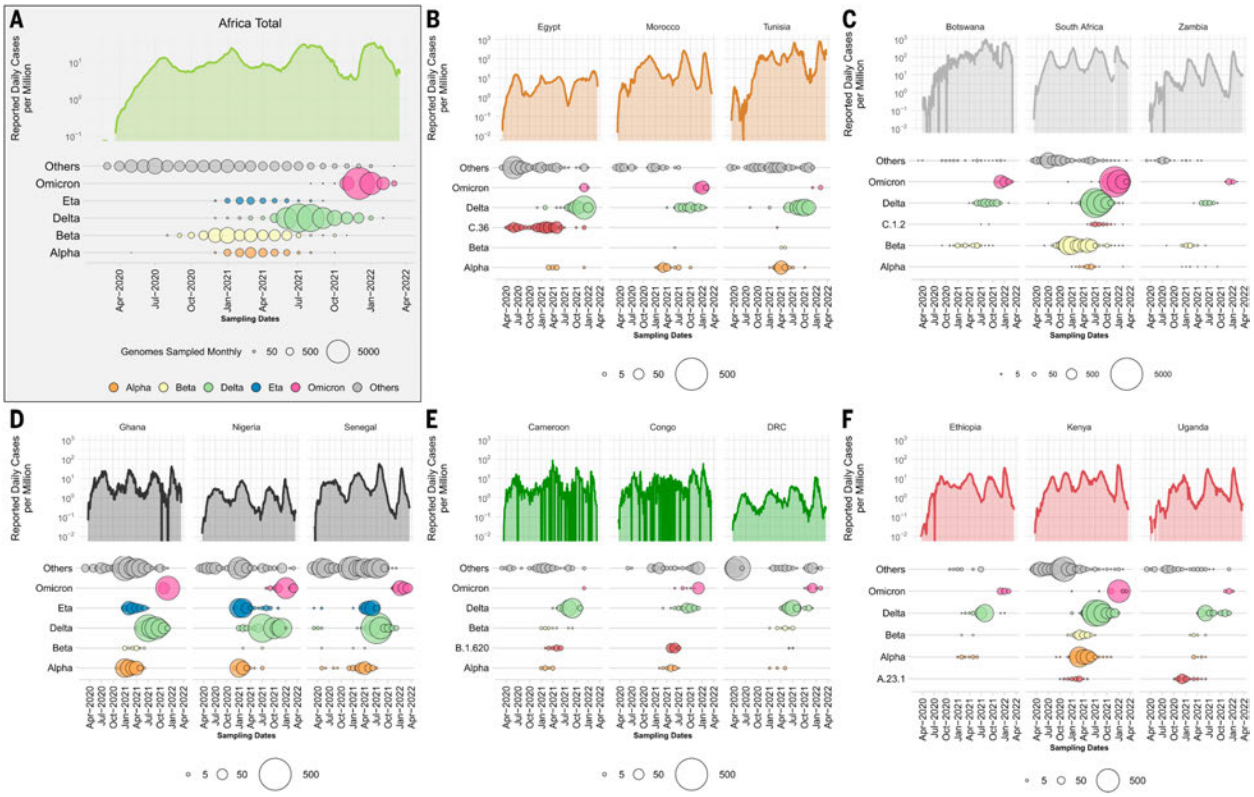


Figure 1: *Epidemiological progression of the COVID-19 pandemic on the African continent. A) Total reported new case counts per million inhabitants in Africa (Data Source: Our World in Data; OWID; log-transformed) along with the distribution of VOCs, the Eta VOI and other lineages through time (size of circles proportional to the number of genomes sampled per month for each category). (B-F) Breakdown of reported new cases per million (Data Source: Our World in Data; OWID; log-transformed) and monthly sampling of VOCs, regional variant or lineage of interest and other lineages for three selected countries for North, Southern, West, Central and East Africa respectively. For each region, a different variant or lineage of interest is shown, relevant to that region (C.36, C.1.2, Eta, B.1.620 and A.23.1, respectively).*

Optimizing surveillance coverage in Africa

By mapping and comparing the locations of specimen sampling laboratories to the sequencing laboratories, a number of aspects regarding the expansion of genomic surveillance on the continent became clear. First, even though several countries in Africa started sequencing SARS-CoV-2 in the first months of the pandemic, local sequencing capacity was initially limited. However, local sequencing capabilities slowly expanded over time, particularly after the emergence of VOCs (**Figure 7.2A**). The fact that almost half of all SARS-CoV-2 sequencing in Africa was performed using the Oxford Nanopore technology (ONT), which is relatively low-cost compared to other sequencing technologies and better adapted to modest laboratory infrastructures, illustrates one component of how this rapid scale-up of local sequencing was achieved (**Appendix F - Supp Fig S5**). Yet, to rely only on local sequencing would have thwarted the continent's chance at a reliable genomic surveillance program. At the time of writing, there were 52/55 countries in Africa with SARS-CoV-2 genomes deposited in GISAID, however, there were still 16 countries with no reported local sequencing capacity (**Figure 7.2A**) and undoubtedly many with limited capacity to meet demand during pandemic waves.

To tackle this, three centers of excellence and various regional sequencing hubs were established to maximize resources available in a few countries to assist in genomic surveillance across the continent. This sequencing is done either as the sole source of viral genomes for those countries (e.g. Angola, South Sudan and Namibia) or concurrently with local efforts to increase capacity during resurgences (**Figure 7.2B**). Sequencing is further supplemented by a number of countries utilizing facilities outside of Africa. Ultimately, a mix of strategies from local sequencing, collaborative resource sharing among African countries and sequencing with academic collaborators outside the continent helped close surveillance blindspots (**Figure 7.2C**). Countries in sub-Saharan Africa, particularly in Southern and East Africa, most benefited from the regional sequencing networks, while countries in West and North Africa often partnered with collaborators outside of Africa.

The success of pathogen genomic surveillance programs relies on how representative it is of the epidemic under investigation. For SARS-CoV-2, this is often measured in terms of the percentage of reported cases sequenced and the regularity of sampling. African countries were positioned across a range of different combinations of overall proportion and frequency of genomic sampling (**Figure 7.2D**). While the ultimate goal would be to optimize both of these parameters, a lower proportion of sampling can also be useful if frequency of sampling is maintained as high as possible. For instance, South Africa and Nigeria, who have both sequenced ~1% of cases overall, can be considered to have successful genomic surveillance programs on the basis that sampling is representative over time, and has enabled the timely detection of variants (Beta, Eta, Omicron).

Additionally, for genomic surveillance to be most useful for rapid public health response during a pandemic, sequencing would ideally be done in real-time or in a framework as close as possible to that. We show a general trend of decreasing sequencing turnaround time in Africa (**Appendix F - Supp Fig S6**), particularly from a mean of 182 days between October to December 2020 to a mean of 50 days over the same period a year later, although this does come with several caveats. Firstly, we measure sequencing turnaround time in the most accessible manner, which is by comparing the date of sampling of a specimen to the date its sequence was deposited in GISAID. Generally, the genomic data potentially informs the public health response more rapidly than reflected here, particularly when it comes to local outbreak investigations or variant detection. This analysis is also confounded by various factors such as country-to-country variation in these trends (**Appendix F - Supp Fig S7**), delays in data sharing, and potential retrospective sequencing, particularly by countries joining sequencing efforts at later stages of the pandemic. The most critical caveat is the fact that sequencing from the most recently collected samples (e.g. over the last six months) may still be ongoing. The shortening duration between sampling and genomic data sharing is nevertheless a positive takeaway, given that this data also feeds into continental and global genomic monitoring networks. Overall, the continental average delay from specimen collection to sequencing submission is 87 days with 10 countries having an average turnaround time of less than 60 days and Botswana of less than 30 days (**Appendix F - Supp Fig S8**).

Most importantly in the context of optimizing genomic surveillance, we found that the route taken to sequencing impacts the speed of data generation. Local sequencing has significantly faster sequencing turn-around times of the three frameworks we investigated (median of 51 days), followed by sequencing within regional sequencing networks in Africa (median of 93 days) and finally outsourced sequencing to countries outside Africa (median of 113 days) (**Figure 7.2E**). This finding strongly supports the investments in local genomic surveillance, to generate timely and regular data for local and regional decision making. Finally, we show that it is beneficial in several ways for countries to undertake genomic surveillance through several sequencing laboratories, rather than centralizing efforts. For instance, we estimate strong correlations between the numbers of sequencing laboratories per country with the total number of genomes produced by that country (method, correlation value), the total number of *epiweeks* for which sequencing data was produced (method, correlation value), and importantly, sequencing turnaround time (method, correlation value) (**Figure 7.2F**).

With the increase in sequencing capacity on the continent, a decrease in the time taken to detect new variants was observed. For example, the Beta variant was identified in December 2020 in South Africa (*4*), but sampling and molecular clock analyses suggest the variant originated in September 2020. This three-month lag in detection means that a new variant, like Beta, has ample time to spread over a large geographic region prior to its detection. However, by the end of 2021, the time to detect a new variant was substantially improved. Phylogenetic and molecular clock analyses suggest that the Omicron variant originated around 9 October 2021 (95% Highest

posterior density or HPD: 30 September - 20 October 2021) and the variant was described on 23rd November 2021 (*3*). Thus, Omicron was detected within ~5 weeks from origin compared to the Beta variant (~16 weeks) and the Alpha variant, detected in the UK (~10 weeks). More importantly, the time from sequence deposition to the WHO declaring the new variant a VOC was substantially shortened to 72 hours for the Omicron variant.

To interpret insights from the described genomic surveillance in Africa, it is important to understand the context of epidemiological reporting and sampling strategies utilized for sequencing on the continent (**Appendix F - Supp Table S3**). Most countries provided daily reports of newly recorded cases, while a few provided weekly and monthly reports. For most countries, surveillance was mainly focused on the major cities, suggesting potential cryptic circulation in rural areas. We find that at the onset of the pandemic, surveillance was focused on identification of imported cases from incoming travelers or local residents returning from various countries. As community transmissions began to emerge, the focus shifted towards regular surveillance and outbreak investigations. Together, these three strategies account for the vast majority of samples generated on the continent and analyzed here. As the pandemic progressed and vaccines were made available, some countries on the continent began to explore other sampling strategies such as reinfections, environmental samples such as waste water samples, and vaccine breakthrough cases to gain new insights into the evolutionary dynamics of SARS-CoV-2. The utility of sequencing for viral evolution tracking and VOC detection in the way described above is obviously also dependent on sampling proportions, especially within sampling for regular surveillance.

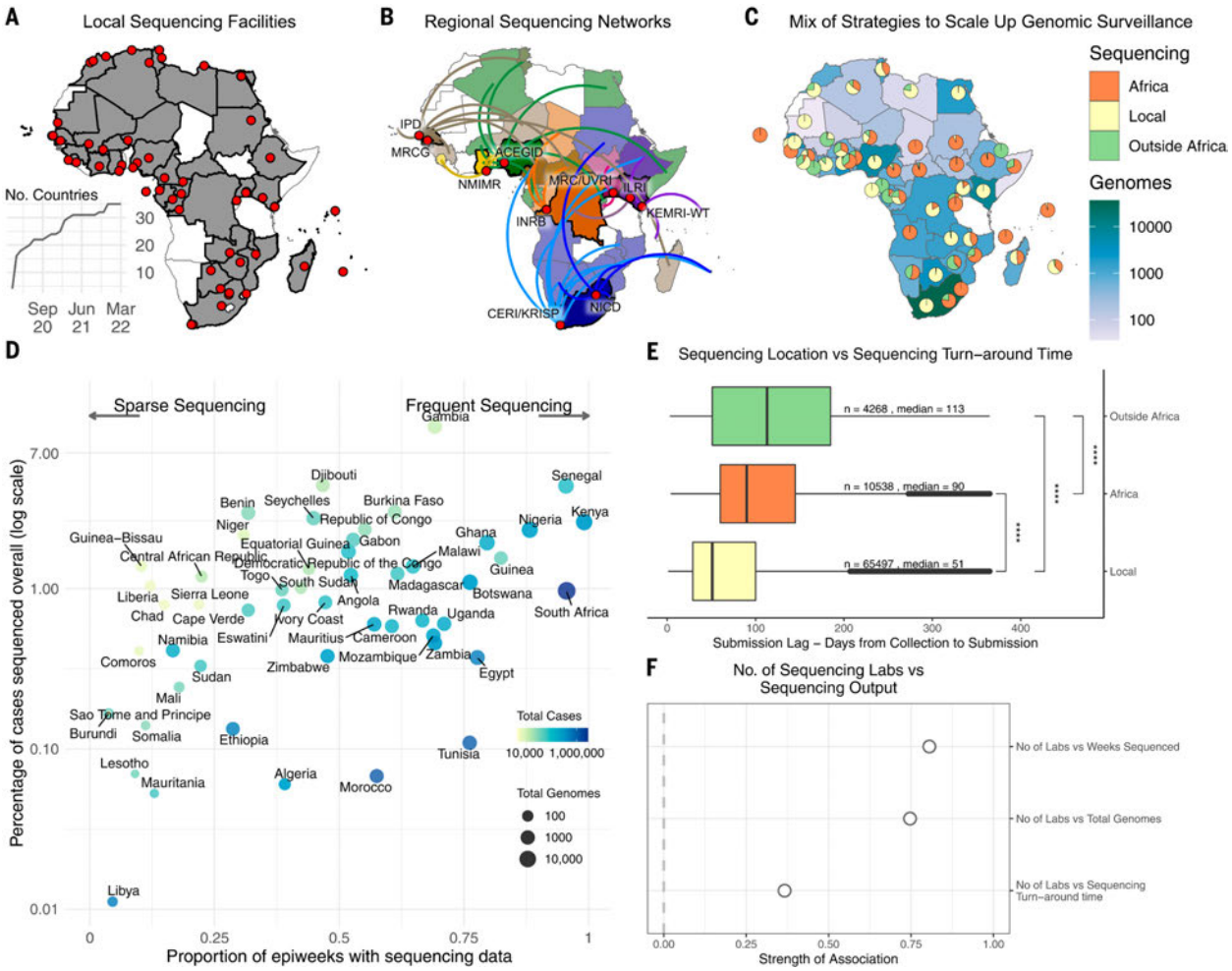


Figure 2: *Sequencing strategies and outputs in Africa. A) Geographical representation of all countries (shaded in gray) and institutions (red dots) in Africa with their own on-site sequencing facilities. The inset graph shows the number of countries in Africa able to carry out sequencing locally over time. B) Key regional sequencing hubs and networks in Africa showing countries (shaded in bright colors) and institutions (red dots) that have sequenced for other countries (shaded in corresponding light colors and linking curves) on the continent. CERI: Centre for Epidemic Response and Innovation; KRISP: KwaZulu-Natal Research Innovation and Sequencing Platform; NICD: National Institute for Communicable Diseases; KEMRI-WT: Kenya Medical Research Institute - Wellcome Trust; ILRI: International Livestock Research Institute; MRC/UVRI: Medical Research Council/Uganda Virus Research Institute; INRB: Institut National de Recherche Biomédicale; ACEGID: African Centre of Excellence for Genomics of Infectious Diseases; NMIMR: Noguchi Memorial Institute for Medical Research; MRCG: Medical Research Council Unit - The Gambia; IPD: Institut Pasteur de Dakar C) Geographical representation of the total number of SARS-CoV-2 whole genomes produced over the course of the pandemic in each country, as well as the proportion of those sequences that were produced locally, regionally or abroad. D) Correlation of the proportion of COVID-19 positive cases that have been sequenced and the corresponding number of epidemiological weeks since the start of the pandemic that are*

represented with genomes for each African country. The color of each circle represents the number of cases and its size the number of genomes. E) Comparison of sequencing turn-around times (lag times from sample collection to sequence submission) for the three strategies of sequencing in Africa, showing a significant difference in the means (p-value<0.0001). The box and whisker plot denote the lower quartile, median and upper quartile (box), the minimum and maximum values (whisker), and outliers (black dots). F) Pearson correlations of the total number of sequencing laboratories per country against key sequencing outputs.

The speed of SARS-CoV-2 evolution has complicated sequencing efforts. Common methods of RNA sequencing include reverse transcription followed by double stranded DNA amplification using sequence-specific primer sets (*39*). Ongoing SARS-CoV-2 evolution has necessitated the continual evaluation and updating of these primer sets to ensure their sustained utility during genomic surveillance efforts. Here, we examined the current set of genomes to determine aspects of the sequencing that might be improved in the future. Many of the primer sets used were designed using viral sequences from the start of the pandemic and may require updating to keep pace with evolution. Indeed, the ARTIC primer sets are currently in version 4.1 (*40*). The Entebbe primer set was designed mid-2020 well into the first year of the epidemic and used an algorithm and design that accommodates evolution (*41*). The effects of viral evolution on sequencing patterns can be seen with low median unspecified nucleotide (N) values (a consequence of primer dropout or low coverage at that site) observed for the first 12 months of the epidemic with an increase from October 2020 (**Figure 7.3A**). Additional challenges appear (indicated by increasing median N values) as the virus further evolved into Delta and Omicron lineages from January 2021 onward (**Figure 7.3A**). Examining the role of sequencing technology, it appears that the two major technologies used (Illumina and ONT) have similar gap profiles (as measured by N content) while Ion Torrent, MGI and Sanger show reduced N content (**Figure 7.3B**). Likely factors for this pattern are the primers used in sequencing, with primer choice playing a key role in the quantity of gaps (**Figure 7.3C**). The N content varies with viral lineage (**Figure 7.3D**). There was a modest difference in N content across the lineages. Lineages that returned no classification with Pangolin ("None") showed the highest N content, suggesting that high N content was probably the basis for failed classification. The more recent lineages Delta (e.g. AY.39, AY.75) and Omicron (BA.1.1, BA.2) also showed higher N content consistent with virus evolution impairing primer function.This pattern is further explored in **Appendix F - Supp Fig S9** with position of gaps showing an enrichment in the genome regions after position 19 000 with frequent gaps disrupting the spike coding region.

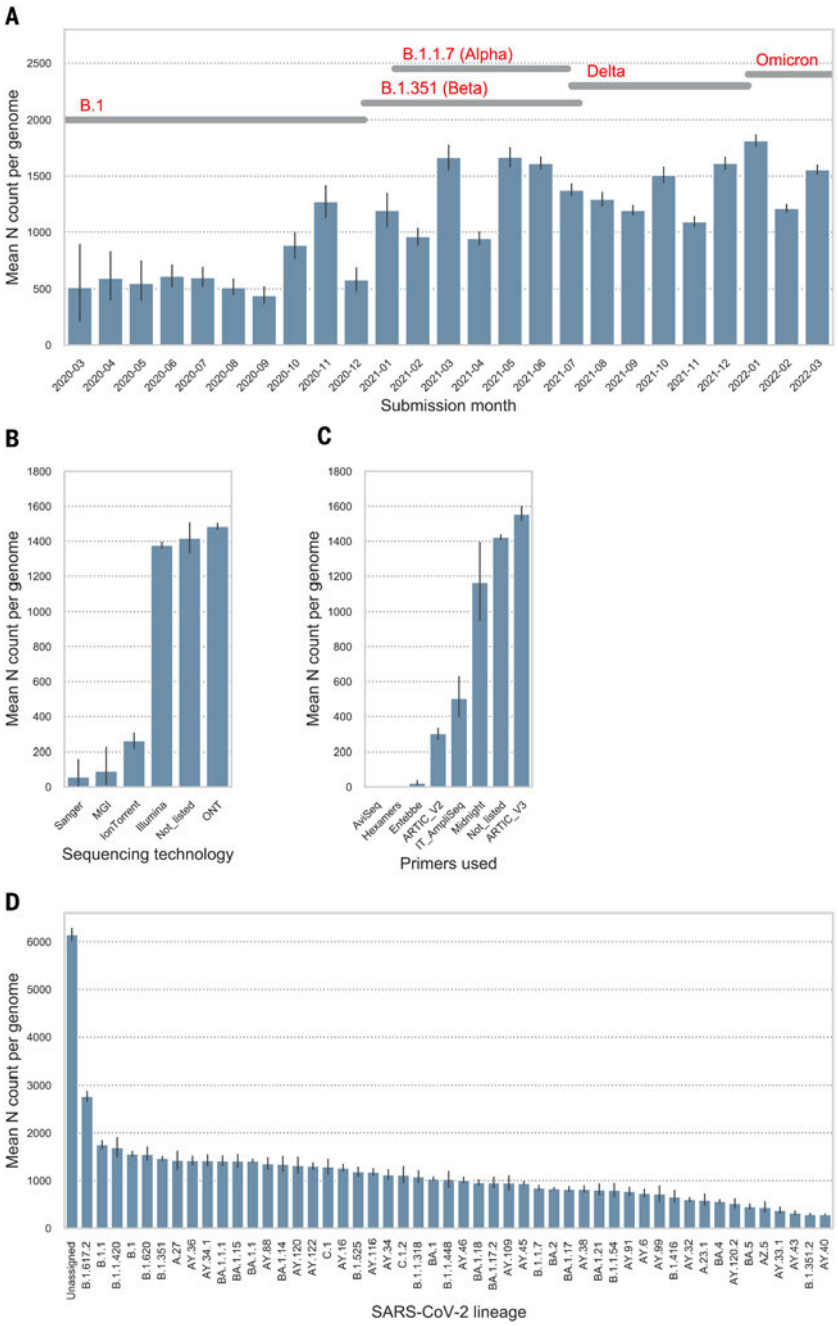


Figure 3: *Genome gap analysis. A) Shows the mean N count per genome by month of submission to GISAID. The dates for the detection of important SARS-CoV-2 lineages are indicated at the top of the figure. B) Illustrates the mean N count per genome stratified by sequencing technology. C) Shows the mean N count per genome stratified by the sequencing primers sets used. For panels A to C, error bars indicate 95% confidence intervals. D) Gapped genomes by lineage. The mean N data were stratified by SARS-CoV-2 lineages to investigate lineage-specific frequency of genome gaps, an indirect measure of primer mismatch. All lineages present at least 100 times in the genome data were presented.*

Phylogenetic insights into the rise and spread of Variants of Concern in Africa

During the first wave of infections in 2020 in Africa, as was the case globally, the majority of corresponding genomes were classified as PANGO B.1 (n=2 456) or B.1.1 viruses (n=1 329). Towards the end of 2020, more distinct viral lineages started to appear. The most important of which that impacted the African continent are: B.1.525 (n=797), B.1.1.318 (n=398) (*42*), B.1.1.418 (n=395), A.23.1 (n=358) (*15*, *29*, *31*, *33*), C.1 (n=446) (*29*), C.1.2 (n=300) (*31*), C.36 (n=305) (*30*, *43*), B.1.1.54 (n=287) (*15*, *29*, *31*, *33*), B.1.416 (n=272), B.1.177 (n=203), B.1.620 (n=138), and B.1.160 (n=61), (*32*) (**Appendix F - Supp Fig S10A,B**). Our discrete state phylogeographic inference from phylogenetic reconstruction of non-VOC African sequences and an equal number of external references revealed that African countries were primarily seeded by multiple introductions of viral lineages from abroad (mainly Europe) at the beginning of the pandemic. The observed pattern of non-VOC viral lineage movement then consistently shifted towards more intercontinental exchanges (**Appendix F - Supp Fig S10C**). Mapping out the spatial routes of dissemination shows that various countries in all subregions of the continent acted as sources of these viral lineages at one point or another (**Appendix F - Supp Fig S10D**). While uneven testing rates and proportions of samples sequenced on the continent may have influenced these inferences (discussed below), the results presented here are in line with the fact that these most predominant non-VOC lineages in Africa, except B.1.177, emerged and circulated widely in different sub-regions (**Figure 7.1**).

Similar to the pandemic globally, VOCs became increasingly important in Africa towards the end of 2020. The Alpha, Beta, Delta and Omicron variants demonstrate many similarities as well as differences in the way they spread on the continent. For all these VOCs, we observe large regional monophyletic transmission clusters in each of their phylogenetic reconstructions in Africa (**Appendix F - Supp Fig S11**). This suggests an important extent of continental dissemination within Africa. Alpha and Beta were epidemiologically important in distinct regions of the continent with Alpha primarily circulating in West, North and most of Central Africa, Beta in southern and most of East Africa, and only substantially co-circulated in a few countries such as Angola, Kenya, Comoros, Burundi and Ghana (**Fig 7.1, Appendix F - Supp Fig S12**). However, we may not have enough resolution in the geospatial data to know how much they were truly co-circulating throughout these countries, or whether there were regional outbreaks of Alpha and Beta within these countries. In Kenya, for example, Beta was detected more in coastal regions, and Alpha more inland (*26*, *44*). In contrast, Delta and Omicron variants sequentially dominated the majority of infections on the entire continent shortly after their emergence (**Figure 7.4A, Appendix F - Supp Fig S12**).

The Alpha variant was first identified in December 2020 in the UK and has since spread globally. In Africa, Alpha was detected in 43 countries with evidence of community transmission, based

on phylogenetic clustering, in many countries including Ghana, Nigeria, Kenya, Gabon and Angola (**Appendix F - Supp Fig S11**). Discrete state maximum likelihood reconstruction from a globally case-sensitive genomic subsampling inferred at least 80 introductions (95% CI: 78 - 82) into Africa with the bulk of imports attributed to the US (>47%) and the UK (>25%) (**Figure 7.4B**). Only 1% of imports into any particular African country were attributed to another African nation. Phylogeographic reconstruction enriched in African sequences revealed that of those, >85% of the intercontinental Alpha exchanges in Africa originated from West African countries (**Figure 7.4C**). This occurred in spite of initial importations of the Alpha variant from Europe into all regions of the continent (**Appendix F - Supp Fig S13B**), but is in line with Alpha having dominated circulation mostly in West Africa (**Appendix F - Supp Fig S12**). In countries where Alpha was introduced but did not grow and cause an expansion of cases, this can be explained by competition with the already established Beta variant, which simultaneously circulated. The characteristics of multiple introductions of Alpha intro Africa and between African countries is similar to the spread of Alpha documented in the UK, Scotland and Ireland (*45–47*).

The second VOC, Beta, was identified in December 2020 in South Africa (*4*). However, sampling and molecular clock analyses suggest that the variant originated around September 2020 (**Appendix F - Supp Fig S11**). At the end of 2020 and beginning of 2021, Beta was driving a second wave of infection in South Africa and quickly spread to other countries within the region. The concurrent introductions and spread of Alpha and other variants (Eta, A.23.1) in other regions of the continent may have reduced the Beta variant's initial growth, limiting its spread to largely southern Africa, and to a lesser extent the East Africa region. Beta spread to at least 114 countries globally, including 37 countries and territories in Africa. For this variant, viral circulation and geographical exchanges occurred predominantly within the continent. Indeed, phylogeographic reconstruction from a globally case-sensitive sampling revealed that of the 810 (95% CI: 803 - 818) inferred introductions of the Beta variant into African countries, only 110 (95% CI: 105 - 115; 13%) were attributed to sources outside the continent (**Appendix F - Supp Fig 13C**), while more than half of introductions were attributed to South Africa (63%) (**Figure 7.4C**). This is in line with expectations as the variant originated in South Africa. Beyond southern Africa, most of the introductions back into the continent were attributed to France and other EU countries into the French overseas territories, Mayotte and Reunion, and other Francophone African countries. Africa-focused phylogeographic analysis revealed a similar spatial pattern showing southern countries as substantial sources of the variant, followed in small numbers by countries in East Africa (**Figure 7.4C**).

The fourth VOC observed was Delta (*13*), which rose to prominence in April 2021 in India, where it fueled an explosive second wave. Since its emergence, Delta was detected in >170 countries, including 37 African countries and territories (**Appendix F - Supp Fig S11**). Our global case-sensitive subsampled analysis infers at least 100 (95% CI: 93 - 106) introductions of the Delta variant into Africa, with the bulk attributed to India (~72%), mainland Europe (~8%), the UK

(~5%), and the US (~2.5%). Viral introductions of Delta also occurred from one African country to others, in 7% of inferred introductions. From our Africa-focused phylogeographic inferences, we infer that viral dissemination of Delta within Africa was not restricted to or dominated by any particular region unlike Alpha and Beta, but rather spread across the entire continent (**Figure 7.4C**). Following introductions from Asia in the middle of 2021, Delta rapidly replaced the other circulating variants (**Figure 7.4A**). For example, in southern African countries, the Delta variant rapidly displaced Beta and by June-2021 was circulating at very high (>90%) frequencies (*48*).

The latest VOC, Omicron, was identified and characterized in November 2021, in southern Africa (*3*). At the time of writing, the variant has been detected and caused waves of infections in >160 countries including 39 African countries and two overseas territories (**Appendix F - Supp Fig S11**). Due to the genetic distance between them and their sequential epidemic expansion globally (rather than simultaneous), phylogenies were reconstructed separately for Omicron BA.1 and BA.2. Our discrete ancestral state reconstruction from a global case-sensitive sampling for Omicron BA.1 infers at least 55 (95% CI: 47 - 62) viral exports of BA.1 out of various African countries, of which 31 (95% CI: 25 - 36) were towards Europe and 8 (95% CI: 6 - 10) towards North America (**Figure 7.4B**). Following explosive expansion of Omicron around the world, we inferred even more reintroductions of the variant back into Africa, at least 69 (95% CI: 60 - 78) from Europe and 102 (95% CI: 92 - 112) from North America (**Figure 7.4B**). From our Africa-focused phylogeographic reconstructions, we determine that, as with Delta, routes of dissemination of this variant involved all regions of the continent spatially (**Figure 7.4C**). Yet, ~75% of all BA.1 viral movement volume in Africa happened between southern African countries, likely due to rapid epidemic expansion in the region soon after its detection (*3*). Omicron BA.2's reach in Africa was limited at the time of writing, with only 3 260 sequences from 19 countries attributed to BA.2 on GISAID (Date of access: 2022-03-31) (15% of all Omicron sequences from Africa). Our discrete ancestral state reconstruction from a global case-sensitive sampling for Omicron BA.2 infers at least 68 (95% CI: 53 - 84) viral exports out of African countries, of which the majority were towards Europe (~88%) (**Figure 7.4B**). We also infer at least 99 (95% CI: 87 - 109) separate introduction or reintroduction events of BA.2 back into African countries, of which ~65% are from Europe and ~30% from Asia, primarily from India (**Figure 7.4B**). This is consistent with India having experienced one of the earliest large BA.2 waves globally. In the context of global incidence of BA.2, this case-sensitive phylogeographic analysis revealed that only 0.01% of viral movements of this lineage globally happened from one African country to another. Our Africa-focused analysis inferred a similar pattern of BA.2 spatial diffusion within African to BA.1 (**Figure 7.4C**). However, given that this accounted for such a small percentage of global BA.2 movements, BA.2 diffusion from one African country to another is unlikely to have had a significant impact on epidemiological expansion, compared to introductions from Asia, Europe or North America.

Globally, dissemination of the SARS-CoV-2 virus throughout the pandemic was intricately linked with human mobility patterns (*49–53*). To determine the validity of the VOC movement patterns that we infer into and within the Africa continent in this study, we compared viral import and export events to and from South Africa with travel to the country. In December 2020, the UK accounted for the 5th highest number of passengers entering South Africa, while other countries with the top 9 sources of travelers were all neighboring countries in southern Africa (**Appendix F - Supp Fig S14A**). Considering that incidence of the Alpha variant was insignificant in the region, this supports our inference of the UK contributing 60% of Alpha introductions to South Africa (**Appendix F - Supp Fig S15A**). In March 2021, the US, Germany, the UK and India were among the top 12 sources of travelers to South Africa behind 8 African countries (**Appendix F - Supp Fig S14B**). During this time of Delta dissemination globally, we infer that ~90% of introductions of Delta into South Africa originated in the UK, the US and India (**Appendix F - Supp Fig S15B**). At the end of 2021, most introductions or re-introductions of Omicron to the country came from the UK, the US or Botswana, corresponding to locations of both high Omicron incidence at the time, and high numbers of passengers to South Africa (**Appendix F - Supp Fig S14C, S15C**). These travel patterns also fit the findings that ~89%,~70% and ~75% of Beta, Delta and Omicron exports respectively from South Africa to other African countries were directed to locations of southern Africa (**Appendix F - Supp Fig S14D-E, S15D-E**).

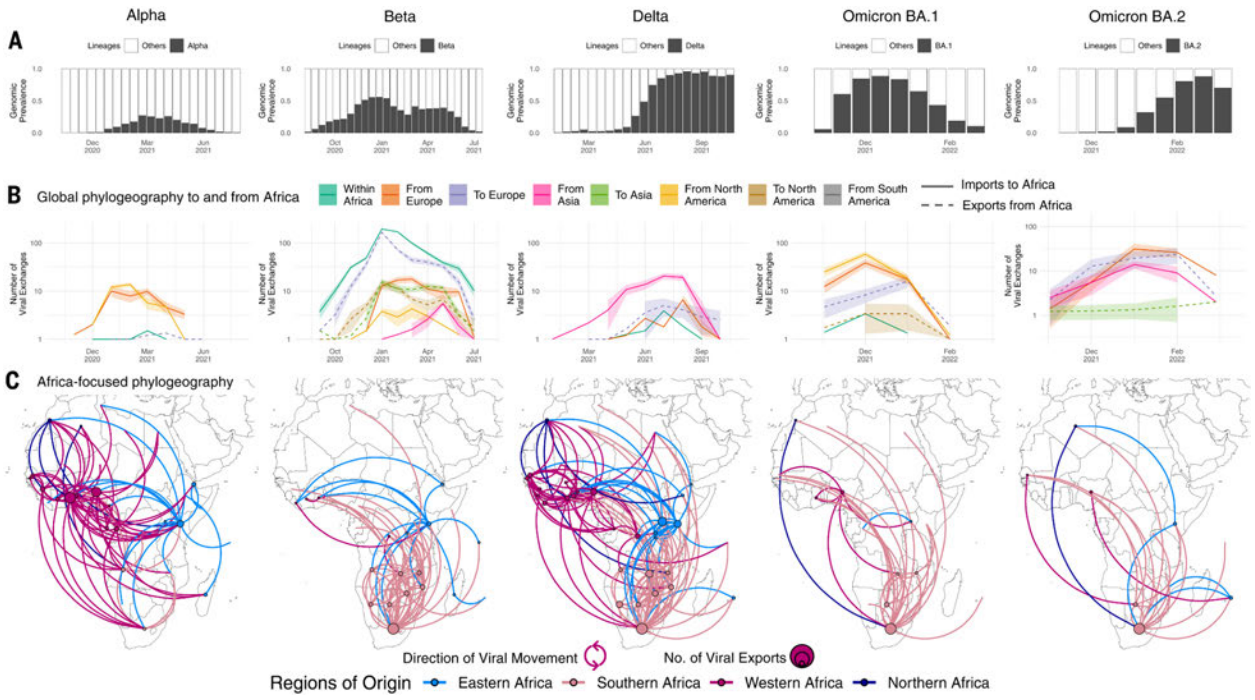


***Figure 4**: Inferred viral dissemination patterns of VOCs within Africa. A) Genomic prevalence of VOCs Alpha, Beta, Delta and Omicron in Africa over time. B) Inferred viral exchange patterns to, from and within the Africa continent for the four VOCs (Omicron as BA.1 and BA.2) based on case-sensitive phylogeographic inference. Introductions and viral transitions within Africa are shown in solid lines and exports from Africa are shown in dotted lines and these are coloured by*

continent. The shaded areas around the lines represent uncertainty of this analysis from ten replicates (+/- s.d.). C) Dissemination patterns of the VOCs within Africa, from inferred ancestral state reconstructions performed on Africa enriched datasets, annotated and coloured by region in Africa. The countries of origin of viral exchange routes are also shown with dots and the curves go from country of origin to destination country in an anti-clockwise direction.

Discussion, Limitations and Conclusions

By April 2020, a total of 20 African countries were able to sequence the virus within their own borders. This was largely made possible by other preexisting sequencing efforts on the continent focused on other human pathogens (e.g. HIV, TB, Ebola and H1N1). However, these efforts were quickly limited by global supply chain issues and in many countries sequencing efforts dramatically slowed down or stopped towards the end of 2020. In order to facilitate more sequencing on the continent over the course of the past year (April 2021 - March 2022) the Africa CDC and partners invested heavily to support genomic surveillance on the continent. This included the transfer of 24 new sequencing platforms (including MinIon, GridIon, MiSeq and NextSeq), the distribution of reagents and flow cells to support the sequencing of 100 000 positive samples, the training of >230 students and technicians in wet laboratory and bioinformatic techniques and additional grants to support 10 regional sequencing hubs. This investment has started bearing fruit and should be intensified as the virus continues to evolve, requiring the adaptation of methodologies locally on the continent to keep pace with the emergence of variants. The continued development of sequencing protocols in Africa is of crucial importance (*41*, *54*, *55*) given the number of variants and lineages that emerged in, and were introduced to, the continent. In Northern Africa, the SARS-CoV-2 pandemic was caused by waves of infections that were similar to those seen in Europe (first wave = B.1 descendents, second wave = Alpha, third wave = Delta and forth wave = Omicron), in southern Africa the pattern was similar but with a Beta wave instead of an Alpha one. In East Africa, the pandemic was more complex, involving both Alpha and Beta as well as its own lineage A.23.1 before the arrival of Delta and Omicron. Central Africa experienced epidemic patterns sometimes mirroring East Africa and other times southern Africa. In West Africa, Eta made a significant contribution to both a second wave (together with alpha) and a third wave (together with Delta). The factors that resulted in these regional differences are not clear but could be due to differences in human mobility, founder effects, competition between lineages or the immunity induced by earlier waves in a region.

Public health benefits of such broadly inclusive genomic surveillance are manifold. The most prominent insight from this expanded genomic surveillance in Africa has been an early warning capacity for the world following the detection of new lineages and variants, most recently relevant in the detection of Omicron BA.1, BA.2, BA.3, BA.4 and BA.5 sub-variants (*3*, *4*, *34*). Furthermore, the reporting of local SARS-CoV-2 sequences made the epidemic more immediate to the Ministries of Health from the reporting African countries. It became clear early on that the

viral evolution is global and the transmission of the virus is extremely rapid which guided mitigation strategies. The generation and the availability of local sequences also validated local diagnostics and allowed investigators to determine if nucleic acid based diagnostics in use could still detect local variants. The detection of SARS-CoV-2 in returning travelers and truck drivers indicated routes that the virus might be using to enter a country and guided early efforts to slow the virus entry and gain time to establish vaccination plans. Later the difficulty of stopping the virus at borders combined with the data that the variants were already in community circulation allowed public health officials to focus efforts and limited resources on vaccination rather than on border controls. The detection and reporting of the more recent lineages with enhanced transmission (i.e. Omicron) and the ability to bypass existing immunity is important information and an early alert to the public health officials globally that the epidemic was still proceeding. As the pandemic progresses in an evolving global context, we provide evidence that with each new variant, transmission dynamics are changing and the use of sequencing with phylogenetics could potentially alter decisions of public health measures. For example, the demonstrated shift away from regional dynamics of Alpha and Beta towards more global patterns with Delta and Omicron can provide insights to public health officials as they anticipate epidemic developments locally. With Omicron it became clear that although the variant expanded first in Africa, the continent ultimately had a minimal role in global dissemination, and continental expansion beyond southern Africa was most influenced by external introductions, in contrast to the Beta variant. All of these public health benefits to sequencing SARS-CoV-2 is primarily amplified, as we show in this study, if the sequencing can be conducted locally within a country, which strongly supports the continued investment into pathogen sequencing on the continent.

In spite of the recent successful expansion of genomics surveillance in Africa, additional work remains necessary. Even with the Africa CDC - Africa PGI's and other investments, there are still 16 countries with no sequencing capacity within their own borders. These countries' only option is to send samples to continental sequencing hubs or to centers outside of the continent, which increases the turnaround times and limits the utility of genomic surveillance for public health decision making. Secondly, not all countries are willing to share data openly in a timely fashion for fear of being subject to travel bans or restrictions which could bring substantial economic harm. Such hesitancy has obvious potential ramifications for the future of genomic surveillance on the continent. Furthermore, with the expansion of sequencing on the continent there is a growing need for more bioinformatics support and knowledge to allow investigators to analyze and report their data in a reasonable timeframe that makes it useful for public health response. It is also clear the SARS-CoV-2 sequencing primers are not a static development and may require updating as the virus evolves. A number of research groups have been addressing the SARS-CoV-2 sequencing primer questions. Issues of gaps in the genomes due to missing amplicons have been discussed (*56*, *57*). The ARTIC primer set has gone through a number of revisions to accommodate virus evolution (*39*, *40*). Additional longer amplicon methods have been published (*58–60*) including methods to use a subset of ARTIC primers (*61*).

The patterns we describe here are of course limited to reported cases, and applies to both the phylogeographic as well as the epidemiology inferences. As such, the results need to be interpreted with these limitations in mind. Our primary phylogeographic inference relied on a sampling strategy considering all high quality African sequences and an equal number of external references. Though this strategy has the advantage of placing all African sequences in a phylogenetic context, it introduces a bias when applied to discrete ancestral state reconstruction as more internal nodes are inferred to be from Africa. To address this we performed an even sampling of global cases, based on reported case counts through time, to compare against our over sampled inference. The even sampling approach has the benefit that the discrete ancestral state reconstruction is not biased by uneven sampling. Comparing the two there are obvious differences, most notably that the number of inferred introductions into Africa is proportional to sampling proportions (**Supp Figure S16**), as we no longer consider all African sequences but just a small subset against a global sample. However, inferences from the two approaches correspond well with one another. For example, considering Alpha we still observed the vast majority of introductions into Africa to originate from Western Europe. Patterns of dissemination within Africa are more robustly comparable between the two, for instance that countries in West Africa were the biggest source of Alpha within the continent. High concordance between the two inference methods were also observed for other VOCs for dispersal routes within Africa which gives us confidence in the inferred patterns we observe here. Although we represent an inference based on over sampling and case sensitive sampling, it is currently not possible to explore how undersampling affects the phylogeographic reconstruction due to uneven testing rates. Additionally, the robustness of the phylogeographic inference can also be affected by the underlying methodology used. Broad consensus would favor the use of Bayesian methods for phylogeographic reconstruction, which is often considered to be the "gold standard" in the field. The main drawbacks of Bayesian methods are that they can only be applied to a relatively small number of sequences at a time (<1,000) and are extremely computationally and time intensive. Given the explosion of sequence data over the past two years, the scientific community will have to adapt or put forth new analytical methods to fully capitalize on the global sequencing efforts for SARS-CoV-2.

Despite our best attempts to consider and minimize genomic sampling bias, the accuracy of the resulting phylogenetic inferences is limited by the available epidemiological and genomic data, leading to unaccounted biases in the estimates of viral movements. This includes limited testing and subsequent sequencing in many African countries. Although the percentage of reported cases sequenced in African countries (0.01 - 10%, mean = 1.27%) is not far from global figures (0.01-16%, mean = 1.31%), testing rates and infection-to-detection ratios in Africa were some of the lowest globally (*38*, *62*). Together with estimates of excess mortality being as much as 20-fold more than the reported numbers in African countries (*63*), these are strong indications of undetected and underreported epidemic sizes in Africa, leading to undersampling of genomic data (*62*) and thus underestimates of viral exchange inferences in our study. Some countries with no

publicly available SARS-CoV-2 sequences are by definition completely missing in our inference. This in turn means that inferred routes of viral transmission within Africa could be missing important intermediate locations, although this is potentially true around the world. Nevertheless, we believe that the viral movement inferences that we discuss in this study provide a likely qualitative description of the patterns of SARS-CoV-2 migration into, out of, and within Africa.

Finally, we should also mention uneven sequencing and reporting standards across the different laboratories on the continent - and globally, for that matter. Different groups use different measures for what constitutes a high quality sequence (e.g. 70% vs 80% sequence coverage) or using different sequencing depth coverage. This lack of standardization globally complicates the direct comparison of sequences that may have been submitted to GISIAD using different criteria further biasing any inference. Given the sheer size of SARS-CoV-2 sequencing, with ~10 million whole genome sequences shared on the GISAID database (31st March 2022), there is an urgent need for global standards with regards to sequence quality and associated metadata.

In conclusion, Africa needs to continue expanding genomic sequencing technologies on the continent in conjunction with diagnostics capabilities. This holds true not just for SARS-CoV-2 but for other emerging or re-emerging pathogens on the continent. For example, WHO announced in February 2022 the re-emergence of wild polio in Africa, while sporadic influenza H1N1, measles and Ebola outbreaks continue to occur on the continent. The Africa CDC has estimated that over 200 pathogen outbreaks are reported across the continent every year. Beyond the current pandemic, continued investment in diagnostic and sequencing capacity for these pathogens could serve the public health of the continent well into the 21st century.

Methods

Ethics statement

This project relied on sequence data and associated metadata publicly shared by the GISAID data repository and adhere to the terms and conditions laid out by GISAID (*16*). The African samples processed in this study were obtained anonymously from material exceeding the routine diagnosis of SARS-CoV-2 in African public and private health laboratories. Individual institutional review board (IRB) references or material transfer agreements (MTAs) for countries are listed below.

Angola - (MTA - CON8260), Botswana - Genomic surveillance in Botswana was approved by the Health Research and Development Committee (Protocol HPDME 13/18/1), Egypt - Surveillance in Egypt was approved by the Research Ethics Committee of the National Research Centre (Egypt) (protocol number 14 155, dated March 22, 2020), Kenya - samples were collected under the Ministry of Health protocols as part of the national COVID-19 public health response. The whole genome sequencing study protocol was reviewed and approved by the Scientific and Ethics

Review Committee (SERU) at Kenya Medical Research Institute (KEMRI), Nairobi, Kenya (SERU protocol #4035), Nigeria – (NHREC/01/01/2007), Mali - study of the sequence of SARS-CoV-2 isolates in Mali - Letter of Ethical Committee (N0-2020 /201/CE/FMPOS/FAPH of 09/17/2020), Mozambique - (MTA - CON7800), Malawi - (MTA - CON8265), South Africa - The use of South African samples for sequencing and genomic surveillance were approved by University of KwaZulu-Natal Biomedical Research Ethics Committee (ref. BREC/00001510/2020); the University of the Witwatersrand Human Research Ethics Committee (HREC) (ref. M180832); Stellenbosch University HREC (ref. N20/04/008_COVID-19); the University of the Free State Research Ethics Committee (ref. UFS-HSD2020/1860/2710) and the University of Cape Town HREC (ref. 383/2020), Tunisia - for sequences derived from sampling in Tunisia, all patients provided their informed consent to use their samples for sequencing of the viral genomes. The ethical agreement was provided to the research project ADAGE (PRFCOVID19GP2) by the Committee of protection of persons (Tunisian Ministry of Health) under the reference (CPP SUD N 0265/2020), Uganda - The use of samples and sequences from Uganda were approved by the Uganda Virus Research Institute - Research and Ethics Committee UVRI-REC Federalwide Assurance [FWA] FWA No. 00001354, study reference - GC/127/20/04/771 and by the Uganda National Council for Science and Technology, reference number - HS936ES) and Zimbabwe (MTA - CON8271).

Epidemiological and genomic data dynamics

We analyzed trends in daily numbers of cases of SARS-CoV-2 in Africa up to 31st March 2022 from publicly released data provided by the Our World in Data repository for the continent of Africa (https://github.com/owid/covid-19-data/tree/master/public/data) as a whole and for individual countries (*2*). To provide a comparable view of epidemiological dynamics over time in various countries, the variable under primary consideration for **Figure 1** was 'new cases per million (smoothed)'. To calculate the genomic sampling proportion and frequency for each country for **Figure 2**, the total number of recorded cases at 31st March was considered, as well as the total length of time for which each country has recorded cases of SARS-CoV-2.

Genomic metadata was downloaded for all African entries on GISAID for the same time period (date of access: 31st March 2022). From this, information extracted from all entries for this study included: date of sampling, country of sampling, viral lineage and clade, originating laboratory, sequencing laboratory, and date of submission to the GISAID database. The geographical locations of the originating and sequencing laboratories were manually curated. Sequences originating and sequenced in the same country were defined as locally sequenced, irrespective of specific laboratory or finer location. Sequences originating in one African country and sequenced in another were defined as sequenced within regional sequencing networks. Sequences sequenced in a location not within Africa were labeled as sequenced outside Africa. Sequencing turnaround time was defined as the number of days elapsed from specimen collection to sequence submission to

GISAID. Sequencing technology information for all African entries was also downloaded from GISAID on 31st March 2022.

Primer choice and sequencing outcomes

All SARS-CoV-2 genomes from African countries were retrieved from GISAID (*16*) for submission dates from 1 December 2019 to 31st March 2022 yielding 100 470 entries. Associated metadata for the entries were also retrieved, including collection date, submission date, country, viral strain and sequencing technology. Data on the primers used for the sequencing were requested from investigators and yielded primer data for 13 973 of the entries (~13%). The total N (bases with low sequence depth) per genome were counted, results from which were then used for genome quality analysis and visualization. Gap locations in the genomes were mapped and visualized compared to the original Wuhan strain (*64*).

Phylogenetic investigation

All African sequences on the GISAID sequence database (*16*) were downloaded on the 31st of March 2022 (n=100 470). Of this, Alpha accounted for 3 851 sequences, Beta accounted for 14 548 sequences, Delta accounted for 35 027 sequences, Omicron for 21 708, while 25 336 sequences were classified as none-VOCs. Prior to any phylogenetic inference we performed some quality assessment on the sequences to exclude incomplete or problematic sequences as well as sequences lacking complete metadata. Briefly, all African sequences were passed through the NextClade analysis pipeline (*65*) in order to identify and exclude: (*i*) sequences missing >10% of the SARS-CoV-2 genome, (*ii*) sequences that deviate by >70 nucleotides from the Wuhan reference strain, (*iii*) sequences with >10 ambiguous bases, (*iv*) clustered mutations, and (*v*) sequences flagged with private mutations as problematic by NextClade. Additionally, Omicron variants were screened for traces of viral recombination with RDP5.23 (*66*) using default settings and a p-value of ≤0.05 as evidence of recombination. using a P-value 0.05 or lower cut-off as evidence of recombination. A large number of sequences were removed (n=57 421) with incomplete sequences (<90% genome coverage) being the biggest contributor. This produced a final African dataset of 43 049 high quality African sequences. Due to the sheer size of the dataset we opted to perform independent phylogenetic inferences on the main VOCs (Alpha, Beta, Delta and Omicron BA.1 and BA.2) that have spread on the African continent, as well as a separate inference for all non-VOC SARS-CoV-2 sequences.

In order to evaluate the spread of the virus on the African continent we aligned the African datasets against a large number of globally representative sequences from around the world. Due to the oversampling of some variants or lineages we performed a random downsampling while retaining the oldest two known variants from each country. Reference sequences were respectively aligned with their African counterparts independently with NextAlign (*65*). Each of the alignments were

then used to infer maximum likelihood (ML) tree topologies in FastTree v 2.0 (*67*) using the General Time Reversible (GTR) model of nucleotide substitution and a total of 100 bootstrap replicates (*68*). The resulting ML tree topologies were first inspected in TempEst (*69*) to identify any sequences that deviate more than 0.0001 from the residual mean. Following the removal of potential outliers in R with the ape package (*70*), the resulting ML-trees were then transformed into time calibrated phylogenies in TreeTime (*71*) by applying a rate of 8x10e-4 substitution per site per year (*72*) in order to transform the branches into units of calendar time. Time calibrated trees were then visualized along with associated metadata in R using ggtree (*73*) and other packages.

We performed a basic viral dispersal analysis for each of the VOCs (excluding Gamma), as well as for the non-VOC dataset. Briefly, a migration model was fitted to each of the time calibrated tree topologies in TreeTime, mapping the country location of sampled sequences to the external tips of the trees. The *mugration* model of TreeTime also infer the most likely location for internal nodes in the trees. Using a custom python script we could then count the number of state changes by iterating over each phylogeny from the root to the external tips. We count state changes when an internal node transitions from one country to a different country in the resulting child-node or tip(s). The timing of transition events is then recorded which serve as the estimated import or export event. To infer some confidence around these estimates, we performed ten replicates for each of the dataset by random selection from the 100 bootstrap trees. Due to the high uncertainty in the inferred locations for deep internal nodes in the trees we truncated state changes to the earliest date of sampling in each dataset. All data analytics were performed using custom python and R scripts and results visualized using the ggplot libraries (*74*). Such phylogeographic methods are always subject to uneven sampling through time (i.e. over the course of the pandemic) and through space (by sampling location). To address this we have performed a case sensitive analysis to investigate the effects of oversampling African locations on the inferred number of viral introductions. Furthermore, in a previous analysis (*15*) we performed a sensitivity analysis to address some of these issues and found no substantial variations in estimates.

Case sensitive phylogeographic inference

To address the potential over sampling of African sequences relative to global reference in the above mentioned analyses, we performed another phylogeographic inference on subsamples based on global case counts to try and eliminate oversampling bias in our inference. To this end, we considered all high quality sequences for each of the VOCs (Alpha, Beta, Delta and Omicron BA.1 and BA.2) globally over the same sampling period (till 31st of March 2022). We used subsampler (https://github.com/andersonbrito/subsampler) to generate subsamples for each variant based on globally reported cases. In short, subsampler uses a case count matrix of daily cases, along with the fasta sequences and GISAID associated metadata to sample a user defined number of sequences. For each VOC and for BA.1 and BA.2 we performed 10 samplings using different

number seeds in order to sample datasets of ~20 000. Once again, sampled sequences were screened for viral recombination as described above and sequences with signs of recombination were removed. Subsampler has the added advantage that it disregards poor quality sequences (e.g. <90% coverage) and sequences with missing metadata (e.g. exact date of sampling). Each dataset was then subjected to the same analytical pipeline as mentioned above to infer the viral transitions between Africa and the rest of the world.

Regional and country specific Nextstrain builds

In order to investigate more granular changes in lineage dynamics within a specific country or region in Africa we utilized the Nextstrain pipeline (https://github.com/nextstrain/ncov) to generate the regional and country-specific builds for African countries (*75*). Firstly, all sequence data and metadata were retrieved from the GISAID sequence database and filtered for Africa based on the 'region' tab, for inclusion in regional- and country-specific African builds. For country-specific builds ~4 000 sequences from a given country were randomly selected and analyzed against ~1 000 randomly selected sequences from the Africa 'nextregions' records that do not match the focal country of interest. For region specific (e.g. West Africa), ~4 000 sequences from the focal region are selected at random and analyzed against ~1 000 randomly selected sequences from the Africa 'nextregions' records that do not match the focal region of interest. The methodological pipeline for NextStrain is well documented and performs all analyses within one workflow, including filtering of sequences, alignment, tree inference, molecular clock and ancestral state reconstruction. For more information please visit, https://docs.nextstrain.org/en/latest/index.html.

All region- and country-specific builds are regularly updated to keep track of the evolving pandemic on the continent. All builds are publicly available under the links provided in **Supp Tables S1** and **S2** as well as on the NextStrain webpage.

References

1. Q. Li, X. Guan, P. Wu, X. Wang, L. Zhou, Y. Tong, R. Ren, K. S. M. Leung, E. H. Y. Lau, J. Y. Wong, X. Xing, N. Xiang, Y. Wu, C. Li, Q. Chen, D. Li, T. Liu, J. Zhao, M. Liu, W. Tu, Z. Feng, Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus-Infected Pneumonia. *N. Engl. J. Med.* **382**, 1199–1207 (2020).

2. J. Hasell, E. Mathieu, D. Beltekian, B. Macdonald, C. Giattino, E. Ortiz-Ospina, M. Roser, H. Ritchie, A cross-country database of COVID-19 testing. *Sci. Data*. **7**, 345 (2020).

3. R. Viana, S. Moyo, D. G. Amoako, H. Tegally, C. Scheepers, C. L. Althaus, U. J.

Anyaneji, P. A. Bester, M. F. Boni, M. Chand, W. T. Choga, R. Colquhoun, M. Davids, K. Deforche, D. Doolabh, L. du Plessis, S. Engelbrecht, J. Everatt, J. Giandhari, M. Giovanetti, T. de Oliveira, Rapid epidemic expansion of the SARS-CoV-2 Omicron variant in southern Africa. *Nature*. **603**, 679–686 (2022).

4. H. Tegally, E. Wilkinson, M. Giovanetti, A. Iranzadeh, V. Fonseca, J. Giandhari, D. Doolabh, S. Pillay, E. J. San, N. Msomi, K. Mlisana, A. von Gottberg, S. Walaza, M. Allam, A. Ismail, T. Mohale, A. J. Glass, S. Engelbrecht, G. Van Zyl, W. Preiser, T. de Oliveira, Detection of a SARS-CoV-2 variant of concern in South Africa. *Nature*. **592**, 438–443 (2021).

5. D. P. Martin, S. Weaver, H. Tegally, J. E. San, S. D. Shank, E. Wilkinson, A. G. Lucaci, J. Giandhari, S. Naidoo, Y. Pillay, L. Singh, R. J. Lessells, NGS-SA, COVID-19 Genomics UK (COG-UK), R. K. Gupta, J. O. Wertheim, A. Nekturenko, B. Murrell, G. W. Harkins, P. Lemey, S. L. Kosakovsky Pond, The emergence and ongoing convergent evolution of the SARS-CoV-2 N501Y lineages. *Cell*. **184**, 5189-5200.e7 (2021).

6. F. Campbell, B. Archer, H. Laurenson-Schafer, Y. Jinnai, F. Konings, N. Batra, B. Pavlin, K. Vandemaele, M. D. Van Kerkhove, T. Jombart, O. Morgan, O. le Polain de Waroux, Increased transmissibility and global spread of SARS-CoV-2 variants of concern as at June 2021. *Euro Surveill.* **26** (2021), doi:10.2807/1560-7917.ES.2021.26.24.2100509.

7. B. Korber, W. M. Fischer, S. Gnanakaran, H. Yoon, J. Theiler, W. Abfalterer, N. Hengartner, E. E. Giorgi, T. Bhattacharya, B. Foley, K. M. Hastie, M. D. Parker, D. G. Partridge, C. M. Evans, T. M. Freeman, T. I. de Silva, Sheffield COVID-19 Genomics Group, C. McDanal, L. G. Perez, H. Tang, D. C. Montefiori, Tracking Changes in SARS-CoV-2 Spike: Evidence that D614G Increases Infectivity of the COVID-19 Virus. *Cell*. **182**, 812-827.e19 (2020).

8. E. Hacisuleyman, C. Hale, Y. Saito, N. E. Blachere, M. Bergh, E. G. Conlon, D. J. Schaefer-Babajew, J. DaSilva, F. Muecksch, C. Gaebler, R. Lifton, M. C. Nussenzweig, T. Hatziioannou, P. D. Bieniasz, R. B. Darnell, Vaccine Breakthrough Infections with SARS-CoV-2 Variants. *N. Engl. J. Med.* **384**, 2212–2218 (2021).

9. D. Planas, D. Veyer, A. Baidaliuk, I. Staropoli, F. Guivel-Benhassine, M. M. Rajah, C. Planchais, F. Porrot, N. Robillard, J. Puech, M. Prot, F. Gallais, P. Gantner, A. Velay, J. Le Guen, N. Kassis-Chikhani, D. Edriss, L. Belec, A. Seve, L. Courtellemont, O. Schwartz, Reduced sensitivity of SARS-CoV-2 variant Delta to antibody neutralization. *Nature*. **596**, 276–280 (2021).

10. S. Yue, Z. Li, Y. Lin, Y. Yang, M. Yuan, Z. Pan, L. Hu, L. Gao, J. Zhou, J. Tang, Y. Wang, Q. Tian, Y. Hao, J. Wang, Q. Huang, L. Xu, B. Zhu, P. Liu, K. Deng, L. Wang, X. Chen, Sensitivity of SARS-CoV-2 Variants to Neutralization by Convalescent Sera and a VH3-30 Monoclonal Antibody. *Front. Immunol.* **12**, 751584 (2021).

11. S. Cele, I. Gazy, L. Jackson, S.-H. Hwa, H. Tegally, G. Lustig, J. Giandhari, S.

Pillay, E. Wilkinson, Y. Naidoo, F. Karim, Y. Ganga, K. Khan, M. Bernstein, A. B. Balazs, B. I. Gosnell, W. Hanekom, M.-Y. S. Moosa, Network for Genomic Surveillance in South Africa, COMMIT-KZN Team, A. Sigal, Escape of SARS-CoV-2 501Y.V2 from neutralization by convalescent plasma. *Nature*. **593**, 142–146 (2021).

12. B. Meng, S. A. Kemp, G. Papa, R. Datir, I. A. T. M. Ferreira, S. Marelli, W. T. Harvey, S. Lytras, A. Mohamed, G. Gallo, N. Thakur, D. A. Collier, P. Mlcochova, COVID-19 Genomics UK (COG-UK) Consortium, L. M. Duncan, A. M. Carabelli, J. C. Kenyon, A. M. Lever, A. De Marco, C. Saliba, R. K. Gupta, Recurrent emergence of SARS-CoV-2 spike deletion H69/V70 and its role in the Alpha variant B.1.1.7. *Cell Rep.* **35**, 109292 (2021).

13. P. Mlcochova, S. A. Kemp, M. S. Dhar, G. Papa, B. Meng, I. A. T. M. Ferreira, R. Datir, D. A. Collier, A. Albecka, S. Singh, R. Pandey, J. Brown, J. Zhou, N. Goonawardane, S. Mishra, C. Whittaker, T. Mellan, R. Marwal, M. Datta, S. Sengupta, R. K. Gupta, SARS-CoV-2 B.1.617.2 Delta variant replication and immune evasion. *Nature*. **599**, 114–119 (2021).

14. N. R. Faria, T. A. Mellan, C. Whittaker, I. M. Claro, D. da S. Candido, S. Mishra, M. A. E. Crispim, F. C. S. Sales, I. Hawryluk, J. T. McCrone, R. J. G. Hulswit, L. A. M. Franco, M. S. Ramundo, J. G. de Jesus, P. S. Andrade, T. M. Coletti, G. M. Ferreira, C. A. M. Silva, E. R. Manuli, R. H. M. Pereira, E. C. Sabino, Genomics and epidemiology of the P.1 SARS-CoV-2 lineage in Manaus, Brazil. *Science*. **372**, 815–821 (2021).

15. E. Wilkinson, M. Giovanetti, H. Tegally, J. E. San, R. Lessells, D. Cuadros, D. P. Martin, D. A. Rasmussen, A.-R. N. Zekri, A. K. Sangare, A.-S. Ouedraogo, A. K. Sesay, A. Priscilla, A.-S. Kemi, A. M. Olubusuyi, A. O. O. Oluwapelumi, A. Hammami, A. A. Amuri, A. Sayed, A. E. O. Ouma, et al., A year of genomic surveillance reveals how the SARS-CoV-2 pandemic unfolded in Africa. *Science*. **374**, 423–431 (2021).

16. Y. Shu, J. McCauley, GISAID: Global initiative on sharing all influenza data - from vision to reality. *Euro Surveill.* **22**, 30494 (2017).

17. C. Kuiken, B. Korber, R. W. Shafer, HIV sequence databases. *AIDS Rev.* **5**, 52–61 (2003).

18. D. L. Bugembe, J. Kayiwa, M. V. T. Phan, P. Tushabe, S. Balinandi, B. Dhaala, J. Lexow, H. Mwebesa, J. Aceng, H. Kyobe, D. Ssemwanga, J. Lutwama, P. Kaleebu, M. Cotten, Main Routes of Entry and Genomic Diversity of SARS-CoV-2, Uganda. *Emerging Infect. Dis.* **26**, 2411–2415 (2020).

19. T. Mashe, F. T. Takawira, L. de Oliveira Martins, M. Gudza-Mugabe, J. Chirenda, M. Munyanyi, B. V. Chaibva, A. Tarupiwa, H. Gumbo, A. Juru, C. Nyagupe, V. Ruhanya, I. Phiri, P. Manangazira, A. Goredema, S. Danda, I. Chabata, J. Jonga, R. Munharira, K. Masunda, SARS-CoV-2 Research Group, Genomic epidemiology and the role of international and regional travel in the SARS-CoV-2 epidemic in Zimbabwe: a

retrospective study of routinely collected surveillance data. *Lancet Glob. Health*. **9**, e1658–e1666 (2021).

20. A. Chouikha, W. Fares, A. Laamari, S. Haddad-Boubaker, Z. Belaiba, K. Ghedira, W. Kammoun Rebai, K. Ayouni, M. Khedhiri, S. Ben Halima, H. Krichen, H. Touzi, I. Ben Dhifallah, F. Z. Guerfali, C. Atri, S. Azouz, O. Khamessi, M. Ardhaoui, M. Safer, N. Ben Alaya, H. Triki, Molecular Epidemiology of SARS-CoV-2 in Tunisia (North Africa) through Several Successive Waves of COVID-19. *Viruses*. **14** (2022), doi:10.3390/v14030624.

21. F. Ntoumi, C. C. Mfoutou Mapanguy, A. Tomazatos, S. R. Pallerla, L. T. K. Linh, N. Casadei, A. Angelov, M. Sonnabend, S. Peter, P. G. Kremsner, T. P. Velavan, Genomic surveillance of SARS-CoV-2 in the Republic of Congo. *Int. J. Infect. Dis.* **105**, 735–738 (2021).

22. Y. Butera, E. Mukantwari, M. Artesi, J. D'Arc Umuringa, Á. N. O'Toole, V. Hill, S. Rooke, S. L. Hong, S. Dellicour, O. Majyambere, S. Bontems, B. Boujemla, J. Quick, P. C. Resende, N. Loman, E. Umumararungu, A. Kabanda, M. M. Murindahabi, P. Tuyisenge, M. Gashegu, N. Rujeni, Genomic Sequencing of SARS-CoV-2 in Rwanda: evolution and regional dynamics. *medRxiv* (2021), doi:10.1101/2021.04.02.21254839.

23. C. N Agoti, G. Githinji, K. S Mohammed, A. W Lambisia, Z. R de Laurent, M. W Mburu, E. M Ong'era, J. M Morobe, E. Otieno, H. Abdou Azali, K. Said Abdallah, A. Diarra, A. Ahmed Yahaya, P. Borus, N. Gumede Moeletsi, D. Fred Athanasius, B. Tsofa, P. Bejon, D. James Nokes, L. Isabella Ochola-Oyier, Detection of SARS-CoV-2 variant 501Y.V2 in Comoros Islands in January 2021. *Wellcome Open Res.* **6**, 192 (2021).

24. J. M. Morobe, D. Didon, B. Pool, A. W. Lambisia, T. Makori, K. S. Mohammed, Z. R. de Laurent, L. Ndwiga, M. W. Mburu, E. Moraa, N. Murunga, J. Musyoki, J. Mwacharo, L. Nyamako, D. Riako, P. Ephnatus, F. Gambo, J. Naimani, J. Namulondo, F. A. Dratibi, C. N. Agoti, Genomic Epidemiology of SARS-CoV-2 in Seychelles, 2020-2021. *medRxiv* (2022), doi:10.1101/2022.03.18.22272503.

25. C. Morang'a, J. Ngoi, J. Gyamfi, D. Amuzu, B. Nuertey, P. Soglo, V. Appiah, I. Asante, P. Owusu-Oduro, S. Armoo, D. Adu-Gyasi, N. Amoako, J. Oliver-Commey, M. Owusu, A. Sylverken, E. Fenteng, V. M'cormack, F. Tei-Maya, E. Quansah, R. Ayivor-Djanie, G. Awandare, Tracking genetic diversity of SARS-CoV-2 infections in Ghana after one year of surveillance. *Res. Sq.* (2021), doi:10.21203/rs.3.rs-1088719/v1.

26. C. N. Agoti, L. I. Ochola-Oyier, K. S. Mohammed, A. W. Lambisia, Z. R. de Laurent, J. M. Morobe, M. W. Mburu, D. O. Omuoyo, E. M. Ongera, L. Ndwiga, E. Maitha, B. Kitole, T. Suleiman, M. Mwakinangu, J. Nyambu, J. Otieno, B. Salim, J. Musyoki, N. Murunga, E. Otieno, G. Githinji, Genomic surveillance reveals the spread patterns of SARS-CoV-2 in coastal Kenya during the first two waves. *medRxiv* (2021), doi:10.1101/2021.07.01.21259583.

27. S. P. C. Brand, J. Ojal, R. Aziza, V. Were, E. A. Okiro, I. K. Kombe, C. Mburu, M. Ogero, A. Agweyu, G. M. Warimwe, J. Nyagwange, H. Karanja, J. N. Gitonga, D. Mugo, S. Uyoga, I. M. O. Adetifa, J. A. G. Scott, E. Otieno, N. Murunga, M. Otiende, E. Barasa, COVID-19 transmission dynamics underlying epidemic waves in Kenya. *Science*. **374**, 989–994 (2021).

28. G. Githinji, Z. R. de Laurent, K. S. Mohammed, D. O. Omuoyo, P. M. Macharia, J. M. Morobe, E. Otieno, S. M. Kinyanjui, A. Agweyu, E. Maitha, B. Kitole, T. Suleiman, M. Mwakinangu, J. Nyambu, J. Otieno, B. Salim, K. Kasera, J. Kiiru, R. Aman, E. Barasa, C. N. Agoti, Tracking the introduction and spread of SARS-CoV-2 in coastal Kenya. *Nat. Commun.* **12**, 4809 (2021).

29. H. Tegally, E. Wilkinson, R. J. Lessells, J. Giandhari, S. Pillay, N. Msomi, K. Mlisana, J. N. Bhiman, A. von Gottberg, S. Walaza, V. Fonseca, M. Allam, A. Ismail, A. J. Glass, S. Engelbrecht, G. Van Zyl, W. Preiser, C. Williamson, F. Petruccione, A. Sigal, T. de Oliveira, Sixteen novel lineages of SARS-CoV-2 in South Africa. *Nat. Med.* **27**, 440–446 (2021).

30. W. H. Roshdy, M. K. Khalifa, J. E. San, H. Tegally, E. Wilkinson, S. Showky, D. P. Martin, M. Moir, A. Naguib, N. Elguindy, M. R. Gomaa, M. Fahim, H. A. Elsood, A. A. Mohsen, R. Galal, M. Hassany, R. J. Lessells, A. A. Al Karmalawy, R. EL Shesheny, A. M. Kandeil, T. de Oliveira, SARS-CoV-2 Genetic diversity and lineage dynamics of in Egypt. *medRxiv* (2022), doi:10.1101/2022.01.05.22268646.

31. C. Scheepers, J. Everatt, D. G. Amoako, A. Mnguni, A. Ismail, B. Mahlangu, C. K. Wibmer, E. Wilkinson, H. Tegally, J. E. San, J. Giandhari, N. Ntuli, S. Pillay, T. Mohale, Y. Naidoo, Z. Khumalo, Z. Makatini, Network for Genomic Surveillance South Africa (NGS-SA), A. Sigal, C. Williamson, J. N. Bhiman, The continuous evolution of SARS-CoV-2 in South Africa: a new lineage with rapid accumulation of mutations of concern and global detection. *medRxiv* (2021), doi:10.1101/2021.08.20.21262342.

32. G. Dudas, S. L. Hong, B. I. Potter, S. Calvignac-Spencer, F. S. Niatou-Singa, T. B. Tombolomako, T. Fuh-Neba, U. Vickos, M. Ulrich, F. H. Leendertz, K. Khan, C. Huber, A. Watts, I. Olendraitė, J. Snijder, K. N. Wijnant, A. M. J. J. Bonvin, P. Martres, S. Behillil, A. Ayouba, G. Baele, Emergence and spread of SARS-CoV-2 lineage B.1.620 with variant of concern-like mutations and deletions. *Nat. Commun.* **12**, 5769 (2021).

33. D. L. Bugembe, M. V. T. Phan, I. Ssewanyana, P. Semanda, H. Nansumba, B. Dhaala, S. Nabadda, Á. N. O'Toole, A. Rambaut, P. Kaleebu, M. Cotten, Emergence and spread of a SARS-CoV-2 lineage A variant (A.23.1) with altered spike protein in Uganda. *Nat. Microbiol.* **6**, 1094–1101 (2021).

34. H. Tegally, M. Moir, J. Everatt, M. Giovanetti, C. Scheepers, E. Wilkinson, K. Subramoney, Z. Makatini, S. Moyo, D. G. Amoako, C. Baxter, C. L. Althaus, U. J. Anyaneji, D. Kekana, R. Viana, J. Giandhari, R. J. Lessells, T. Maponga, D. Maruapula,

W. Choga, T. de Oliveira, Emergence of SARS-CoV-2 Omicron lineages BA.4 and BA.5 in South Africa. *Nat. Med.* (2022), doi:10.1038/s41591-022-01911-2.

35. S. A. Madhi, G. Kwatra, J. E. Myers, W. Jassat, N. Dhar, C. K. Mukendi, A. J. Nana, L. Blumberg, R. Welch, N. Ngorima-Mabhena, P. C. Mutevedzi, Population Immunity and Covid-19 Severity with Omicron Variant in South Africa. *N. Engl. J. Med.* **386**, 1314–1326 (2022).

36. N. Wolter, W. Jassat, S. Walaza, R. Welch, H. Moultrie, M. Groome, D. G. Amoako, J. Everatt, J. N. Bhiman, C. Scheepers, N. Tebeila, N. Chiwandire, M. du Plessis, N. Govender, A. Ismail, A. Glass, K. Mlisana, W. Stevens, F. K. Treurnicht, Z. Makatini, C. Cohen, Early assessment of the clinical severity of the SARS-CoV-2 omicron variant in South Africa: a data linkage study. *Lancet*. **399**, 437–446 (2022).

37. H. C. Lewis, H. Ware, M. Whelan, L. Subissi, Z. Li, X. Ma, A. Nardone, M. Valenciano, B. Cheng, K. Noel, C. Cao, M. Yanes-Lane, B. Herring, A. Talisuna, N. Nsenga, T. Balde, D. A. Clifton, M. Van Kerkhove, D. L. Buckeridge, N. Bobrovitz, the UNITY Studies Collaborator Group, SARS-CoV-2 infection in Africa: A systematic review and meta-analysis of standardised seroprevalence studies, from January 2020 to December 2021. *medRxiv* (2022), doi:10.1101/2022.02.14.22270934.

38. COVID-19 Cumulative Infection Collaborators, Estimating global, regional, and national daily and cumulative infections with SARS-CoV-2 through Nov 14, 2021: a statistical analysis. *Lancet*. **399**, 2351–2380 (2022).

39. J. Quick, nCoV-2019 sequencing protocol v3 (LoCost) (2020).

40. J. R. Tyson, P. James, D. Stoddart, N. Sparks, A. Wickenhagen, G. Hall, J. H. Choi, H. Lapointe, K. Kamelian, A. D. Smith, N. Prystajecky, I. Goodfellow, S. J. Wilson, R. Harrigan, T. P. Snutch, N. J. Loman, J. Quick, Improvements to the ARTIC multiplex PCR method for SARS-CoV-2 genome sequencing using nanopore. *BioRxiv* (2020), doi:10.1101/2020.09.04.283077.

41. M. Cotten, D. Lule Bugembe, P. Kaleebu, M. V T Phan, Alternate primers for whole-genome SARS-CoV-2 sequencing. *Virus Evol.* **7**, veab006 (2021).

42. H. Tegally, M. Ramuth, D. Amoaka, C. Scheepers, E. Wilkinson, M. Giovanetti, R. J. Lessells, J. Giandhari, A. Ismail, D. Martin, E. J. San, M. Crawford, R. S. Daniels, R. Harvey, S. Bahadoor, J. Sonoo, M. Timol, L. Veerapa-Mangroo, A. von Gottberg, J. Bhiman, S. Manraj, A Novel and Expanding SARS-CoV-2 Variant, B.1.1.318, dominates infections in Mauritius. *medRxiv* (2021), doi:10.1101/2021.06.16.21259017.

43. A.-R. N. Zekri, A. A. Bahnasy, M. M. Hafez, Z. K. Hassan, O. S. Ahmed, H. K. Soliman, E. R. El-Sisi, M. H. S. E. Dine, M. S. Solimane, L. S. A. Latife, M. G. Seadawy, A. S. Elsafty, M. Abouelhoda, Characterization of the SARS-CoV-2 genomes in Egypt in first and second waves of infection. *Sci. Rep.* **11**, 21632 (2021).

44. C. Nasimiyu, D. Matoke-Muhia, G. K. Rono, E. Osoro, D. O. Obado, J. M. Mwangi, N. Mwikwabe, K. Thiongâ O, J. Dawa, I. Ngere, J. Gachohi, S. Kariuki, E. Amukoye, M. Mureithi, P. Ngere, P. Amoth, I. Were, L. Makayotto, V. Nene, E. O. Abworo, S. O. Oyola, Imported SARS-COV-2 Variants of Concern Drove Spread of Infections Across Kenya During the Second Year of the Pandemic. *medRxiv* (2022), doi:10.1101/2022.02.28.22271467.

45. M. U. G. Kraemer, V. Hill, C. Ruis, S. Dellicour, S. Bajaj, J. T. McCrone, G. Baele, K. V. Parag, A. L. Battle, B. Gutierrez, B. Jackson, R. Colquhoun, Á. O'Toole, B. Klein, A. Vespignani, COVID-19 Genomics UK (COG-UK) Consortium, E. Volz, N. R. Faria, D. M. Aanensen, N. J. Loman, O. G. Pybus, Spatiotemporal invasion dynamics of SARS-CoV-2 lineage B.1.1.7 emergence. *Science*. **373**, 889–895 (2021).

46. S. J. Lycett, J. Hughes, M. P. McHugh, A. D. S. Filipe, R. Dewar, L. Lu, T. Doherty, A. Shepherd, R. Inward, G. Rossi, D. Balaz, R. R. Kao, S. Rooke, S. Cotton, M. D. Gallagher, C. E. Balcazar-Lopez, A. O'Toole, E. Scher, V. Hill, J. T. McCrone, D. L. Robertson, Epidemic waves of COVID-19 in Scotland: a genomic perspective on the impact of the introduction and relaxation of lockdown on SARS-CoV-2. *medRxiv* (2021), doi:10.1101/2021.01.08.20248677.

47. P. W. G. Mallon, F. Crispie, G. Gonzalez, W. Tinago, A. A. Garcia Leon, M. McCabe, E. de Barra, O. Yousif, J. S. Lambert, C. J. Walsh, J. G. Kenny, E. Feeney, M. Carr, P. Doran, P. D. Cotter, Whole-genome sequencing of SARS-CoV-2 in the Republic of Ireland during waves 1 and 2 of the pandemic. *medRxiv* (2021), doi:10.1101/2021.02.09.21251402.

48. H. Tegally, E. Wilkinson, C. L. Althaus, M. Giovanetti, J. E. San, J. Giandhari, S. Pillay, Y. Naidoo, U. Ramphal, N. Msomi, K. Mlisana, D. G. Amoako, J. Everatt, T. Mohale, A. Nguni, B. Mahlangu, N. Ntuli, Z. T. Khumalo, Z. Makatini, N. Wolter, T. de Oliveira, Rapid replacement of the Beta variant by the Delta variant in South Africa. *medRxiv* (2021), doi:10.1101/2021.09.23.21264018.

49. S. Chang, E. Pierson, P. W. Koh, J. Gerardin, B. Redbird, D. Grusky, J. Leskovec, Mobility network models of COVID-19 explain inequities and inform reopening. *Nature*. **589**, 82–87 (2021).

50. M. Chinazzi, J. T. Davis, M. Ajelli, C. Gioannini, M. Litvinova, S. Merler, A. Pastore Y Piontti, K. Mu, L. Rossi, K. Sun, C. Viboud, X. Xiong, H. Yu, M. E. Halloran, I. M. Longini, A. Vespignani, The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak. *Science*. **368**, 395–400 (2020).

51. M. U. G. Kraemer, C.-H. Yang, B. Gutierrez, C.-H. Wu, B. Klein, D. M. Pigott, Open COVID-19 Data Working Group, L. du Plessis, N. R. Faria, R. Li, W. P. Hanage, J. S. Brownstein, M. Layan, A. Vespignani, H. Tian, C. Dye, O. G. Pybus, S. V. Scarpino, The effect of human mobility and control measures on the COVID-19 epidemic in China.

Science. **368**, 493–497 (2020).

52. P. Nouvellet, S. Bhatia, A. Cori, K. E. C. Ainslie, M. Baguelin, S. Bhatt, A. Boonyasiri, N. F. Brazeau, L. Cattarino, L. V. Cooper, H. Coupland, Z. M. Cucunuba, G. Cuomo-Dannenburg, A. Dighe, B. A. Djaafara, I. Dorigatti, O. D. Eales, S. L. van Elsland, F. F. Nascimento, R. G. FitzJohn, C. A. Donnelly, Reduction in mobility and COVID-19 transmission. *Nat. Commun.* **12**, 1090 (2021).

53. C. Xiong, S. Hu, M. Yang, W. Luo, L. Zhang, Mobile device data reveal the dynamics in a positive relationship between human mobility and COVID-19 infections. *Proc Natl Acad Sci USA*. **117**, 27087–27089 (2020).

54. S. Pillay, J. Giandhari, H. Tegally, E. Wilkinson, B. Chimukangara, R. Lessells, Y. Moosa, S. Mattison, I. Gazy, M. Fish, L. Singh, K. S. Khanyile, J. E. San, V. Fonseca, M. Giovanetti, L. C. Alcantara, T. de Oliveira, Whole Genome Sequencing of SARS-CoV-2: Adapting Illumina Protocols for Quick and Accurate Outbreak Investigation during a Pandemic. *Genes (Basel)*. **11** (2020), doi:10.3390/genes11080949.

55. L. Singh, J. E. San, H. Tegally, P. M. Brzoska, U. J. Anyaneji, E. Wilkinson, L. Clark, J. Giandhari, S. Pillay, R. J. Lessells, D. P. Martin, M. Furtado, A. M. Kiran, T. de Oliveira, Targeted Sanger sequencing to recover key mutations in SARS-CoV-2 variant genome assemblies produced by next-generation sequencing. *Microb. Genom.* **8** (2022), doi:10.1099/mgen.0.000774.

56. A. J. Page, A. E. Mather, T. Le-Viet, E. J. Meader, N.-F. Alikhan, G. L. Kay, L. de Oliveira Martins, A. Aydin, D. J. Baker, A. J. Trotter, S. Rudder, A. P. Tedim, A. Kolyva, R. Stanley, M. Yasir, M. Diaz, W. Potter, C. Stuart, L. Meadows, A. Bell, The Covid-Genomics Uk Cog-Uk Consortium, Large-scale sequencing of SARS-CoV-2 genomes from one region allows detailed epidemiology and enables local outbreak management. *Microb. Genom.* **7** (2021), doi:10.1099/mgen.0.000589.

57. Issues with SARS-CoV-2 sequencing data - nCoV-2019 Genomic Epidemiology - Virological, (available at https://virological.org/t/issues-with-sars-cov-2-sequencing-data/473).

58. N. E. Freed, M. Vlková, M. B. Faisal, O. K. Silander, Rapid and inexpensive whole-genome sequencing of SARS-CoV-2 using 1200 bp tiled amplicons and Oxford Nanopore Rapid Barcoding. *Biol. Methods Protoc.* **5**, bpaa014 (2020).

59. J.-S. Eden, R. Rockett, I. Carter, H. Rahman, J. de Ligt, J. Hadfield, M. Storey, X. Ren, R. Tulloch, K. Basile, J. Wells, R. Byun, N. Gilroy, M. V. O'Sullivan, V. Sintchenko, S. C. Chen, S. Maddocks, T. C. Sorrell, E. C. Holmes, D. E. Dwyer, 2019-nCoV Study Group, An emergent clade of SARS-CoV-2 linked to returned travellers from Iran. *Virus Evol.* **6**, veaa027 (2020).

60. A. S. Gonzalez-Reiche, M. M. Hernandez, M. J. Sullivan, B. Ciferri, H.

Alshammary, A. Obla, S. Fabre, G. Kleiner, J. Polanco, Z. Khan, B. Alburquerque, A. van de Guchte, J. Dutta, N. Francoeur, B. S. Melo, I. Oussenko, G. Deikus, J. Soto, S. H. Sridhar, Y.-C. Wang, H. van Bakel, Introductions and early spread of SARS-CoV-2 in the New York City area. *Science*. **369**, 297–301 (2020).

61. K. Itokawa, T. Sekizuka, M. Hashino, R. Tanaka, M. Kuroda, Disentangling primer interactions improves SARS-CoV-2 genome sequencing by multiplex tiling PCR. *PLoS ONE*. **15**, e0239403 (2020).

62. A. X. Han, A. Toporowski, J. A. Sacks, M. Perkins, S. Briand, M. van Kerkhove, E. Hannay, S. Carmona, B. Rodriguez, E. Parker, B. E. Nichols, C. A. Russell, Low testing rates limit the ability of genomic surveillance programs to monitor SARS-CoV-2 variants: a mathematical modelling study. *medRxiv* (2022), doi:10.1101/2022.05.20.22275319.

63. COVID-19 Excess Mortality Collaborators, Estimating excess mortality due to the COVID-19 pandemic: a systematic analysis of COVID-19-related mortality, 2020-21. *Lancet*. **399**, 1513–1536 (2022).

64. F. Wu, S. Zhao, B. Yu, Y.-M. Chen, W. Wang, Z.-G. Song, Y. Hu, Z.-W. Tao, J.-H. Tian, Y.-Y. Pei, M.-L. Yuan, Y.-L. Zhang, F.-H. Dai, Y. Liu, Q.-M. Wang, J.-J. Zheng, L. Xu, E. C. Holmes, Y.-Z. Zhang, A new coronavirus associated with human respiratory disease in China. *Nature*. **579**, 265–269 (2020).

65. I. Aksamentov, C. Roemer, E. Hodcroft, R. Neher, Nextclade: clade assignment, mutation calling and quality control for viral genomes. *JOSS*. **6**, 3773 (2021).

66. D. P. Martin, A. Varsani, P. Roumagnac, G. Botha, S. Maslamoney, T. Schwab, Z. Kelz, V. Kumar, B. Murrell, RDP5: a computer program for analyzing recombination in, and removing signals of recombination from, nucleotide sequence datasets. *Virus Evol.* **7**, veaa087 (2021).

67. M. N. Price, P. S. Dehal, A. P. Arkin, FastTree 2 — approximately maximum-likelihood trees for large alignments. *PLoS ONE*. **5**, e9490 (2010).

68. J. Felsenstein, Confidence limits on phylogenies: an approach using the bootstrap. *Evolution*. **39**, 783–791 (1985).

69. A. Rambaut, T. T. Lam, L. Max Carvalho, O. G. Pybus, Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus Evol.* **2**, vew007 (2016).

70. A.-A. Popescu, K. T. Huber, E. Paradis, ape 3.0: New tools for distance-based phylogenetics and evolutionary analysis in R. *Bioinformatics*. **28**, 1536–1537 (2012).

71. P. Sagulenko, V. Puller, R. A. Neher, TreeTime: Maximum-likelihood phylodynamic analysis. *Virus Evol.* **4**, vex042 (2018).

72. S. Wang, X. Xu, C. Wei, S. Li, J. Zhao, Y. Zheng, X. Liu, X. Zeng, W. Yuan, S.

Peng, Molecular evolutionary characteristics of SARS-CoV-2 emerging in the United States. *J. Med. Virol.* **94**, 310–317 (2022).

73. G. Yu, Using ggtree to Visualize Data on Tree-Like Structures. *Curr. Protoc. Bioinformatics*. **69**, e96 (2020).

74. H. Wickham, ggplot2. *WIREs Comp Stat*. **3**, 180–185 (2011).

75. J. Hadfield, C. Megill, S. M. Bell, J. Huddleston, B. Potter, C. Callender, P. Sagulenko, T. Bedford, R. A. Neher, Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics*. **34**, 4121–4123 (2018).

Acknowledgments

First and foremost, we acknowledge authors in institutions in Africa and beyond who have made invaluable contributions towards specimen collection and sequencing to produce and share, via GISAID, SARS-CoV-2 genomic data from Africa. A full list of these supplementary authors can be found in the Supplementary Information section. We also acknowledge the authors from the originating and submitting laboratories worldwide, who generated and shared SARS-CoV-2 sequence data, via GISAID, from other regions in the world, which was used to contextualize the African genomic data (**Supp Table S4**).

Funding

Sequencing efforts in the African Union Member States were supported by the Africa Center for Disease Control (Africa CDC) - Africa Pathogen Genomics Initiative (Africa PGI), and the World Health Organization Regional Office for Africa (WHO AFRO) through the transfer of laboratory infrastructure, the provision of reagents and training. The Africa PGI is supported by the African Union, Centers for Disease Control and Prevention (CDC), Bill and Melinda Gates Foundation (BMGF), Illumina Inc, Oxford Nanopore Technologies (ONT) and other partners. In addition, all Institut Pasteur organizations and CERMES in Niger are part of the PEPAIR COVID-19-Africa project which is funded by the French Ministry for European and Foreign Affairs.

KRISP and CERI is supported in part by grants from WHO, the Abbott Pandemic Defense Coalition (APDC), the National Institute of Health USA (U01 AI151698) for the United World

Antivirus Research Network (UWARN) and the INFORM Africa project through IHVN (U54 TW012041), H3BioNet Africa (Grant # 2020 HTH 062), the South African Department of Science and Innovation (SA DSI) and the South African Medical Research Council (SAMRC) under the BRICS JAF #2020/049. ILRI is also supported by the Ministry for Economic Cooperation and Federal Development of Germany (BMZ). Work conducted at ACEGID is made possible by support provided to ACEGID by a cohort of generous donors through TED's Audacious Project, including the ELMA Foundation, MacKenzie Scott, the Skoll Foundation, and Open Philanthropy. Work at ACEGID was also partly supported by grants from the National Institute of Allergy and Infectious Diseases (https://www.niaid.nih.gov), NIH-H3Africa (https://h3africa.org) (U01HG007480 and U54HG007480), the World Bank (projects ACE-019 and ACE-IMPACT), the Rockefeller Foundation (Grant #2021 HTH), the Africa CDC through the African Society of Laboratory Medicine (ASLM; Grant #INV018978), the Wellcome Trust (Project 216619/Z/19/Z) and the Science for Africa Foundation. Sequencing efforts at the National Institute for Communicable Diseases (NICD) was also 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 NICD of the National Health Laboratory Service (NHLS) and the United States Centers for Disease Control and Prevention (FAIN# U01IP001048; NU51IP000930); the South African Medical Research Council (SAMRC, project number 96838); the ASLM and the Bill and Melinda Gates Foundation grant number INV-018978; the UK Foreign, Commonwealth and Development Office and Wellcome (Grant no 221003/Z/20/Z); and the UK Department of Health and Social Care and managed by the Fleming Fund and performed under the auspices of the SEQAFRICA project.

Funding for sequencing efforts in Angola was supported through Projecto Bongola (N.º 11/MESCTI/PDCT/2020) and OGE INIS (2020/2021). Botswana's sequencing efforts led by the Botswana Harvard AIDS Institute Partnership was supported by: Foundation for Innovative New Diagnostics(FINDdx); BMGF, H3ABioNet [U41HG006941], Sub-Saharan African Network for TB/HIV Research Excellence (SANTHE) and Fogarty International Center (Grant # 5D43TW009610) . H3ABioNet is an initiative of the Human Health and Heredity in Africa Consortium (H3Africa) programme of the African Academy of Science (AAS). HHS/NIH/National Institute of Allergy and Infectious Diseases (NIAID) (5K24AI131928-04; 5K24AI131924-04); SANTHE is a DELTAS Africa Initiative [grant # DEL-15-006]. The DELTAS Africa Initiative is an independent funding scheme of the African Academy of Sciences (AAS)'s Alliance for Accelerating Excellence in Science in Africa (AESA) and supported by the New Partnership for Africa's Development Planning and Coordinating Agency (NEPADAgency) with funding from the Wellcome Trust [grant #107752/Z/15/Z] and the UK government. From Brazil, Joicymara Santos Xavier was funded by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brazil (CAPES) - Finance Code 001. Sequencing efforts from Cote d'Ivoire were funded by the Robert Koch Institute and the German Federal Ministry of Education and Research (BMBF). Sequencing efforts in the Democratic Republic of the Congo were funded by the Bill &

Melinda Gates Foundation under grant INV-018030 awarded to CBP and further supported by funding from the Africa CDC through ASLM for Accelerating SARS-CoV-2 Genomic Surveillance in Africa, the US Centre for Disease Control and Prevention (US CDC), USAMRIID, IRD/Montepellier, UCLA and SACIDS FIND. Efforts from Egypt were funded by the Egyptian Ministry of Health, the Egyptian Academy for Scientific Research and Technology (ASRT) JESOR project #3046 (Center for Genome and Microbiome Research), the Cairo University anti COVID-19 fund and the Science and Technology Development Fund (STDF), Project ID: 41907.

The sequencing effort in Equatorial Guinea was supported by a public–private partnership, the Bioko Island Malaria Elimination Project, composed of the government of Equatorial Guinea Ministries of Mines and Hydrocarbons, and Health and Social Welfare, Marathon EG Production Limited, Noble Energy, Atlantic Methanol Production Company, and EG LNG. Analysis for the Gabon strains was supported by the Science and Technology Research Partnership for Sustainable Development (SATREPS), Japan International Cooperation Agency (JICA), and Japan Agency for Medical Research and Development (AMED) (grant number JP21jm0110013) and a grant from AMED (grant number JP21wm0225003). CIRMF (Gabon) is funded by the Gabonese Government and TOTAL Energy inc. CIRMF is a member of CANTAM supported by EDCTP. The work at WACCBIP (Ghana) was funded by a grant from the Rockefeller Foundation (2021 HTH 006), an Institut de Recherche pour le Développement (IRD) grant (ARIACOV), African Research Universities Alliance (ARUA) Vaccine Development Hubs grant with funds from Open Society Foundation, National Institute of Health Research (NIHR) (17.63.91) grants using UK aid from the UK Government for a global health research group for Genomic surveillance of malaria in West Africa (Wellcome Sanger Institute, UK) and the World Bank African Centers of Excellence Impact grant (WACCBIP-NCDs: Awandare).

In addition to the funding sources from ILRI, KEMRI (Kenyan) contributions to sequencing efforts was supported in part by the National Institute for Health Research (NIHR) (project references 17/63/82 and 16/136/33) using UK aid from the UK Government to support global health research, and The UK Foreign, Commonwealth and Development Office (FCDO) and Wellcome (grant# 220985/Z/20/Z) and the Kenya Medical Research Institute Grant # KEMRI/COV/SPE/012. Contributions from Lesotho were supported by the Africa CDC, ALSM and SA NICD. Liberian efforts was funded by the Africa CDC through a subaward from the Bill and Melinda Gates Foundations, while efforts from Madagascar were funded by the French Ministry for Europe and Foreign Affairs through the REPAIR COVID-19-Africa project coordinated by the Pasteur International Network association. Sequencing from Malawi was supported by Wellcome Trust. Contributions from Mali was supported by Fogarty International Center and National Institute of Allergy and Infectious Diseases sections of the National Institutes of Health under Leidos-15X051, award numbers U2RTW010673 for the West African Center of Excellence for Global Health Bioinformatics Research Training and U19AI089696 and U19AI129387 for the West Africa International Center of Excellence for Malaria Research. Funding for surveillance, sampling and

testing in Madagascar: World Health Organization (WHO), the US Centers for Disease Control and Prevention (US CDC: Grant#U5/IP000812-05), the United States Agency for International Development (USAID: Cooperation Agreement 72068719CA00001), the Office of the Assistant Secretary for Preparedness and Response in the U.S. Department of Health and Human Services (DHHS: grant number IDSEP190051-01-0200). Funding for sequencing: Bill & Melinda Gates Foundation (GCE/ID OPP1211841), Chan Zuckerberg Biohub, and the Innovative Genomics Institute at UC Berkeley.

Mozambique acknowledges support from the Mozambican Ministry of Health and the President's Emergency Plan for AIDS Relief (PEPFAR) through the U.S. Centers for Disease Control and Prevention (CDC) under the terms of [grant number GH002021, GH001944], and the Bill & Melinda Gates Foundation, #OPP1214435. Namibian efforts was supported by Africa CDC through a subaward from the Bill and Melinda Gates Foundations. Efforts from the country Niger were supported by the French Ministry for Europe and Foreign Affairs through the REPAIR COVID-19-Africa project coordinated by the Pasteur International Network association. In addition to the funding support for ACEGID already listed, Nigeria's contributions were made possible by support from Flu Lab and a cohort of donors through the Audacious Project, a collaborative funding initiative housed at TED, including the ELMA Foundation, MacKenzie Scott, the Skoll Foundation, and Open Philanthropy.

Efforts from the Republic of the Congo was supported by the European and Developing Countries Clinical Trials Partnership (EDCTP) IDs: PANDORA, CANTAM and German Academic Exchange Service (DAAD) IDs: PACE-UP; DAAD Project ID: 5759234. Rwanda's contributions were made possible by funding from the African Network for improved Diagnostics, Epidemiology and Management of common Infectious Agents (ANDEMIA) was granted by the German Federal Ministry of Education and Research (BMBF grant 01KA1606, 01KA2021 and 01KA2110B) and the National Institute of Health Research (NIHR) Global Health Research programme (16/136/33) using UK aid from the UK Government. In addition to the South African institutions listed above, the University of Cape Town's work was supported by the Wellcome Trust [Grant # 203135/Z/16/Z], EDCTP RADIATES (RIA2020EF-3030), the South African Department of Science and Innovation (SA DSI) and the South African Medical Research Council (SAMRC), Stellenbosch University's contributions by the South African Medical Research Council (SA-MRC), and the University of Pretoria's contributions funded by the G7 Global Health Fund (REF#) and the BMBF ANDEMIA grant (REF#).

Funding from the Fleming Fund supported sequencing in Sudan. The Ministry of Higher Education and Scientific Research of Tunisia provided funding for sequencing from Tunisia. UVRI (Uganda) acknowledge support from the Wellcome Trust and FCDO - Wellcome Epidemic Preparedness – Coronavirus (AFRICO19, grant agreement number 220977/Z/20/Z), from the MRC (MC_UU_1201412) and from the UK Medical Research Council (MRC/UKRI) and FCDO

(DIASEQCO, grant agreement number NC_PC_19060). Research at the FredHutch institute which supported bioinformatics analyses of sequences in the present study was supported by the Bill and Melinda Gates foundation (#INV-018979). Research support from Broad Institute colleagues was made possible by support from Flu lab and a cohort of generous donors through TED's Audacious Project, including the ELMA Foundation, MacKenzie Scott, the Skoll Foundation, Open Philanthropy, the Howard Hughes Medical Institute and NIH (U01AI151812 and U54HG007480) (P.C.S.). Work from Quadram Institute Bioscience was funded by The Biotechnology and Biological Sciences Research Council Institute Strategic Programme Microbes in the Food Chain BB/R012504/1 and its constituent projects BBS/E/F/000PR10348, BBS/E/F/000PR10349, BBS/E/F/000PR10351, and BBS/E/F/000PR10352 and by the Quadram Institute Bioscience BBSRC funded Core Capability Grant (project number BB/CCG1860/1). Sequences generated in Zambia through PATH were funded by the BMGF and Africa CDC.

The content and findings reported herein are the sole deduction, view and responsibility of the researcher/s and do not reflect the official position and sentiments of the funding agencies.

Supplementary Materials

Supplementary Figure S1
Supplementary Figure S2
Supplementary Figure S3
Supplementary Figure S4
Supplementary Figure S5
Supplementary Figure S6
Supplementary Figure S7
Supplementary Figure S8
Supplementary Figure S9
Supplementary Figure S10
Supplementary Figure S11
Supplementary Figure S12
Supplementary Figure S13
Supplementary Figure S14
Supplementary Figure S15
Supplementary Figure S16
Supplementary Table S1
Supplementary Table S2
Supplementary Table S3
Supplementary Table S4

Author contribution

Conceptualization: HT, CB, SKT, TdO, RL, EW

Methodology: HT, JES, MC, BT, GM, DPM, AWL, DAR, LMK, GG, TdO, RL, EW

Genomic Data Generation: HT, JES, MC, MM, BT, GM, DPM, AWL, AD, DGA, MMD, AS, ANZ, ASG, AKS, AO, AS, AOM, AKS, AGA, AL, AK, AEA, AAJ, AF, AOO, AAA, AJ, AK, AM, AR, AS, AK, AB, AC, AJT, AC, AKK, AK, AB, AS, AA, AN , AVG, AN , AJP, AY, AV, ANH, AC, AI, AM, ALB, AI, AAS, AG, AF, AES, BM, BLS, BSO, BB, BD, BLH, BT, BL, BM, BN, BTM, BAK, BK, BA, BP, BM, CB, CW, CN, CA, CBP, CS, CGA, CNA, CMM, CL, CKO, CI, CNM, CP, CG, CEO, CDR, CMM, CE, DBL, DJB, DM, DP, DB, DJN, DS, DT, DSA, DG, DSG, DOO, DM, DWW, EF, EKL, ES, EMO, ENN, EOA, EO, ES, EB, EBA, EAA, EL, EM, EP, EB, ES, EAA, FL, FMT, FW, FA, FTT, FD, FVA, FT, FO, FN, FMM, FER, FAD, FI, GKM, GT, GLK, GOA, GUvZ, GAA, GS, GPM, HCR, HEO, HO, HA, HK, HN, HT, HAAK, HE, HG, HM, HK, IS, IBO, IMA, IO, IBB, IAM, IS, IW, ISK, JWAH, JA, JS, JCM, JMT, JH, JGS, JG, JM, JN, JNU, JNB, JY, JM, JK, JDS, JH, JKO, JMM, JOG, JTK, JCO, JSX, JG, JFW, JHB, JN, JE, JN, JMN, JN, JUO, JCA, JJL, JJHM, JO, KJS, KV, KTA, KAT, KSC, KSM, KD, KGM, KOD, LF, LS, LMK, LB, LdOM, LC, LO, LDO , LLD, LIO, LT, MM, MR, MM, ME, MM, MIM, MK, MD, MM, MdLLM, MV, MFP, MF, MMN, MM, MD, MWM, MGM, MO, MRW, MYT, MOA, MA, MAB, MGS, MKK, MMM, MK, MS, MBM, MM, MA, MVP, NA, NR, NA, NI, NE, NMT, ND, NM, NH, NBS, NMF, NS, NB, NM, NG, NW, NS, NN, NAA, NT, NM, NHR, NI, NM, OCK, OS, OF, OMA, OT, OAO, OF, OEO, O-EO, OF, PS, PO, PC, PN, PS, PEO, PA, PKQ, POO, PB, PD, PAB, PKM, PK, PA, RE, RJ, RKA, RGE, RA, RN, ROP, RG, RAK, RMND, RAA, RAC, SG, SM, SB, SS, SIM, SF, SM, SH, SKK, SM, ST, SHA, SWM, SD, SM, SA, SSA, SMA, SE, SM, SL, SG, SJ, SFA, SO, SG, SL, SP, SO, SvW, SFS, SK, SA, SR, SP, SN, SB, SLB, SvdW, TM, TM, TL, TPV, TS, TGM, TB, UJA, UC, UR, UEG, VE, VN, VG, WHR, WAK, WKA, WP, WTC, YAA, YR, YB, YN, YB, ZRdL, AEO, AvG, GG, MM, OT, PCS, AAS, SOO, YKT, SKT, TdO, CH, RL, JN, EW

Data Analysis: HT, JES, MC, MM, BT, GM, DPM, AWL, AIE, DAR, EM, GSK, SvW, GG, TdO, RL, EW

Funding acquisition: AEO, AvG, GG, MM, OT, AAS, SOO, YKT, SKT, TdO, CH

Project administration: GM, AD, DGA, MMD, AC, DWW, HO, SWM, AEO, AvG, GG, MM, OT, PCS, AAS, SOO, YKT, SKT, TdO, CH, RL, JN, EW

Supervision: AEO, AvG, GG, MM, OT, PCS, AAS, SOO, YKT, SKT, TdO, CH, RL, JN, EW

Writing – original draft: HT, JES, MC, GM, DPM, CB, SKT, TdO, RL, EW

Writing – review & editing: HT, JES, MC, MM, BT, GM, DPM, CB, AWL, AD, DGA, MMD, AS, ANZ, ASG, AKS, AO, AS, AOM, AKS, AIE, AL, AK, AEA, AAJ, AF, AOO, AAA, AJ, AK, AM, AR, AS, AK, AB, AC, AJT, AC, AKK, AK, AB, AS, AA, AVG, AJP, AY, AV, ANH, AC, AI, AM, ALB, AI, AAS, AG, AF, AES, BM, BLS, BSO, BB, BD, BLH, BT, BL, BM, BN, BTM, BAK, BK, BA, BP, BM, CB, CW, CA, CBP, CS, CGA, CNA, CMM, CL, CKO, CI, CNM, CP, CEO, CDR, CMM, CE, DBL, DJB, DM, DP, DB, DJN, DS, DT, DSA, DG, DSG, DOO, DM, DWW, EF, EKL, ES, EMO, ENN, EOA, EO, ES, EB, EBA, EL, EM, EP, EB, ES, EAA, EM, FL, FMT, FW, FA, FTT, FD, FVA, FT, FO, FN, FMM, FER, FAD, FI, GKM, GT, GLK, GOA, GUvZ, GAA, GSK, GS, GPM, HCR, HEO, HO, HA, HK, HN, HT, HAAK, HE, HG, HM, HK, IS, IBO, IMA, IO, IBB, IS, IW, ISK, JWAH, JA, JS, JCM, JMT, JH, JGS, JG, JM, JNU, JNB, JY, JM, JK, JDS, JH, JKO, JMM, JOG, JTK, JCO, JSX, JG, JHB, JN, JE, JN, JMN, JN, JUO, JCA, JJL, JO, KJS, KV, KTA, KAT, KSC, KSM, KD, KGM, KOD, LF, LS, LB, LdOM, LC, LO, LLD, LIO, MM, MR, MM, ME, MM, MIM, MK, MD, MM, MdLLM, MV, MFP, MF, MMN, MM, MD, MWM, MGM, MO, MRW, MYT, MOA, MA, MAB, MGS, MKK, MMM, MK, MS, MBM, MM, MVP, NA, NR, NI, NMT, ND, NM, NH, NBS, NMF, NS, NB, NM, NG, NW, NS, NN, NAA, NT, NM, NHR, NI, NM, OCK, OS, OF, OMA, OT, OAO, OF, OEO, OF, PS, PO, PC, PN, PS, PEO, PA, PKQ, POO, PB, PD, PAB, PKM, PK, PA, RE, RJ, RKA, RGE, RA, RN, ROP, RG, RAK, RAA, RAC, SG, SM, SS, SIM, SF, SM, SH, SKK, SM, ST, SHA, SWM, SD, SM, SA, SSA, SMA, SE, SM, SL, SG, SJ, SFA, SO, SG, SL, SP, SO, SvW, SFS, SK, SA, SR, SP, SN, SB, SLB, SvdW, TM, TM, TL, TPV, TS, TGM, TB, UJA, UC, UR, UEG, VE, VN, VG, WHR, WAK, WKA, WP, WTC, YAA, YR, YB, YN, YB, ZRdL, AEO, AvG, GG, MM, OT, PCS, AAS, SOO, YKT, SKT, TdO, CH, RL, JN, EW

Competing interests: With the exception of Pardis Sabeti who is a co-founder of and consultant to Sherlock Biosciences and a Board Member of Danaher Corporation and who holds equity in the companies, we the authors have no conflicts of interest to declare.

Data and Code availability

All of the SARS-CoV-2 whole genome sequences that were analyzed in the present study are all publicly available on the GISAID sequence database. A full list of the African sequences as well as global references are presented and acknowledged in **Supp Table S4** and in our github repository (https://github.com/CERI-KRISP/SARS-CoV-2-epidemic-in-Africa) and under DOI: https://zenodo.org/badge/latestdoi/513935576 at Zenodo. The github repositories also contain all of the metadata, raw and time scaled ML tree topologies, annotated tree topologies as well as data analysis and visualization scripts used here, which will allow for the independent reproduction of results. Furthermore, the repositories also contain all Institutional Review Board (IRB) and Material Transfer Agreements (MTA). Please refer to the Ethics Statement in the Methods section for more details.

Chapter 8: Conclusion

In this thesis, real-time genomic monitoring of SARS-CoV-2 is showcased, utilising national and continental efforts towards regular genomic surveillance programs during the pandemic, with the application of various bioinformatics, phylogenetics, and epidemiological mapping techniques. Research work towards this thesis enabled the rapid detection and monitoring of multiple variants of concern and lineages in South Africa and Africa, often with global public health impact. The knowledge acquired from the results of Chapter 2 established the SARS-CoV-2 surveillance baselines in South Africa during the first wave of infections, and was ultimately key in rapidly recognizing the evolutionary emergence of the Beta variant of concern after the up-tick of cases in October 2020, described in Chapter 3. Chapter 4 and 5 extends methods developed towards the first two chapters to rapidly investigate a sharp acceleration in cases at the beginning of the fourth and fifth waves in South Africa. This enabled the characterisation and phylodynamics analysis of the Omicron variant of concern and its BA.4 and BA.5 lineages in record time, which also meant that the world in turn received quick warning of an upcoming threat. While the Beta variant only circulated regionally mostly in southern and East Africa, the Omicron lineages BA.1, BA.2 (Chapter 4) and BA.4 and BA.5 (Chapter 5), all went on to consecutively dominate infections in many parts of the world in 2022. Finally, Chapter 6 and 7 present continental genomic surveillance efforts in Africa. In addition to giving insights into the establishment of epidemics from external introductions, and cross-border viral movements, they highlight the expansion of genomic surveillance on the continent to cover blindspots and the benefits this brings for fast public health action. Most importantly, they help to recognize that genomic surveillance for highly transmissible pathogens cannot be done in silo, and must be integrated within strong regional and global collaborative networks for most effective response.

Other arms of this research have also contributed to answering additional key questions during the pandemic as well as to epidemiological and genomic monitoring of pathogens in several other ways. Genome assembly and phylogenetic analysis of sequenced diagnostic specimens helped to investigate the genomic epidemiology of SARS-CoV-2 in many instances. This included descriptions of early transmission dynamics of SARS-CoV-2 in the province of KwaZulu-Natal (1), the city of Cape Town in South Africa (2), and the province of the Free States (3), genomic monitoring of infection waves and outbreaks in Brazil (4,5), Mauritius (6), Uganda (7), Malawi (8) and Egypt (9), description of additional lineages of interest in South Africa, including the rapid replacement of Beta with Delta (10) and the C.1.2 lineage (11), and tracking the spread of variants internationally in real-time (12).

Bioinformatics analyses contributed to the development and fine tuning of diagnostics and sequencing methods during the pandemic, including setting up whole genome sequencing protocols (13,14), assessing or adapting qPCR methods to highly mutated variants (15,16), and development of Sanger sequencing gap filling methods to solve regions that become difficult to

sequence with whole genome sequencing due to primer dropouts (17). Genomic analysis methods such as lineage classification and phylogenetics were also integrated in studies of vaccine efficacy (18), antibody neutralisation from naturally acquired or vaccine immunity (19,20), cross neutralisation from infection with other variants (21), and T-cell recognition (22). This was crucial to translating SARS-CoV-2 variant genotypes to phenotypes in actionable ways related to diagnostics and therapeutics. Genomic data generation and bioinformatics analysis also contributed to more thorough investigation of the adaptive evolution and selection processes occurring within SARS-CoV-2 genomes (23,24).

Clinically, sequence assembly and bioinformatics methods were also applied to characterising the intriguing phenomenon of chronic and persistent SARS-CoV-2 infections in immunocompromised patients (25,26). For instance, research from this thesis helped to characterise patients infected with the virus for more than 6 and 9 months respectively, both patients with non-suppressed HIV and both showing advanced evolution of the infecting virus over this length of time (25,26). This is of critical interest to the research and public health communities, given plausible hypotheses that variants of concern could originally evolve during this type of persistent infections. Interestingly, every time these patients have their HIV antiviral treatments reviewed and restored and their HIV viral load goes down, they also soon after test negative for SARS-CoV-2. This points to actionable measures to take charge of HIV patients in Southern Africa to decrease their infection times and improve outcomes from SARS-CoV-2 infections.

Ongoing projects from this research include further work on integrative analysis of genomic, epidemiological and mobility data, particularly a brief research article on how changes in human mobility during the COVID-19 pandemic influenced the synchrony of epidemic peaks across the world (In press: Journal of Travel Medicine), and an in-depth analysis of the global dispersal patterns of SARS-CoV-2 variants of concern and the associated role of international travel volumes (In preparation). Finally, research output from this thesis has extended beyond SARS-CoV-2, having contributed to visualisation for the early tracking of Monkeypox cases during the 2022 outbreak (27), and to an international lobby by viral epidemiologists towards non-discriminatory nomenclature for the monkeypox virus (28).

Methods developed and skills acquired during this thesis will be used to further expand research outcomes in pathogen genomics and epidemic modelling. In future studies, a number of research questions will be tackled. Still around understanding the epidemiology of SARS-CoV-2, a thorough investigation of the co-evolution of SARS-CoV-2 and other pathogens, such as HIV, will be undertaken, which will feed into a systematic analysis to understand the origins of the Omicron variant of concern. Further research will focus on epidemic forecasting and disease susceptibility modelling to inform rapid response and surveillance strategies, especially related to factors that predict expansion of localised outbreaks to large epidemics, and how climate change is shifting epidemic risks for arboviral disease establishment in Africa.

References

1. Giandhari J, Pillay S, Wilkinson E, Tegally H, Sinayskiy I, Schuld M, et al. Early transmission of SARS-CoV-2 in South Africa: An epidemiological and phylogenetic report. Int J Infect Dis. 2021 Feb;103:234–41.

2. Engelbrecht S, Delaney K, Kleinhans B, Wilkinson E, Tegally H, Stander T, et al. Multiple Early Introductions of SARS-CoV-2 to Cape Town, South Africa. Viruses. 2021 Mar 22;13(3).

3. Mwangi P, Okendo J, Mogotsi M, Ogunbayo A, Adelabu O, Sondlane H, et al. SARS-CoV-2 variants from COVID-19 positive cases in the Free State province, South Africa from July 2020 to December 2021. FrontVirol. 2022 Sep 14;2.

4. Giovanetti M, Slavov SN, Fonseca V, Wilkinson E, Tegally H, Patané JSL, et al. Genomic epidemiology of the SARS-CoV-2 epidemic in Brazil. Nat Microbiol. 2022 Sep;7(9):1490–500.

5. Giovanetti M, Fonseca V, Wilkinson E, Tegally H, San EJ, Althaus CL, et al. Replacement of the Gamma by the Delta variant in Brazil: Impact of lineage displacement on the ongoing pandemic. Virus Evol. 2022 Mar 18;8(1):veac024.

6. Tegally H, Ramuth M, Amoaka D, Scheepers C, Wilkinson E, Giovanetti M, et al. A Novel and Expanding SARS-CoV-2 Variant, B.1.1.318, dominates infections in Mauritius. medRxiv. 2021 Jun 16;

7. Bbosa N, Ssemwanga D, Namagembe H, Kiiza R, Kiconco J, Kayiwa J, et al. Rapid Replacement of SARS-CoV-2 Variants by Delta and Subsequent Arrival of Omicron, Uganda, 2021. Emerging Infect Dis. 2022 May;28(5):1021–5.

8. Bandawe G, Chigwechokha P, Kunyenje P, Kazembe Y, Mwale J, Kamaliza M, et al. Management and containment of a SARS-CoV-2 Beta variant outbreak at the Malawi University of Science and Technology. medRxiv. 2021 Nov 28;

9. Roshdy WH, Khalifa MK, San JE, Tegally H, Wilkinson E, Showky S, et al. SARS-CoV-2 Genetic Diversity and Lineage Dynamics in Egypt during the First 18 Months of the Pandemic. Viruses. 2022 Aug 25;14(9).

10. Tegally H, Wilkinson E, Althaus CL, Giovanetti M, San JE, Giandhari J, et al. Rapid replacement of the Beta variant by the Delta variant in South Africa. medRxiv. 2021 Sep 27;

11. Scheepers C, Everatt J, Amoako DG, Tegally H, Wibmer CK, Mnguni A, et al. Emergence and phenotypic characterization of the global SARS-CoV-2 C.1.2 lineage. Nat

Commun. 2022 Apr 8;13(1):1976.

12. O'Toole Á, Hill V, Pybus OG, Watts A, Bogoch II, Khan K, et al. Tracking the international spread of SARS-CoV-2 lineages B.1.1.7 and B.1.351/501Y-V2 with grinch. Wellcome Open Res. 2021 Sep 17;6:121.

13. Pillay S, Giandhari J, Tegally H, Wilkinson E, Chimukangara B, Lessells R, et al. Whole Genome Sequencing of SARS-CoV-2: Adapting Illumina Protocols for Quick and Accurate Outbreak Investigation during a Pandemic. Genes (Basel). 2020 Aug 17;11(8).

14. Tshiabuila D, Giandhari J, Pillay S, Ramphal U, Ramphal Y, Maharaj A, et al. Comparison of SARS-CoV-2 sequencing using the ONT GridION and the Illumina MiSeq. BMC Genomics. 2022 Apr 22;23(1):319.

15. Valley-Omar Z, Marais G, Iranzadeh A, Naidoo M, Korsman S, Maponga T, et al. Reduced amplification efficiency of the RNA-dependent-RNA-polymerase target enables tracking of the Delta SARS-CoV-2 variant using routine diagnostic tests. J Virol Methods. 2022 Apr;302:114471.

16. Vogels CBF, Breban MI, Ott IM, Alpert T, Petrone ME, Watkins AE, et al. Multiplex qPCR discriminates variants of concern to enhance global surveillance of SARS-CoV-2. PLoS Biol. 2021 May 7;19(5):e3001236.

17. Singh L, San JE, Tegally H, Brzoska PM, Anyaneji UJ, Wilkinson E, et al. Targeted Sanger sequencing to recover key mutations in SARS-CoV-2 variant genome assemblies produced by next-generation sequencing. Microb Genom. 2022 Mar;8(3).

18. Madhi SA, Baillie V, Cutland CL, Voysey M, Koen AL, Fairlie L, et al. Efficacy of the ChAdOx1 nCoV-19 Covid-19 Vaccine against the B.1.351 Variant. N Engl J Med. 2021 May 20;384(20):1885–98.

19. Cele S, Gazy I, Jackson L, Hwa S-H, Tegally H, Lustig G, et al. Escape of SARS-CoV-2 501Y.V2 from neutralization by convalescent plasma. Nature. 2021 May;593(7857):142–6.

20. Cele S, Jackson L, Khoury DS, Khan K, Moyo-Gwete T, Tegally H, et al. SARS-CoV-2 Omicron has extensive but incomplete escape of Pfizer BNT162b2 elicited neutralization and requires ACE2 for infection. medRxiv. 2021 Dec 17;

21. Khan K, Karim F, Ganga Y, Bernstein M, Jule Z, Reedoy K, et al. Omicron sub-lineages BA.4/BA.5 escape BA.1 infection elicited neutralizing immunity. medRxiv. 2022 May 1;

22. Riou C, Keeton R, Moyo-Gwete T, Hermanus T, Kgagudi P, Baguma R, et al. Escape from recognition of SARS-CoV-2 variant spike epitopes but overall preservation of T cell immunity. Sci Transl Med. 2022 Feb 9;14(631):eabj6824.

23. Martin DP, Weaver S, Tegally H, San JE, Shank SD, Wilkinson E, et al. The

emergence and ongoing convergent evolution of the SARS-CoV-2 N501Y lineages. Cell. 2021 Sep 30;184(20):5189-5200.e7.

24. Martin DP, Lytras S, Lucaci AG, Maier W, Gruning B, Shank SD, et al. Selection analysis identifies significant mutational changes in Omicron that are likely to influence both antibody neutralization and Spike function (part 1 of 2) [Internet]. 2021 [cited 2021 Dec 11]. Available from: https://virological.org/t/selection-analysis-identifies-significant-mutational-changes-in-omicron-that-are-likely-to-influence-both-antibody-neutralization-and-spike-function-part-1-of-2/771

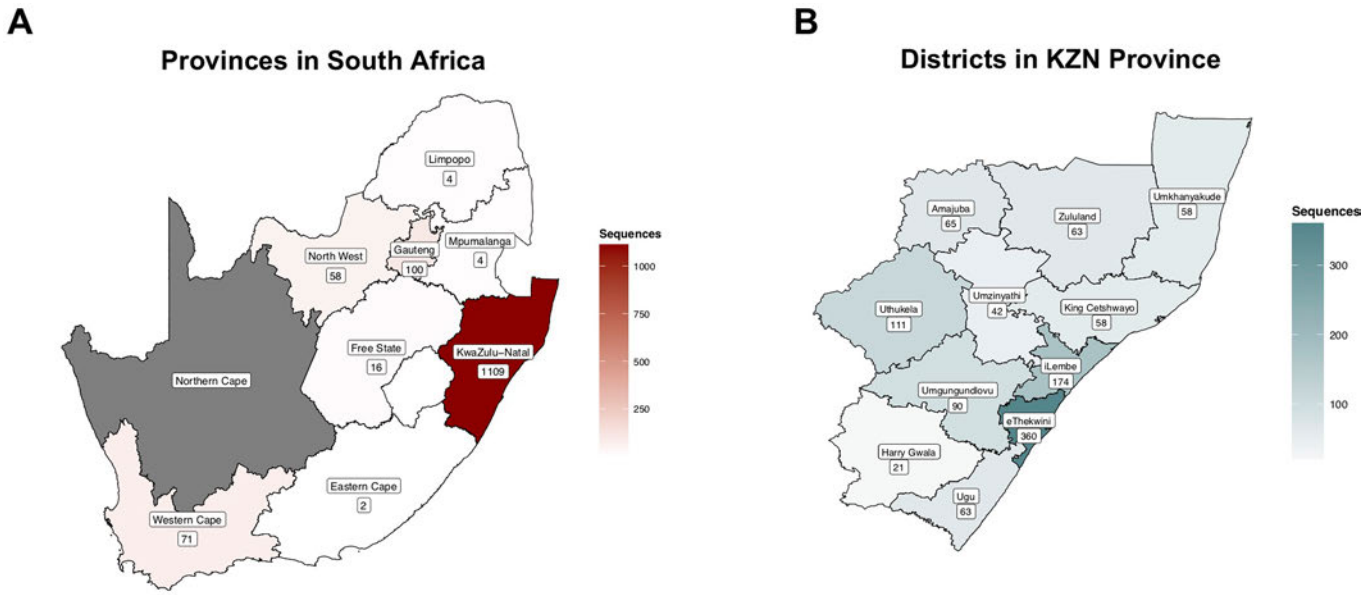
25. Karim F, Moosa MYS, Gosnell BI, Cele S, Giandhari J, Pillay S, et al. Persistent SARS-CoV-2 infection and intra-host evolution in association with advanced HIV infection. medRxiv. 2021 Jun 4;

26. Maponga TG, Jeffries M, Tegally H, Sutherland A, Wilkinson E, Lessells RJ, et al. Persistent SARS-CoV-2 infection with accumulation of mutations in a patient with poorly controlled HIV infection. Clin Infect Dis. 2022 Jul 6;

27. Kraemer MUG, Tegally H, Pigott DM, Dasgupta A, Sheldon J, Wilkinson E, et al. Tracking the 2022 monkeypox outbreak with epidemiological data in real-time. Lancet Infect Dis. 2022 Jul;22(7):941–2.

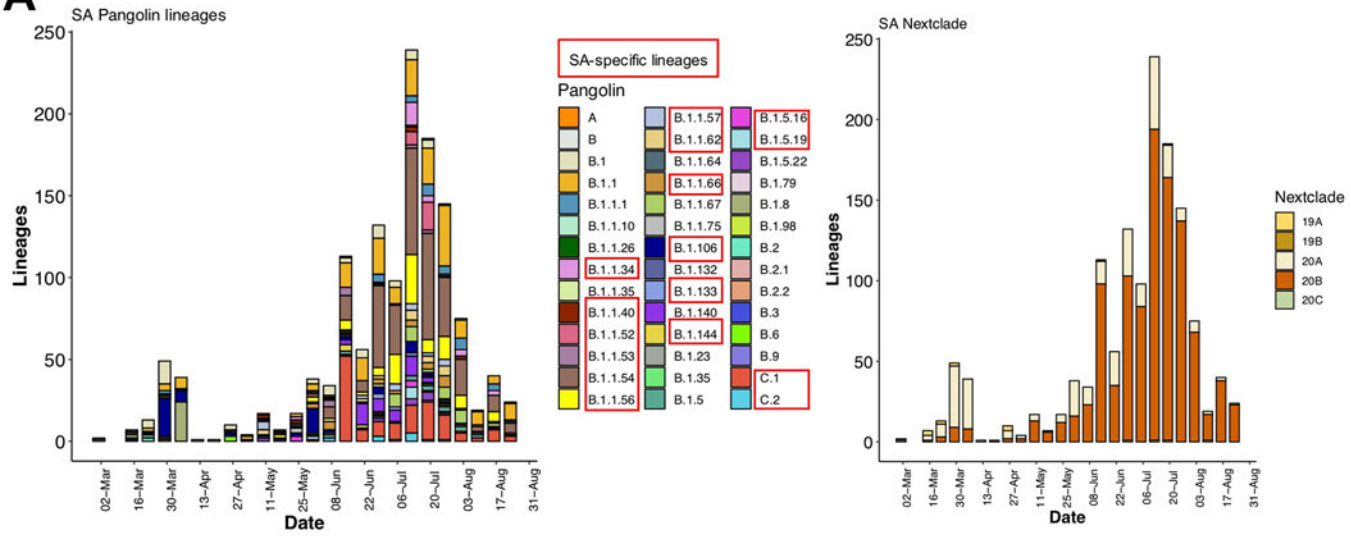
28. Happi C, Adetifa I, Mbala P, Njouom R, Nakoune E, Happi A, et al. Urgent need for a non-discriminatory and non-stigmatizing nomenclature for monkeypox virus. PLoS Biol. 2022 Aug 23;20(8):e3001769.

Appendices



Extended Data Fig. 1 | Map density representation of where the genomes in this study were sampled. a, The number of genomes sampled in each province in South Africa (no genomes from Northern Cape – grey), **b**, The number of genomes sampled in each district of KZN, the most sampled province.

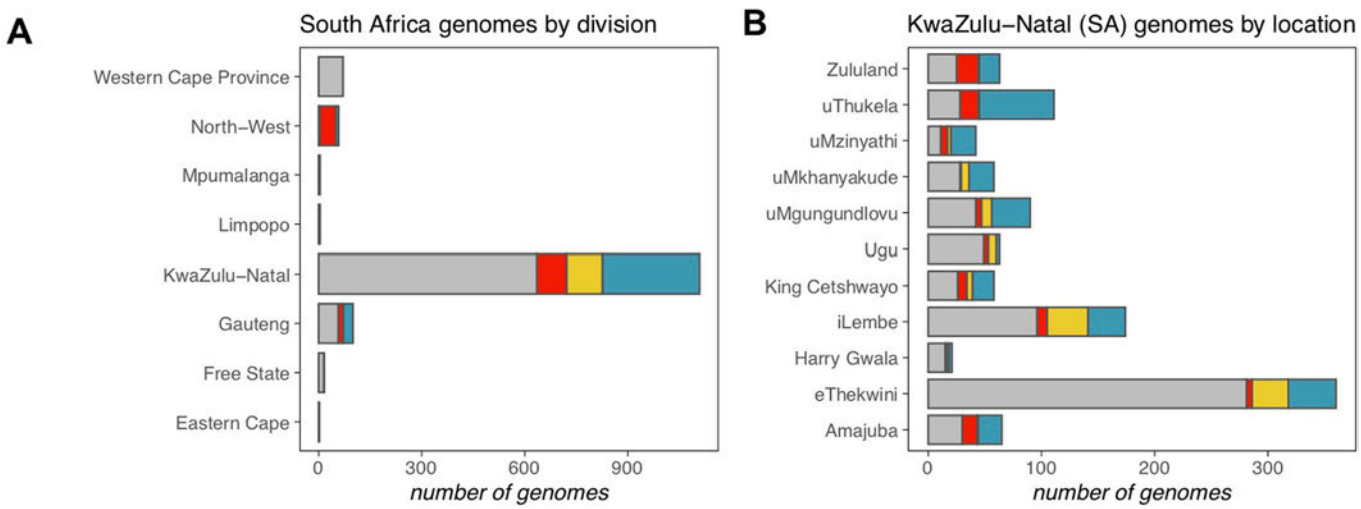
A



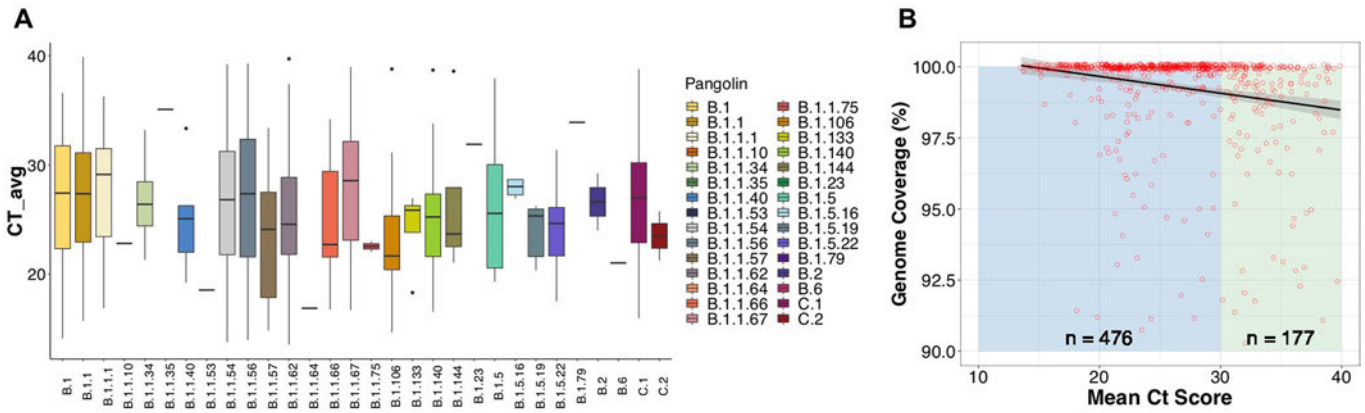
B

	B.1.106	B.1.1.54	C.1	B.1.1.56
Number of Genomes in our dataset	68	320	151	104
Most common countries	ZA (56%), USA (39%), Spain (3%)	ZA (99%), UK (1%)	ZA (99%), UK (1%)	ZA (99%), Australia (1%)
Most Common Provinces	KwaZulu-Natal	North West, Gauteng, KwaZulu-Natal	North West, Gauteng, Limpopo, Free State, KwaZulu-Natal	KwaZulu-Natal
KZN districts	5 districts	All 11 districts	All 11 districts	All 11 districts
Date range	March 16 to July 21	March 19 to August 26	June 03 to August 26	March 21 to August 21
Dates since last sampling in ZA	2020-08-21	2020-08-26	2020-08-21	2020-08-26
Recall value	0.84	0.98	0.99	0.95

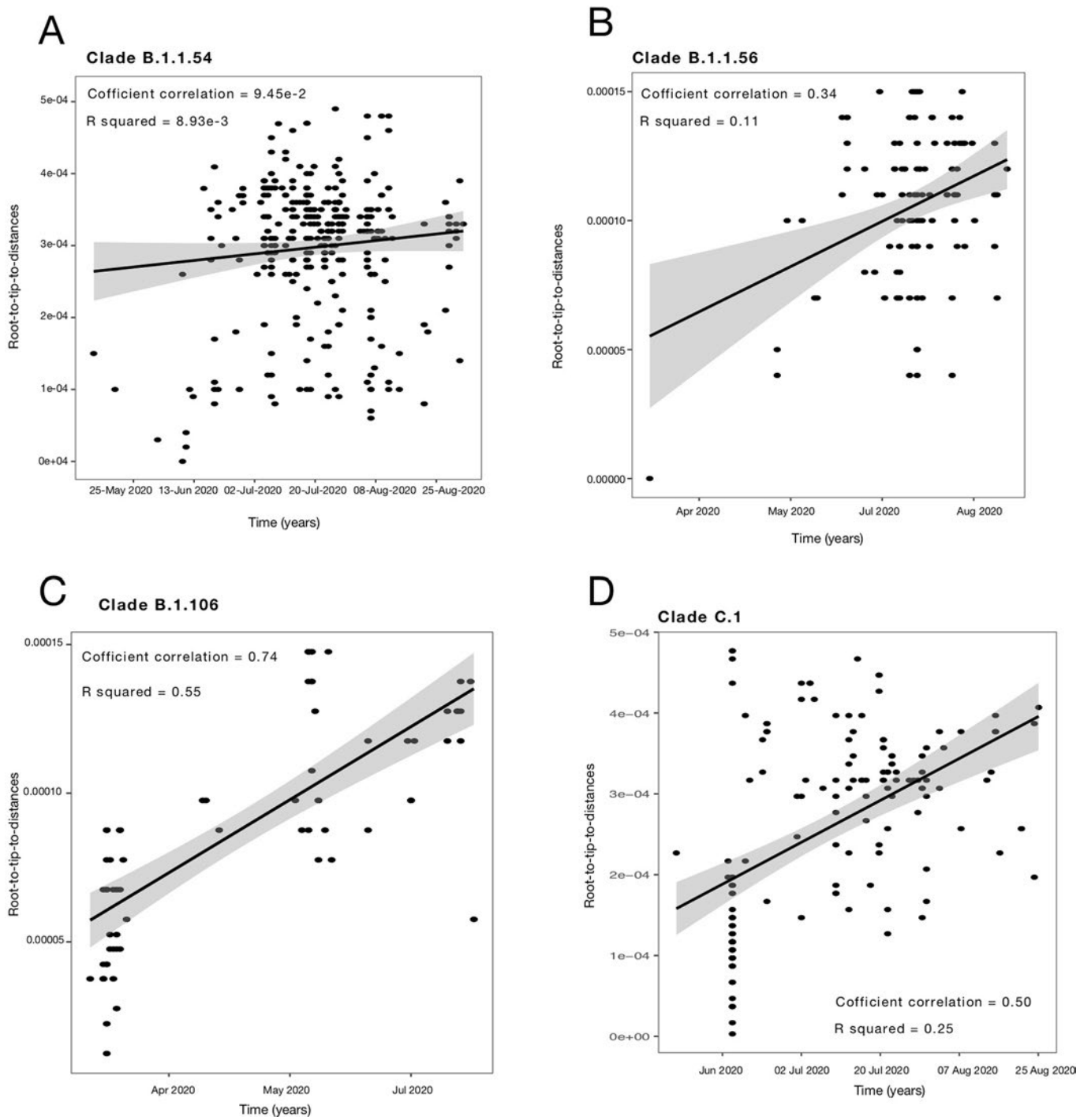
Extended Data Fig. 2 | Classification of viruses circulating in South Africa. a, Classification of South Africa genomes (n = 1365) per date into Pangolin lineages (SA-specific ones specified by red boxes), and into Nextstrain clades. **b**, Detailed sampling information for the four lineages cluster identified to be almost unique to South Africa.



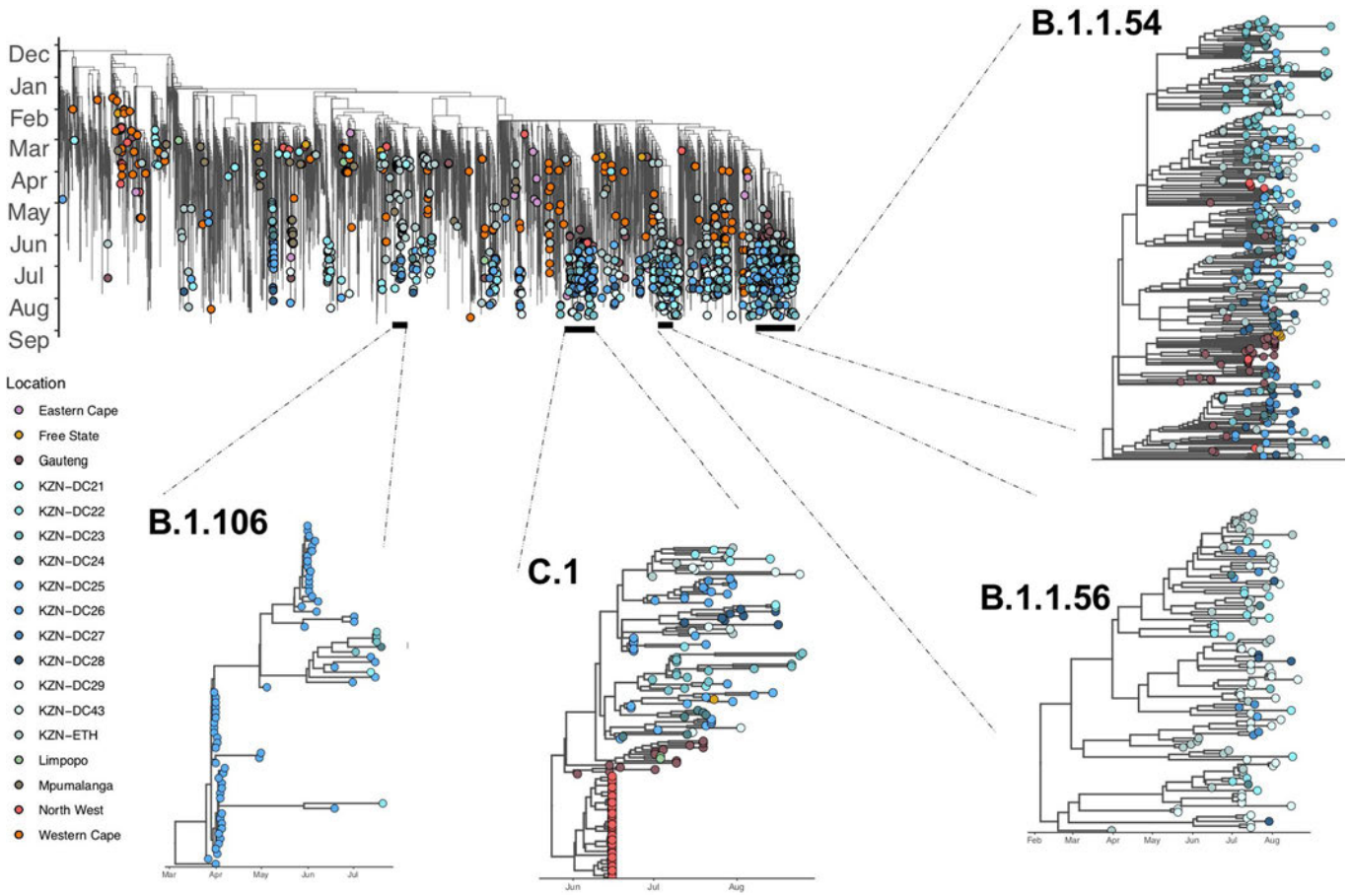
Extended Data Fig. 3 | Prevalence of main lineage clusters in South Africa. a, Distribution of genomes belonging to the lineage clusters by province. **b**, Distribution of genomes belonging to the lineage clusters by district of KZN.



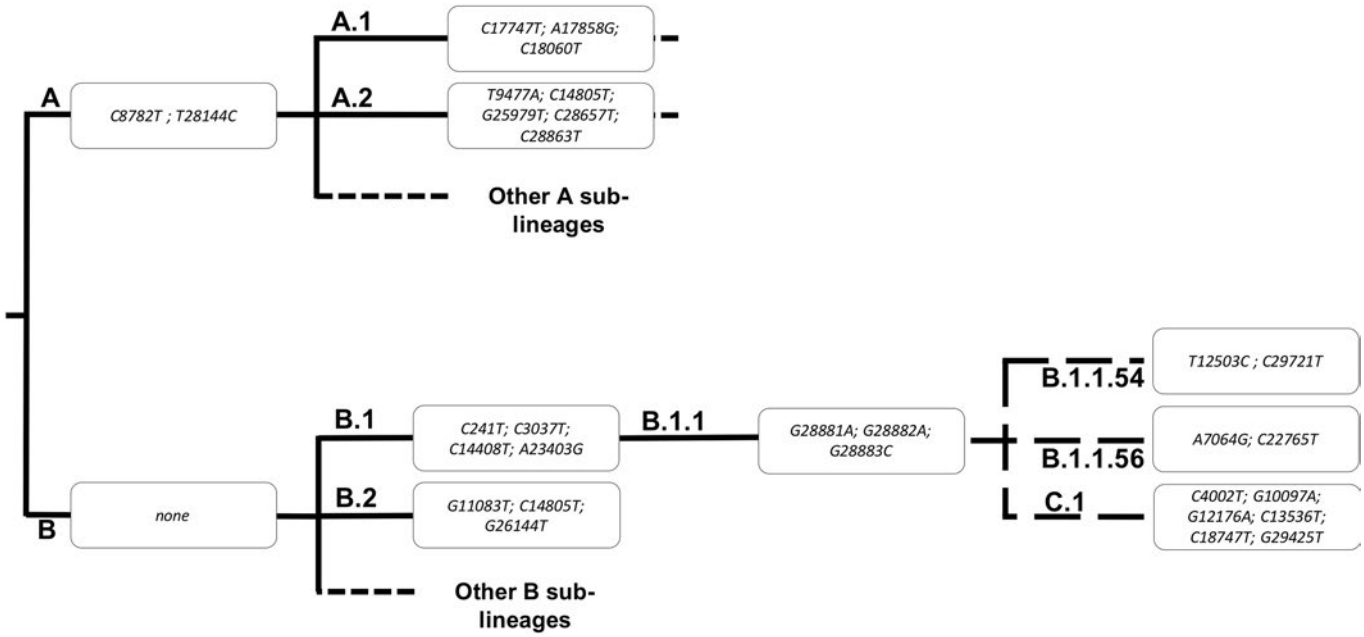
Extended Data Fig. 4 | Ct score investigation of samples in this study. a, Showing the average Ct scores at three target genes for genomes generated at KRISP, and classified into their respective Pangolin lineages. Each box is delimited by two lines at the 25th percentile and 75th percentile, with the line inside the box represents the median, and whisker lines drawn from the box to the whisker boundaries. **b**, Showing the decreasing trend in genome coverage as Ct score increases, with the bulk of genomes > 90% (n = 476) falling in the Ct < 30 category.



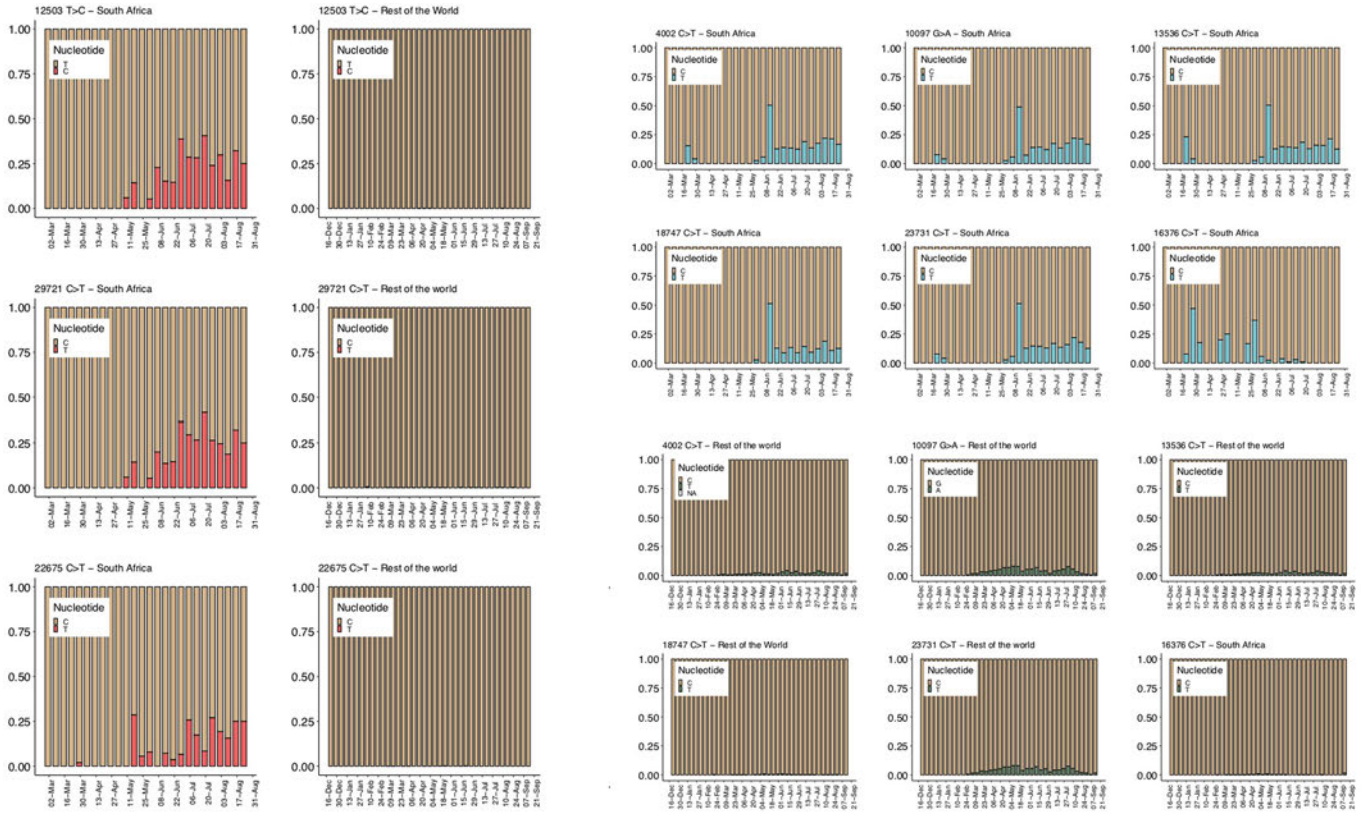
Extended Data Fig. 5 | Temporal signaling for each cluster (Tempest). For SARS-CoV-2, we accept temporal signaling with correlation coefficient > 0.2. Cluster B.1.1.54 (A) had a low correlation coefficient and was therefore rejected from further Bayesian spatiotemporal analyses. Regression lines are shown with error buffers (shaded area) representing 90% confidence intervals.



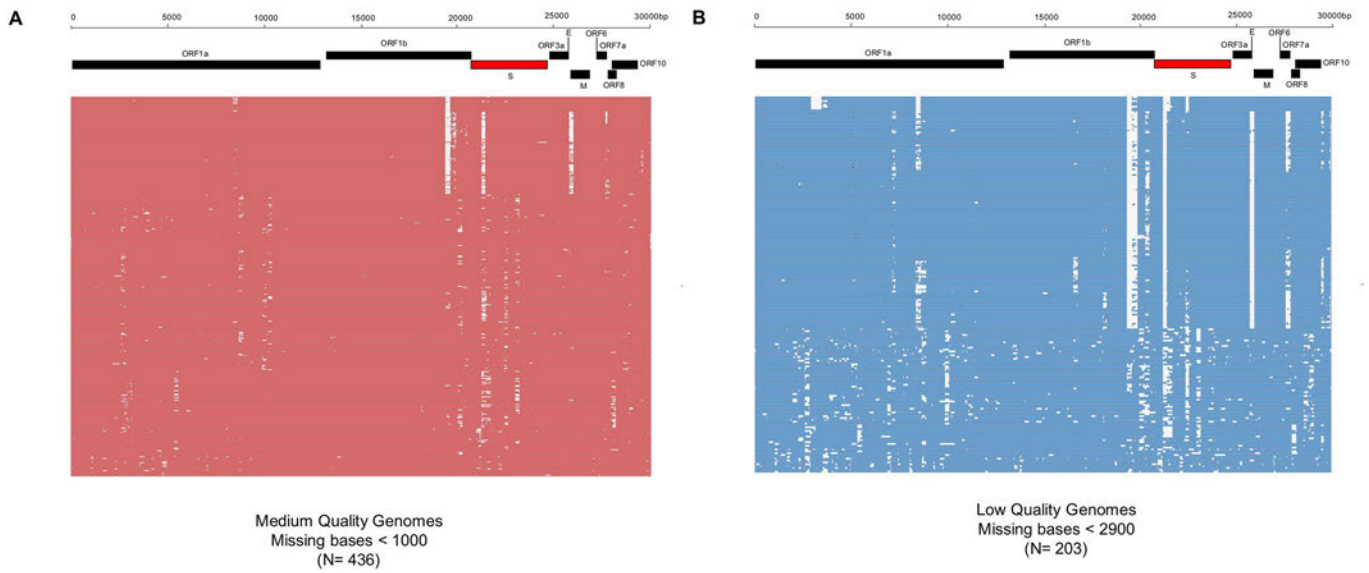
Extended Data Fig. 6 | Maximum likelihood tree of a global dataset showing genomes coloured by sampling location in South Africa. For genomes sampled in KZN, they are further specified by which district they were sampled from. A closer look into cluster B.1.106, C.1 and B.1.1.56 illustrated as trees from BEAST temporal analyses, with a defined time-scale. The zoom-in tree for B.1.1.54 was extracted as a subset of the big ML tree.



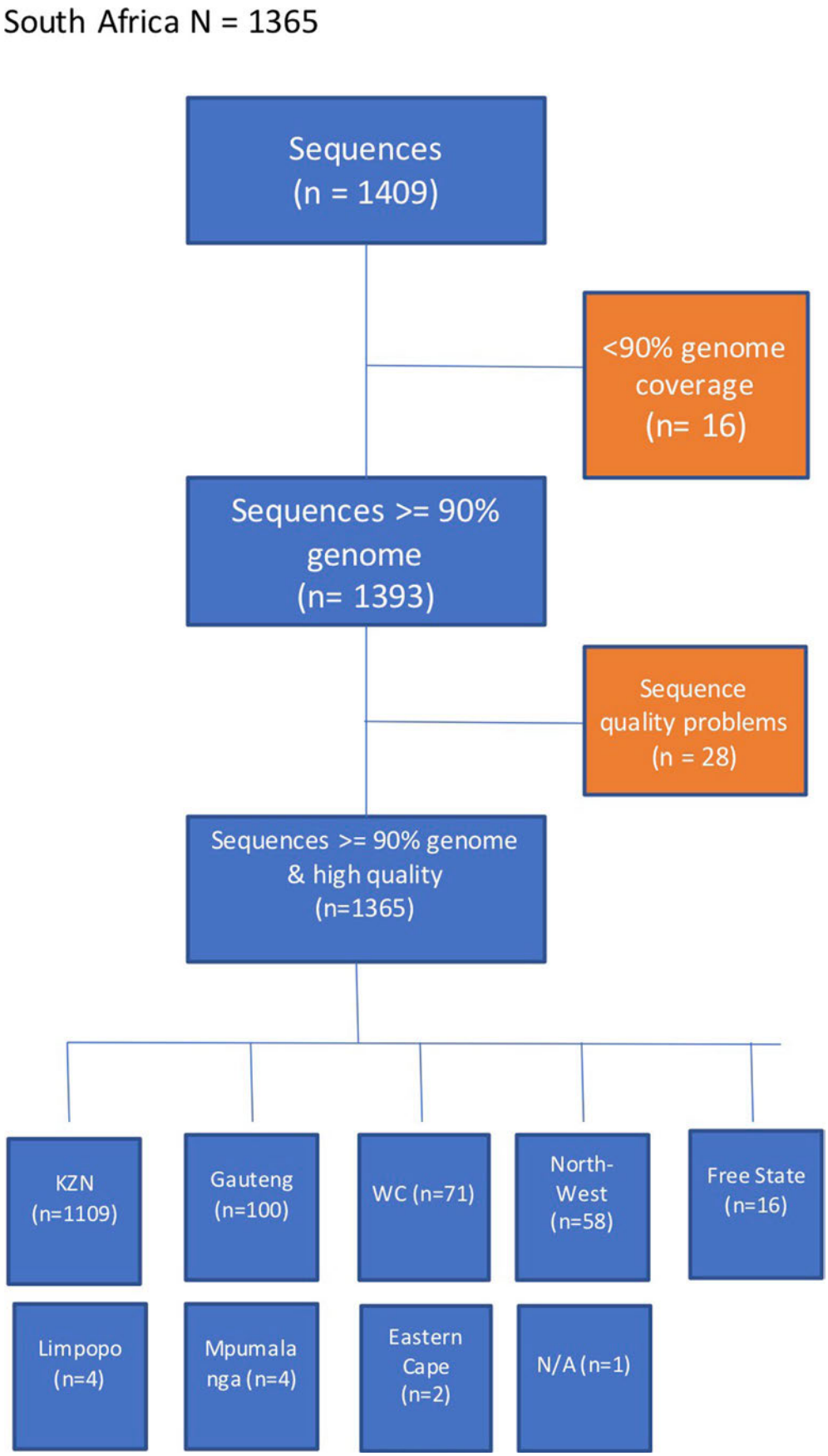
Extended Data Fig. 7 | Flowchart of Pangolin lineage A and B dividing into sub-lineages with their lineage-defining mutations specified. The main lineage-defining mutations of the main lineage cluster investigated in this study are shown in relation to their parent lineage (B.1.1).



Extended Data Fig. 8 | Mutation frequencies in SA vs rest of the world for lineage-defining mutations. Mutation predominantly seen in South Africa are shown in red, whereas the others are shown in blue.

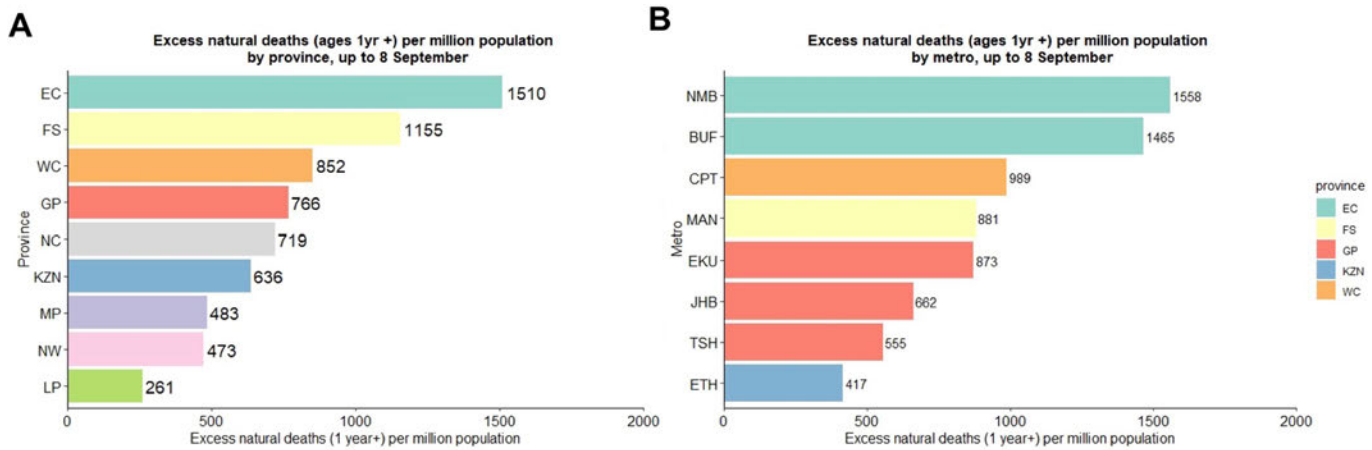


Extended Data Fig. 9 | Mapping of SARS-CoV-2 genome alignment generated in this study onto the genome structure to determine the location of gaps in the sequences. a, An alignment of 436 medium quality genomes (<1000 missing bases) showing small amounts of recurrent gaps (white spaces) in ORF1b, S, ORF3a. **b**, An alignment of 203 low quality genomes (<2900 missing bases) showing a more important amounts of recurrent gaps (white) in ORF1a, ORF1b, S, ORF3a, and ORF7a genes. The rest of our genomes (N = 726) had 100% coverage relative to the reference.

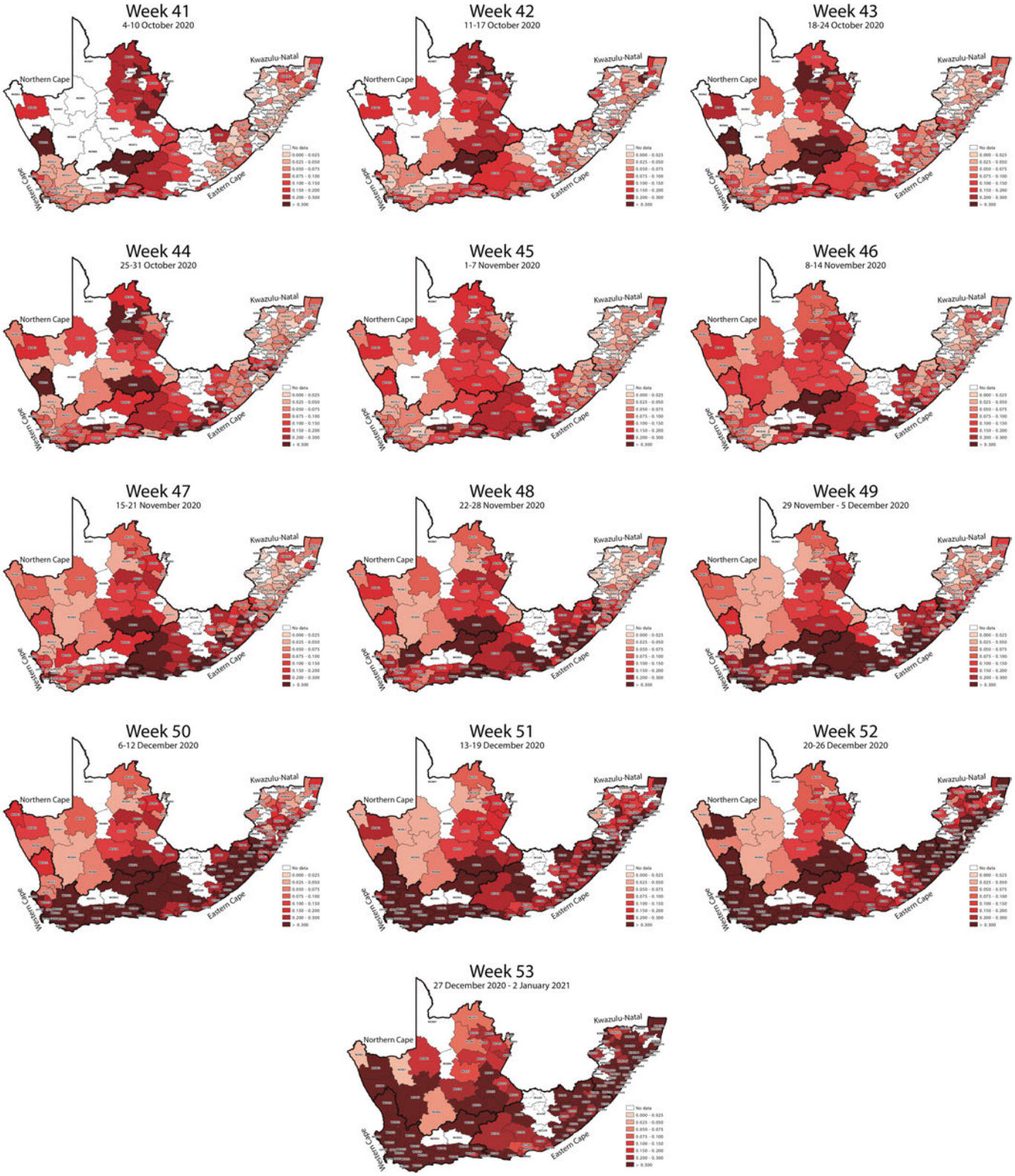


Extended Data Fig. 10 | See next page for caption.

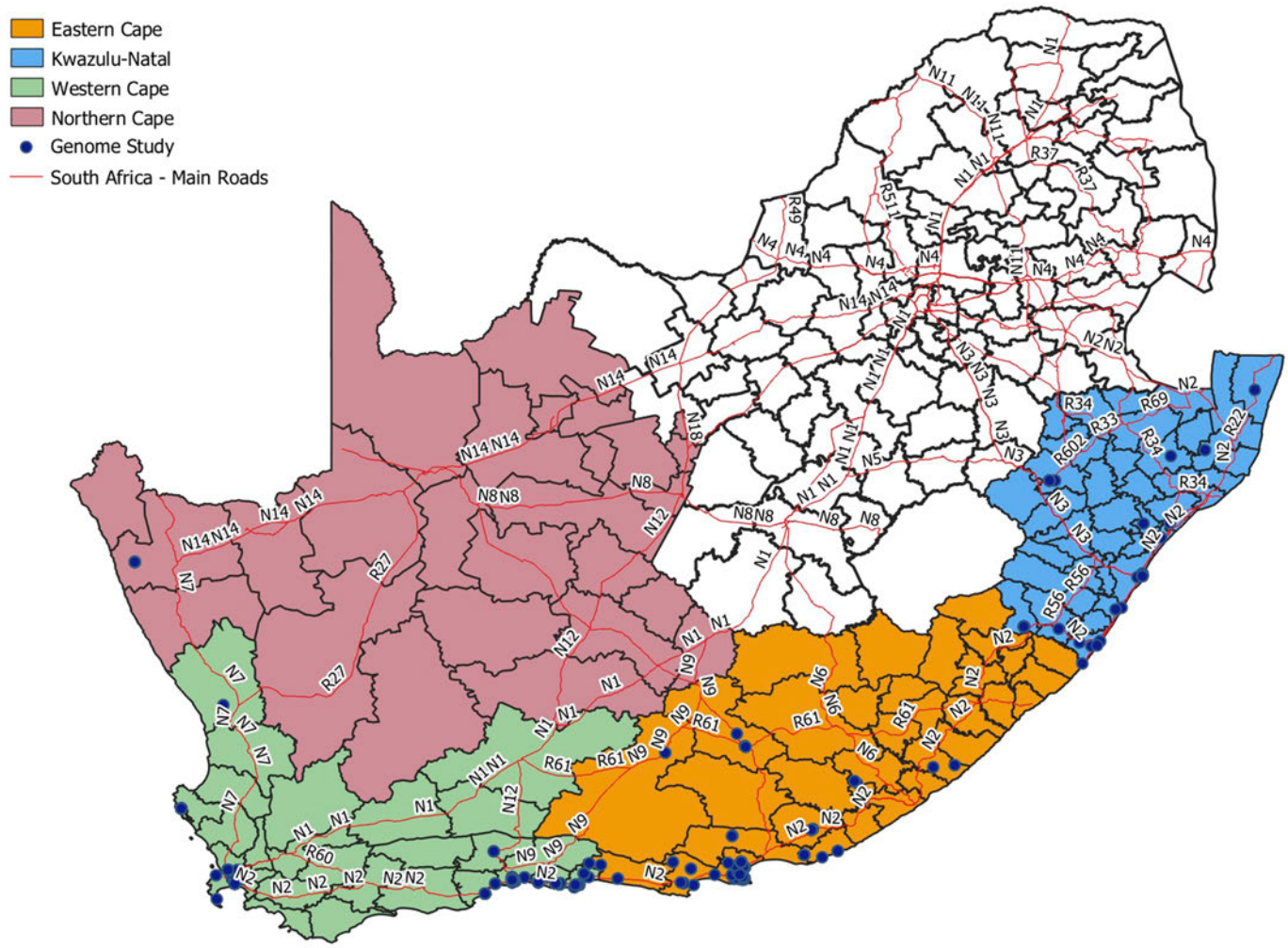
Extended Data Fig. 10 | Curation of South Africa dataset from all available South African genomes available on GISAID as at 15th September 2020. We show initial number of genomes (n = 1409), how many were excluded at each cleaning step and the final number of genomes (n = 1365) with subdivisions into their originating province.



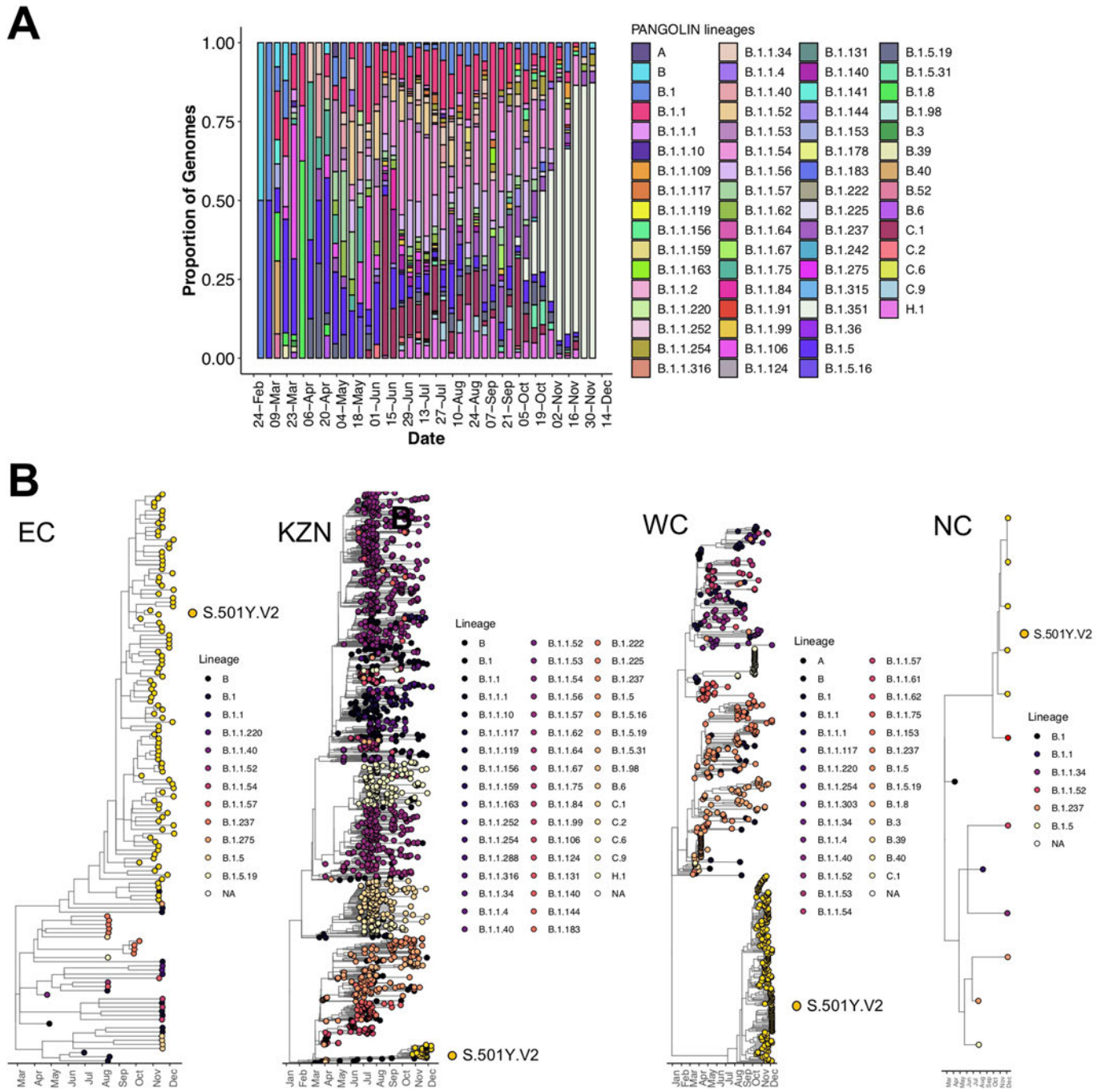
Extended Data Fig. 1 | Excess deaths per million individuals by province and metropolitan municipalities of South Africa. Data are shown for up until the week ending 8 September 2020 (immediately after the first peak of the epidemic peak). **a**, **b**, These graphs indicate the disproportionate effect of the first wave of the epidemic in the province of the Eastern Cape (**a**) and its metropolitan areas (Nelson Mandela Bay and Buffalo City) (**b**). EC, Eastern Cape; FS, Free State; WC, Western Cape; GP, Gauteng province; NC, Northern Cape; KZN, KwaZulu–Natal; MP, Mpumalanga; NW, North West; NMB, Nelson Mandela Bay; BUF, Buffalo City; CPT, Cape Town; MAN, Mangaung; EKU, Ekurhuleni; JHB, Johannesburg; TSH, Tshwane; ETH, Ethekwini.



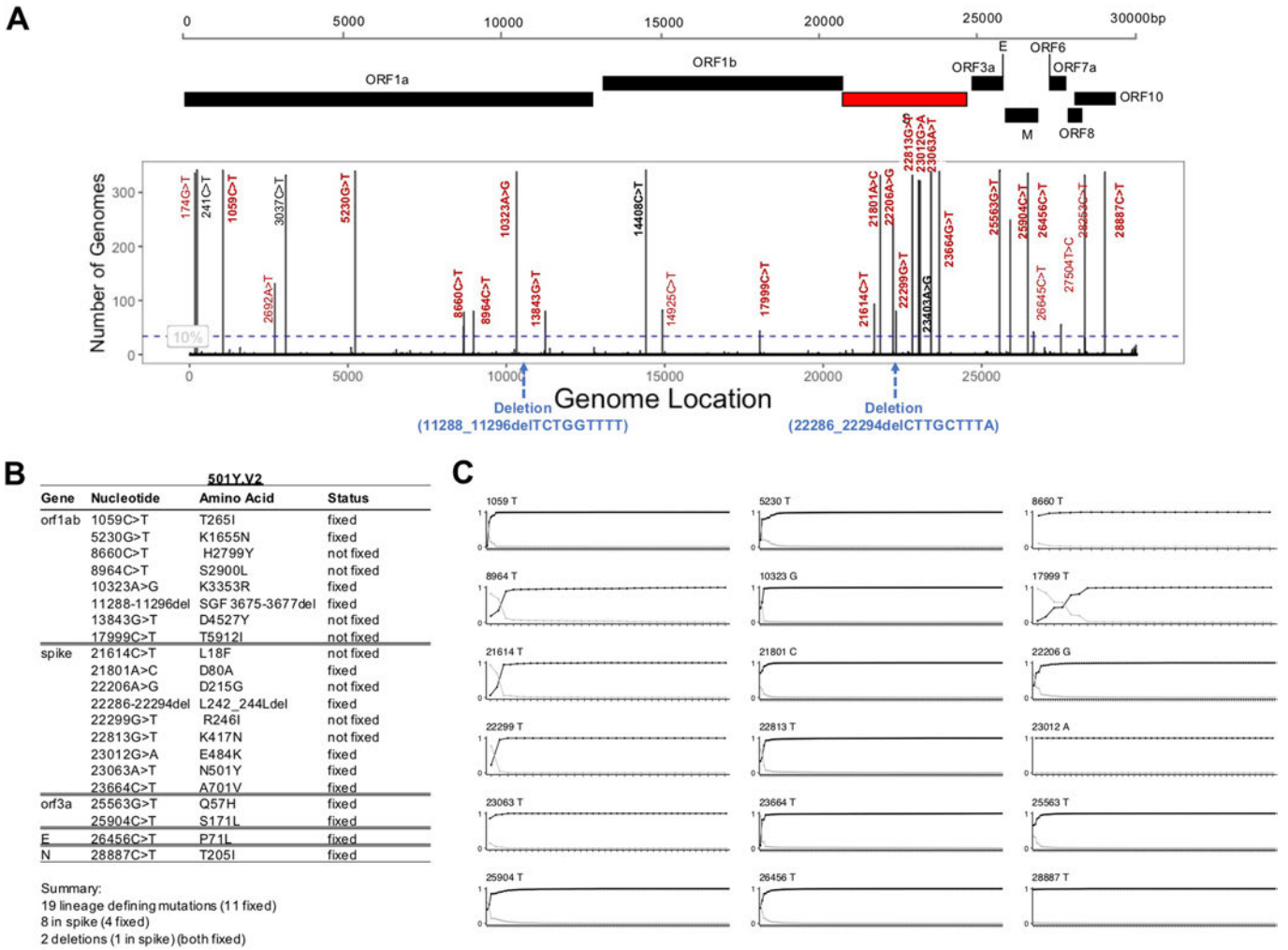
Extended Data Fig. 2 | Positivity test rates across four provinces of South Africa. Maps of the Northern Cape, Western Cape, the Eastern Cape and KwaZulu–Natal (the four provinces investigated in this Article) showing a weekly progression of SARS-CoV-2 prevalence per district, coloured by the rate of positive SARS-CoV-2 PCR tests per district. Data were obtained from the weekly testing report of the National Institute of Communicable Diseases. The .shapefile for this map was obtained from ArcGIS.



Extended Data Fig. 3 | Sampling location of 501Y.V2 genomes. A general map of South Africa, showing the sampling location of the 501Y.V2 genomes in this study (blue dots) in relation to the main road networks of the country, which hints at potential land transmission routes of this lineage along the coast. The .shapefile for this map was obtained from ArcGIS.



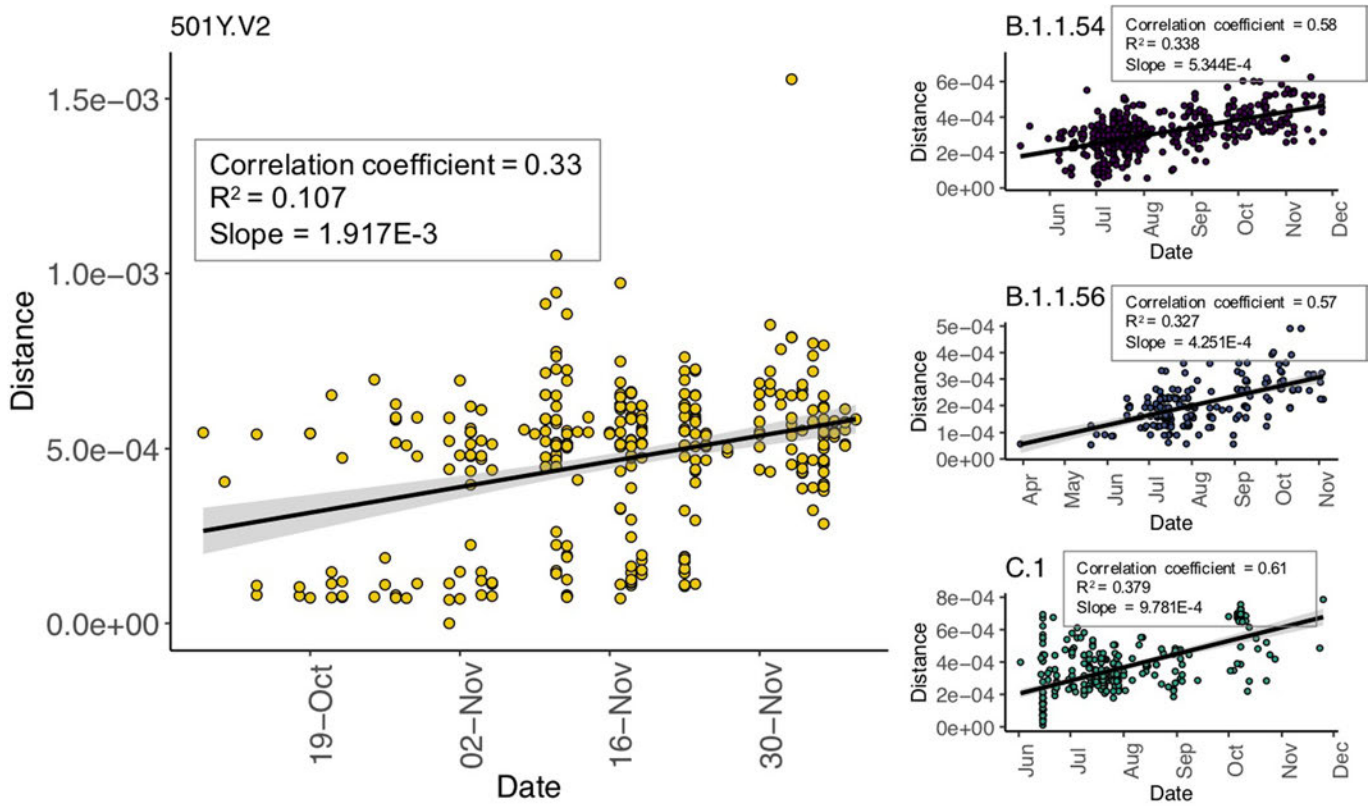
Extended Data Fig. 4 | Replacement of other lineages by the 501Y.V2 lineage. a, Progression of SARS-CoV-2 PANGOLIN lineages circulating in South Africa from March to December 2020, showing the overrepresentation of the 501Y.V2 lineage from October onwards (B.1.351, in off-white). **b**, Independent regional phylogenetic trees for the Eastern Cape, KwaZulu–Natal, Western Cape and Northern Cape, showing a variety of circulating lineages before October and the dominance of 501Y.V2 (in yellow) in late October and November (especially in the Western Cape and Eastern Cape).



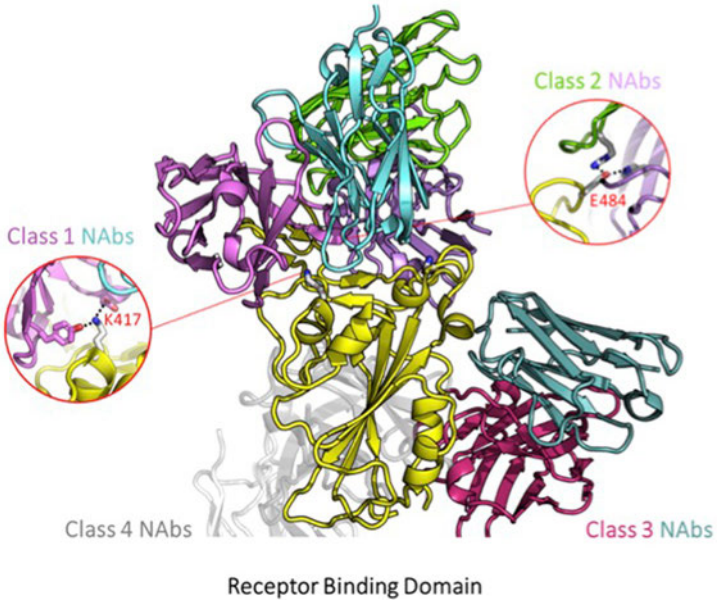
Gene	Nucleotide	Amino Acid	Status
orf1ab	1059C>T	T265I	fixed
	5230G>T	K1655N	fixed
	8660C>T	H2799Y	not fixed
	8964C>T	S2900L	not fixed
	10323A>G	K3353R	fixed
	11288-11296del	SGF 3675-3677del	fixed
	13843G>T	D4527Y	not fixed
	17999C>T	T5912I	not fixed
spike	21614C>T	L18F	not fixed
	21801A>C	D80A	fixed
	22206A>G	D215G	not fixed
	22286-22294del	L242_244Ldel	fixed
	22299G>T	R246I	not fixed
	22813G>T	K417N	not fixed
	23012G>A	E484K	fixed
	23063A>T	N501Y	fixed
	23664C>T	A701V	fixed
orf3a	25563G>T	Q57H	fixed
	25904C>T	S171L	fixed
E	26456C>T	P71L	fixed
N	28887C>T	T205I	fixed

Summary:
19 lineage defining mutations (11 fixed)
8 in spike (4 fixed)
2 deletions (1 in spike) (both fixed)

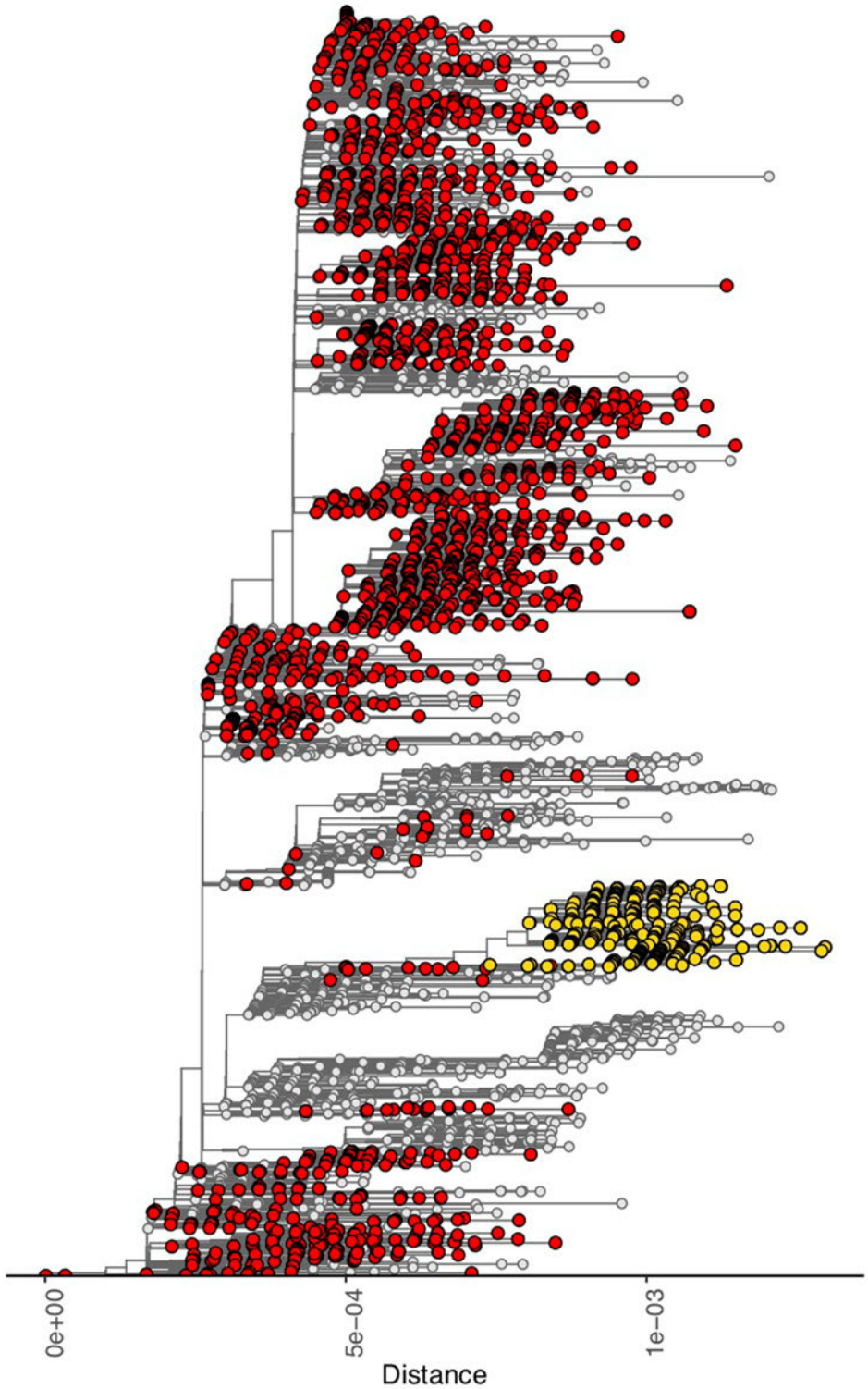
Extended Data Fig. 5 | Overview of mutations associated with the 501Y.V2 lineage. a, All nucleotide substitutions present in more than 10% of the genomes in the 501Y.V2 lineage, mapped on the SARS-CoV-2 genomic structure. Mutations present in the parent lineage (B.1) are marked in black, and mutations specific to the 501Y.V2 lineage are marked in red or blue. All nonsynonymous mutations are in bold (also reported in **b**). Blue, location of deletions on the genomes of the 501Y.V2 lineage. This is an unresolvable ambiguity in the representation of the exact location of the 22286–22294 nucleotide deletion; because of a repeat region that is hard to align (CTTT), the deletion could be any nine-nucleotide segment between 22281–22289 and 22286–22294. This means that, technically, the deletion could also be in amino acids 241–243; however, the resulting amino acid sequence of all of the possibilities are exactly the same (OTLH). **b**, Summary of all nonsynonymous lineage-defining changes in relevant genes that occur in the 501Y.V2 lineage. **c**, Allele proportions at each 501Y.V2 lineage-defining mutation site. Black line and dots, mutant allele proportion; grey line and dots, reference allele proportion in individual samples in three sequencing runs.



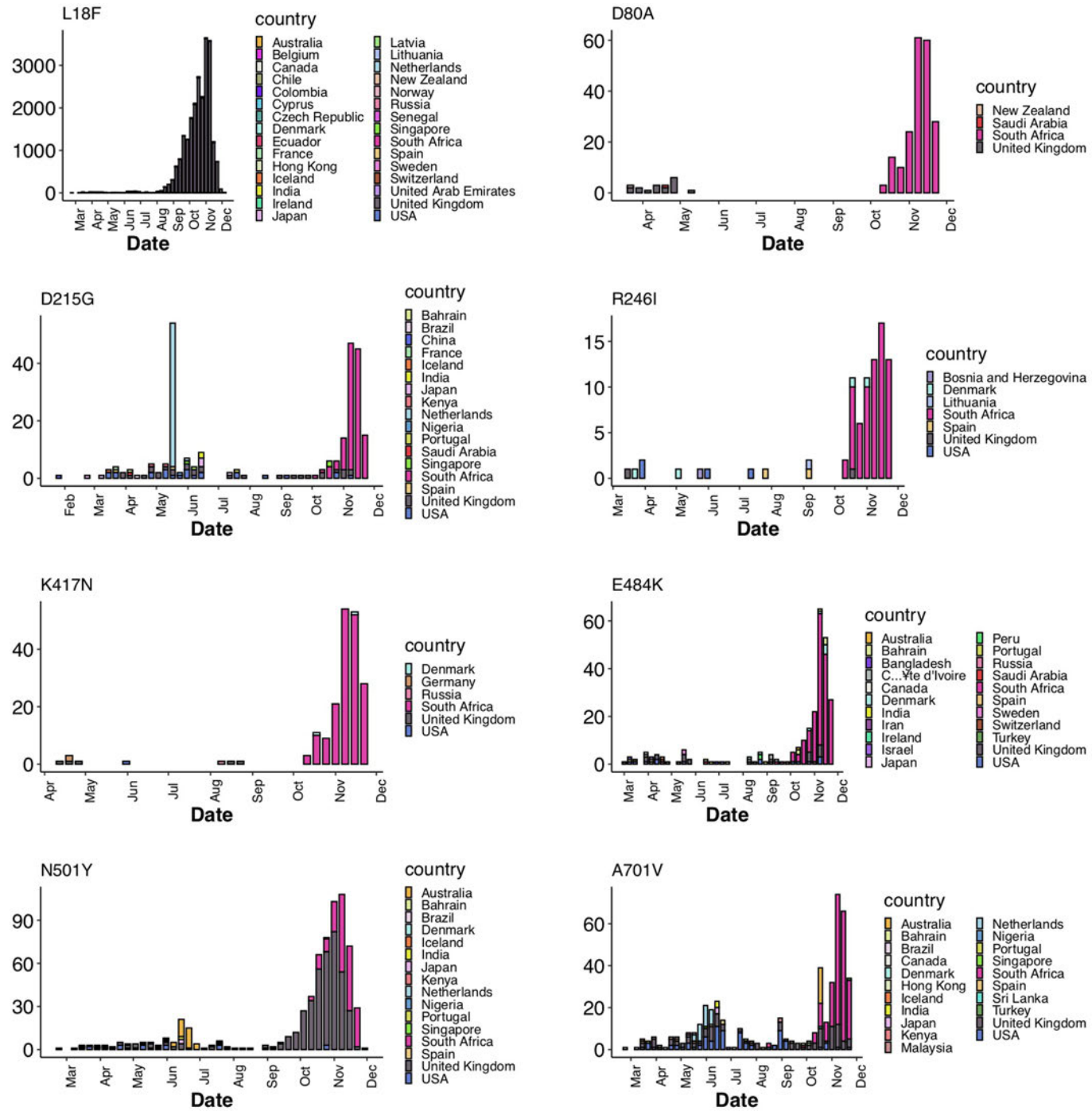
Extended Data Fig. 6 | Molecular clock signal of four main virus clusters that are spreading in South-Africa. Root-to-tip regression obtained from TempEst analysis for the 501Y.V2 lineage cluster ($n = 341$), showing a relatively strong clock-like behaviour (correlation coefficient = 0.33, $R^2 = 0.107$) and a regression line slope, representing mean evolutionary rate, of 1.917×10^{-3} nucleotide changes per site per year. We compare this with the root-to-tip regressions of the B.1.1.54 ($n = 472$), B.1.1.56 ($n = 179$) and C.1 ($n = 271$) lineages, which show estimated mean evolutionary rates of 5.344×10^{-4}, 4.251×10^{-4} and 9.781×10^{-4} respectively. Regression lines are shown with error buffers (shaded area) that represent 90% confidence intervals.



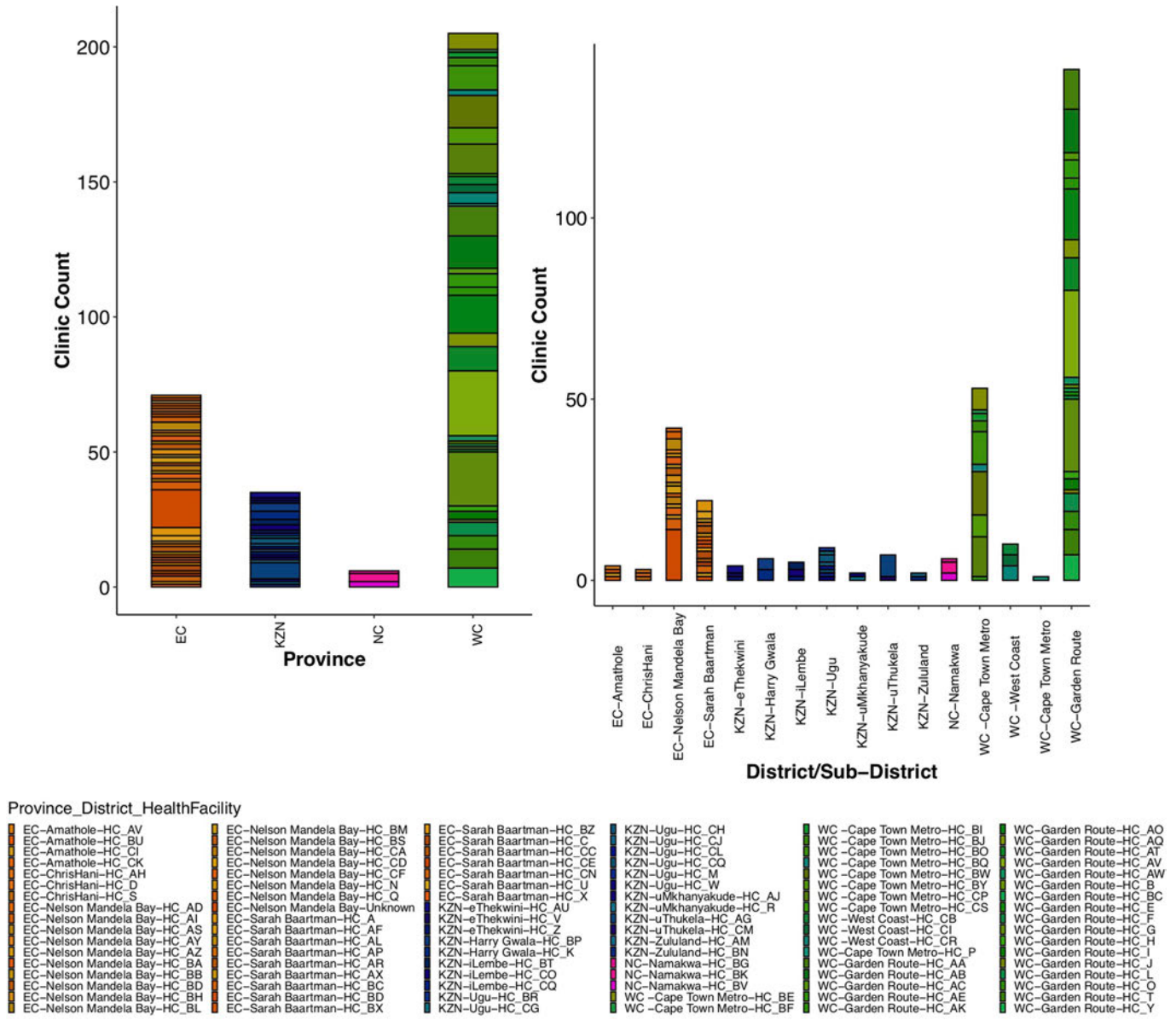
Extended Data Fig. 7 | SARS-CoV-2 RBD interactions with neutralizing antibodies. Model of the SARS-CoV-2 RBD in cartoon view (yellow), showing representative Fab domains for neutralizing antibodies (NAbs) from classes 1, 2, 3 and 4. Two zoomed-in insets show common, key interactions between the RBD residue K417 and class 1 neutralizing antibodies and the RBD residue E484 and class 2 neutralizing antibodies.



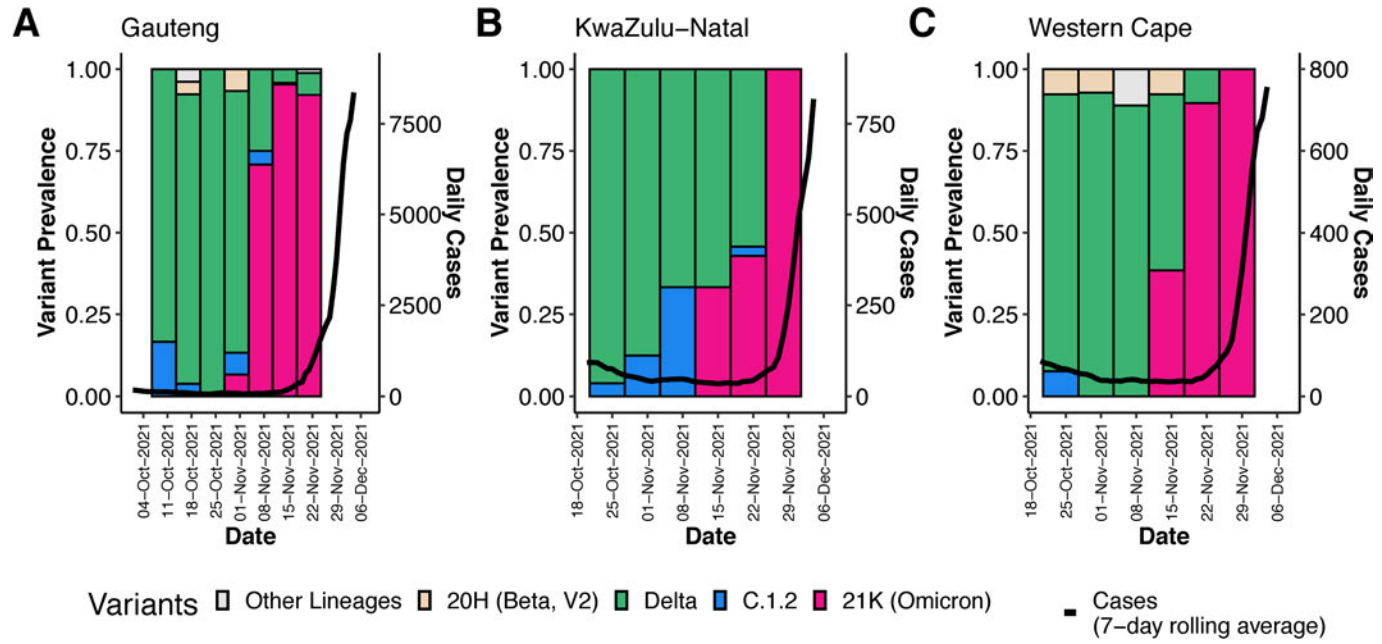
Extended Data Fig. 8 | A maximum-likelihood tree of 5,332 SARS-CoV-2 genomes, of which 2,756 are sampled from South Africa. The branch lengths represent the diversity of the genomes against the Wuhan reference. The 501Y.V2 lineages (in yellow) show relatively longer branches, compared to viral genomes from South Africa that form lineages circulating in the country before the detection of this new lineage.



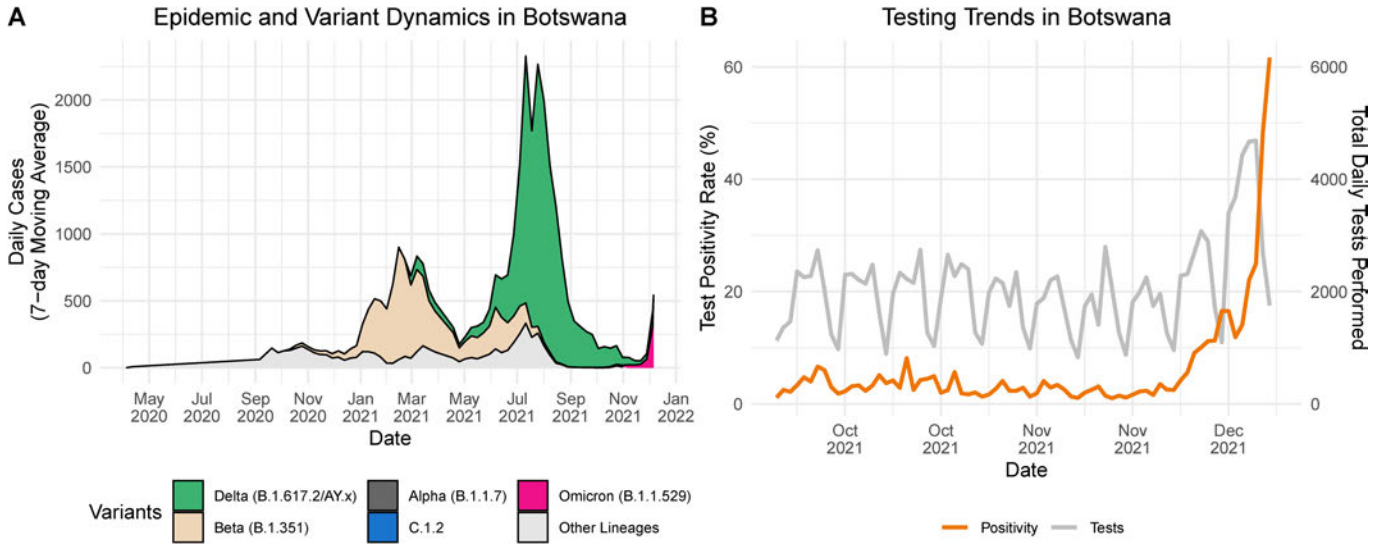
Extended Data Fig. 9 | Worldwide emergence of eight spike mutations. Prevalence of the eight spike mutations around the world, which indicates that several of these mutations have emerged independently in multiple regions.



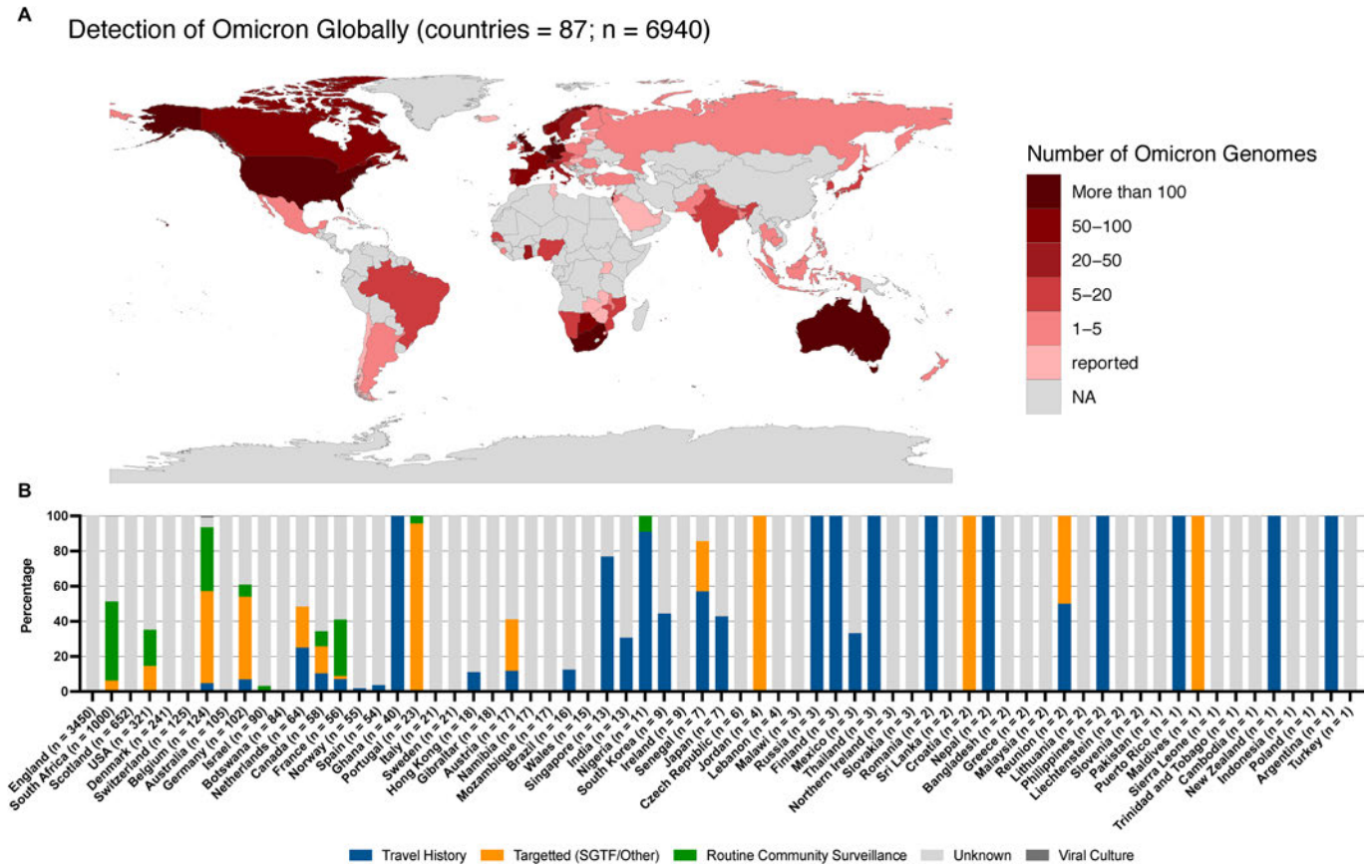
Extended Data Fig. 10 | Random sampling of 501Y.V2 samples across health centres in four provinces. Number of health centres per province in which the 501Y.V2 lineage was detected in sampled genomes (for each of the 4 provinces), showing a total number of 317 samples from 197 health centres. There was no indication of health facility for the remaining 501Y.V2 samples presented in this study.



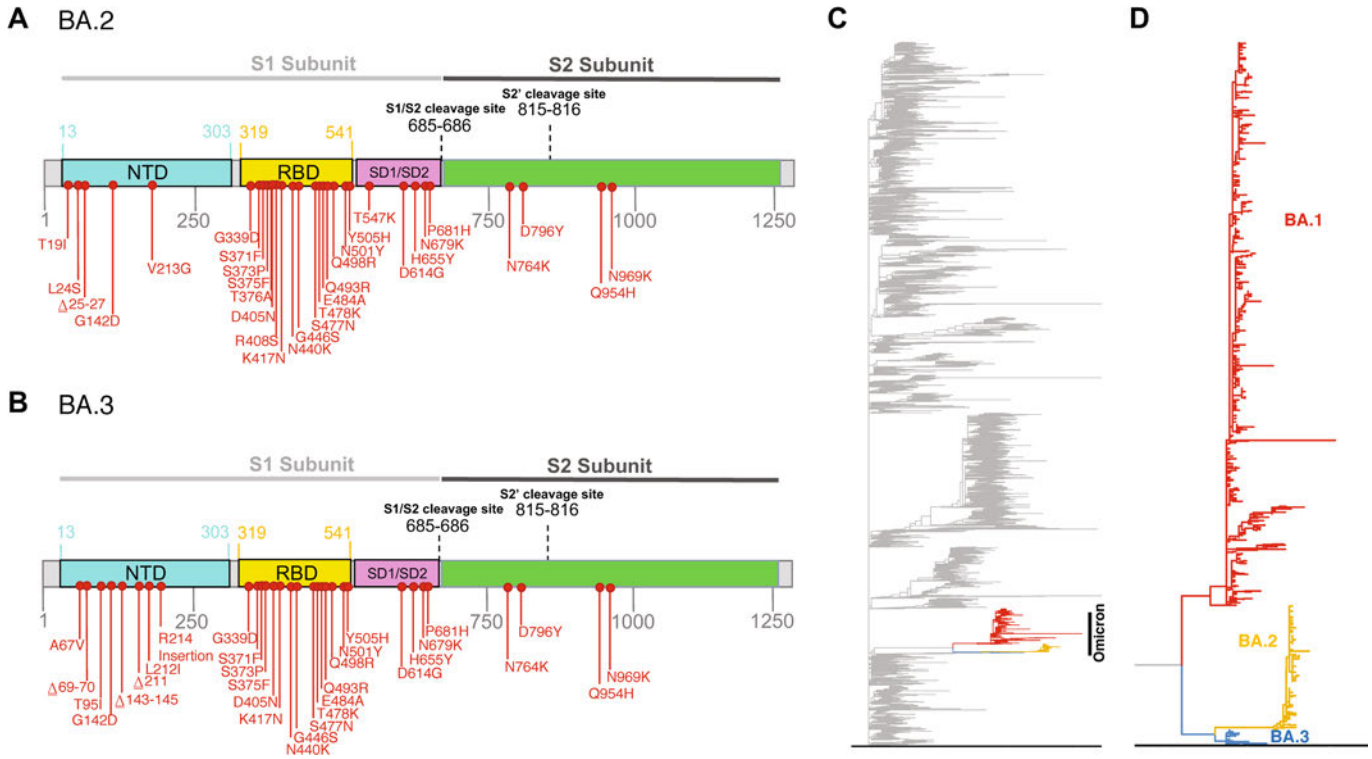
Extended Data Fig. 1 | Progression of daily recorded cases and variant proportions in Gauteng (A), KwaZulu-Natal (B) and Western Cape (C) provinces between October and December 2021. A sharp increase in the 7-day rolling average of the number of cases is observed in all three of the biggest provinces in South Africa at the emergence of the Omicron variant.



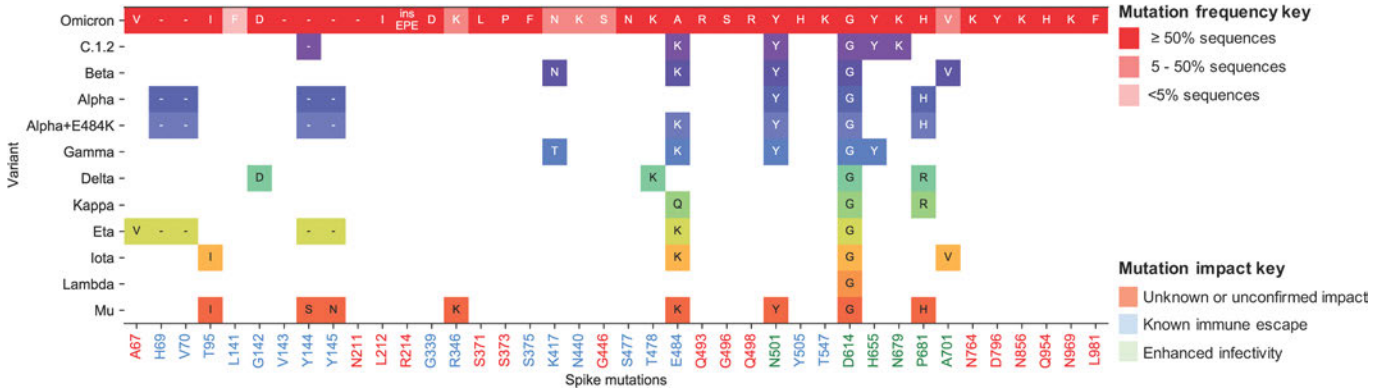
Extended Data Fig. 2 | Epidemic Progression in Botswana. A) Epidemic and variant dynamics in Botswana from May 2020 to December 2021, with the 7-day rolling average of the number of recorded cases coloured by the proportion of variants as inferred by genomic surveillance data available on GISAID. At the end of November 2021, a big Delta-driven wave was coming to its end and an Omicron wave was starting at the end of November 2021. **B**) Trends in testing numbers and positivity rates in Botswana between October and December 2021, showing a sharp increase in positivity rate mid-November 2021.



Extended Data Fig. 3 | Global distribution of Omicron. (**A**) Detection of Omicron globally. Shown are the locations for which Omicron genomes have been deposited on GISAID as of December 16, 2021. Those labelled as "reported" referred to the country from which Omicron has been reported to the WHO but there is currently no sequencing data available in GISAID, all data comes from GISAID and the WHO weekly epidemiology report Edition 70 dated December 14, 2021 (https://reliefweb.int/sites/reliefweb.int/files/resources/20211207_Weekly_Epi_Update_69-%281%29.pdf). Countries are coloured according to the number of genomes deposited with warmer colours representing more genomes. (**B**) Omicron transmission globally. Shown are countries for which Omicron sequencing data is available on GISAID. Proportions of sequences are coloured according to sampling strategy or additional host/location information from either travel history, targeted sequencing (specifically for SGTF, vaccine breakthroughs, outbreaks, contact tracing or other reasons), routine surveillance or unknown if no information has been provided. Countries are ordered by the number of sequences available on GISAID as of December 16, 2021.



Extended Data Fig. 4 | Related Lineages BA.2 and BA.3 Molecular Profile and Evolutionary Origins. A) Amino-acid mutations on the spike gene of the BA.2 **B**) Amino-acid mutations on the spike gene of the BA.3 **C**) Raw maximum likelihood phylogeny of 13,462 SARS-CoV-2 genomes, including 148 BA.2 and 19 BA.3. The newly identified SARS-CoV-2 Omicron variant is shown in colour versus grey for all other lineages. **D**) A zoomed-in view of the Omicron clade showing the evolutionary relationship between BA.1, BA.2 and BA.3.



Extended Data Fig. 5 | BA.1 spike mutations shared with other VOC/VOIs. All spike mutations seen in BA.1 are listed at the top in red and coloured according to prevalence. Prevalence was calculated by number of mutation detections / total number of sequences. However, primer drop-outs have affected the RBD region spanning K417N, N440K and G446S, and so it is likely that these mutations may actually be more prevalent than indicated here. For the VOC/VOIs only mutations that are shared with Omicron and seen in ≥50% of the respective VOC/VOI sequences are shown and are coloured according to Nextstrain clade. The mutations listed at the bottom are shaded according to known immune escape (blue), enhanced infectivity (green) or for unknown/unconfirmed impact (red).

Extended Data Fig. 6 | Maximum-likelihood trees (inferred with RAxML v8.2.12[82]) for genome regions bounding the consensus recombination breakpoints detected in lineages BA.1, BA.2 and BA.3[83]. The trees include SARS-CoV-2 genome sequences sampled in 2021 (N = 221) together with 13 sequences representing the BA.1, BA.2 and BA.3 lineages. Whereas in trees for regions 1 and 3 BA.2 and BA.3 cluster together with high bootstrap support, BA.1 is a well-supported albeit more distantly related sibling lineage. The a 897nt region 2 segment (encoding the N-terminal domain of spike) includes 67 polymorphic sites with a maximum 8nt difference between strains, showing little bootstrap support for any sibling or clade relationships except the membership of certain viruses in WHO-designated clades (Lambda, Omicron, Gamma). Despite Omicron lineages BA.1 and BA.3 clustering with certain Delta and Eta viruses and Omicron BA.2 clustering with a distinct set of Delta viruses (all on the basis of several key nucleotide positions), trees based on region 2 show no statistical support for the three Omicron lineages having distinct evolutionary origins. Bootstrap values are shown on branches with relevant values magnified for readability. All trees were rooted on the Wuhan-Hu-1 sequence.

Extended Data Table 1 | Parameter estimates from BEAST for the full South Africa and Botswana dataset and the reduced data set of only Gauteng Province genomes

Data set	Evolutionary rate $x10^{-3}$ changes/site/year	BA.1 Time of most recent common ancestor (TMRCA)	Exponential growth rate (per day)	Doubling time (days)
South Africa + Botswana 553 Genomes	1.20 (0.92, 1.49)	9 Oct 2021 (30 Sep, 20 Oct)	0.137 (0.099, 0.175)	5.1 (4.0, 7.0)
South Africa + Botswana 553 Genomes	1.1 fixed	8 Oct 2021 (30 Sep, 18 Oct)	0.137 (0.100, 0.173)	5.0 (4.0, 7.0)
South Africa + Botswana 553 Genomes	0.75 fixed	1 Oct 2021 (21 Sep, 13 Oct)	0.139 (0.099, 0.183)	5.0 (3.8, 7.0)
Gauteng Province, South Africa only 626 genomes 2021-11-05, 2021-12-07	0.41 (0.28, 0.54)	01 Oct 2021 (17 Sept, 17 Oct)	2.85 (2.10, 4.23)	2.8 (2.1, 4.2)
Gauteng Province, South Africa only 626 genomes 2021-11-05, 2021-12-07	1.1 fixed	19 Oct 2021 (15 Oct, 26 Oct)	0.29 (0.22, 0.35)	2.42 (1.96, 3.12)

95% HPD intervals in parentheses.

Extended Data Table 2 | Sites in the BA.1 sequences that have been subject to episodic diversifying selection

Coordinate (SARS-CoV-2)	Gene/ORF	Codon (in gene/ORF)	# of selected branches	AA composition	p-value	Notes
3682	ORF1a	1140	1	Q/92, L/2	0.0061	
13423	ORF1a	4387	2	R/34, H/1, N/1	0.0020	
13627	ORF1b	54	1	D/256, -/2, Y/1	0.0098	
18027	ORF1b	1520	1	A/171, -/12, Y/1, V/1	0.0006	
18030	ORF1b	1521	2	T/171, -/12, K/1, I/1	0.0052	
18267	ORF1b	1600	1	E/184, T/1, -/1	0.0001	
18273	ORF1b	1602	1	A/184, C/1, -/1	0.0001	
21534	ORF1b	2689	1	D/85, S/3	0.0066	
22027	S	156	3	E/172, -/11, G/5, P/1	0.0006	
22033	S	158	1	R/165, -/23, S/1	0.0007	
22048	S	163	1	A/168, -/20, L/1	0.0036	
22072	S	171	2	V/167, -/21, K/1	0.0000	
22084	S	175	1	F/161, -/26, Q/2	0.0000	
22576	S	339	3	D/170, -/11, G/8	0.0027	Clade defining
22597	S	346	5	R/151, K/32, -/6	0.0007	Affect Ab binding
22672	S	371	1	L/154, S/18, -/16, F/1	0.0002	Clade defining
22678	S	373	4	P/149, S/26, -/14	0.0009	Clade defining
22684	S	375	5	F/142, S/34, -/13	0.0001	Clade defining
22810	S	417	5	N/113, K/41, -/35	0.0002	Clade defining
22879	S	440	4	K/120, -/36, N/33	0.0018	Clade defining
22897	S	446	5	S/124, -/38, G/27	0.0002	Clade defining
22915	S	452	4	L/138, -/36, R/15	0.0000	Affect Ab binding
22990	S	477	3	N/148, -/23, S/18	0.0005	Clade defining
23011	S	484	3	A/141, -/26, E/21, V/1	0.0016	Clade defining
23047	S	496	3	S/151, G/21, -/17	0.0051	Clade defining
23053	S	498	2	R/148, -/21, Q/20	0.0028	Clade defining
23074	S	505	4	H/142, Y/25, -/22	0.0002	Clade defining
23095	S	512	1	V/170, -/18, T/1	0.0008	
23662	S	701	3	A/156, V/25, -/7, S/1	0.0034	501Y metasignature
23851	S	764	0	K/150, N/23, -/15, H/1	0.0010	Clade defining
24502	S	981	3	F/180, L/6, -/3	0.0084	Clade defining
25548	ORF3a	53	1	L/178, F/2	0.0099	
25707	ORF3a	106	1	L/158, F/22	0.0072	
26528	M	3	2	G/113, -/26, D/9, Y/1	0.0041	
26708	M	63	3	T/110, -/28, A/11	0.0016	Clade defining
26765	M	82	3	I/111, -/28, T/10	0.0019	
27140	M	207	1	N/105, -/42, R/1, S/1	0.0011	
27143	M	208	2	T/104, -/42, S/2, I/1	0.0066	
27146	M	209	1	D/104, -/42, A/2, Y/1	0.0008	
28253	ORF8	121	3	F/271, I/162, -/24, L/10, V/6, K/5, S/4, Q/1, D/1, C/1	0.0013	
28459	N	63	2	D/272, G/11, -/11, Y/1	0.0010	
28471	N	67	2	P/280, -/11, S/3, L/1	0.0070	
28477	N	69	1	G/282, -/11, K/2	0.0001	
28879	N	203	3	K/283, M/8, I/2, -/1, R/1	0.0088	Clade defining
29299	N	343	3	D/253, G/40, C/1, H/1	0.0002	

Extended Data Table 3 | Prior distributions used for the BDSKY analyses

Parameter	Prior distribution	
	South Africa and Botswana (n = 552)	Gauteng Province only (n = 277)
clock rate (x10^{-3} substitutions/site/year)	0.75 fixed; 1.2 fixed	1.1 fixed; 0.3 fixed
kappa	Lognormal(lnMean = 1, lnSd = 1.25)	
gamma shape	Exponential(m = 1)	
effective reproduction number	Lognormal(lnMean = 0.8, lnSd = 0.5)	
becoming non-infectious rate (per year)	36.5 fixed	
sampling proportion	Beta(alpha = 2, beta = 1000)	Beta(alpha = 2, beta = 100)
time of origin	Lognormal(lnMean = -2, lnSd = 0.2)	

The becoming non-infectious rate was fixed to 36.5/year which corresponds to a mean infectious period of 10 days. A less informative prior for the sampling proportion was used for the Gauteng Province only dataset to allow for the possibility of a higher province specific sampling proportion.

Extended Data Table 4 | Time of most recent common ancestor, exponential growth rate and doubling time estimates for the full South Africa and Botswana dataset and the reduced dataset of only Gauteng Province genomes under the 3-epoch BDSKY model in which the sampling proportion was allowed to change at 3 equidistantly spaced time points

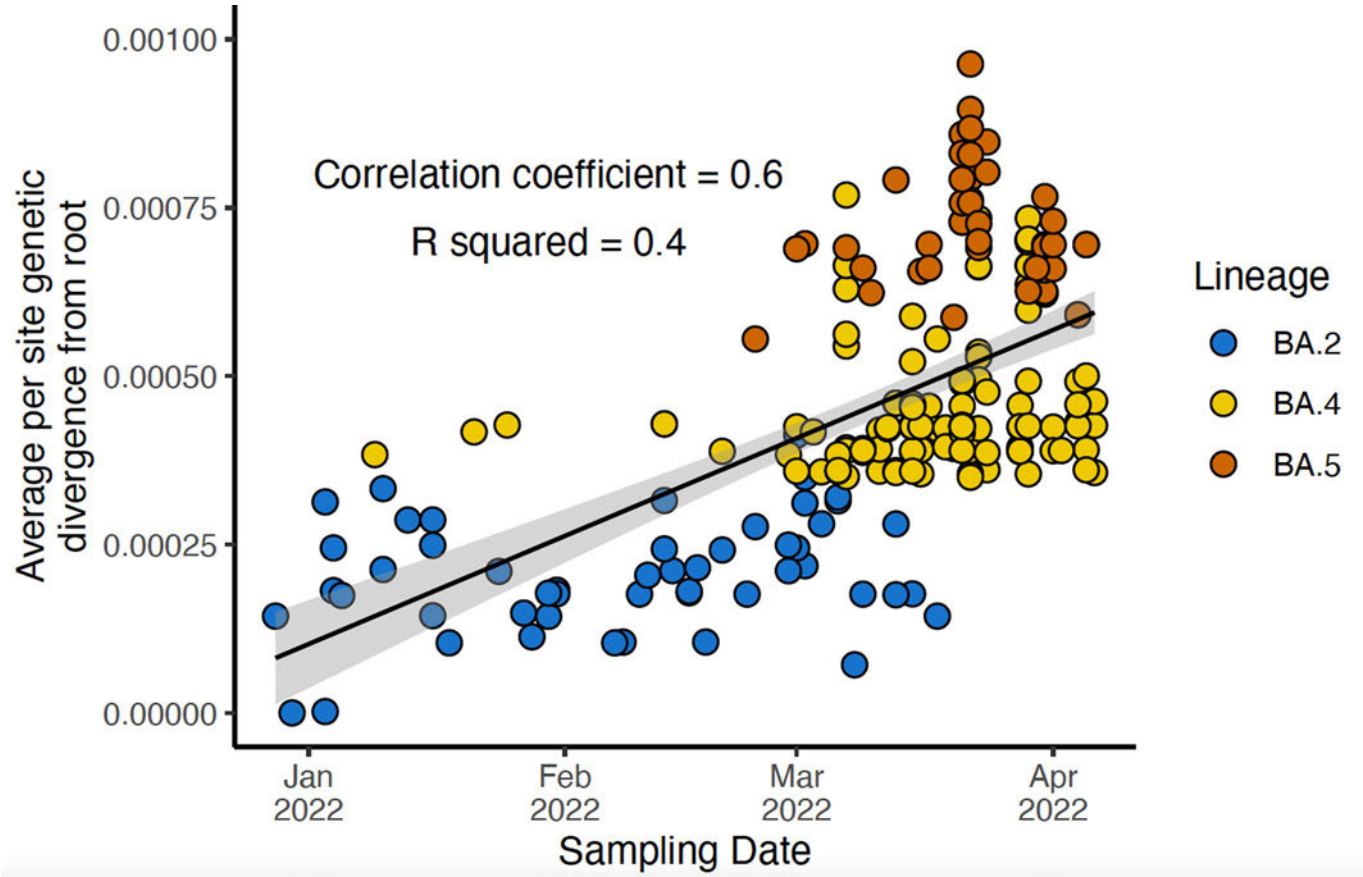
	Fixed clock rate ($x10^{-3}$ substitutions/site/year)	Time of most recent common ancestor (TMRCA)	Exponential growth rate (per day)	Doubling time (days)
South Africa and Botswana (n = 522)	1.20	20 Oct 2021 (13 Oct, 26 Oct)	0.206 (0.188, 0.226)	3.4 (3.0, 3.7)
	0.75	11 Oct 2021 (3 Oct, 18 Oct)	0.174 (0.156, 0.192)	4.0 (3.6, 4.4)
Gauteng Province only (n = 277)	0.30	4 Oct 2021 (24 Sep, 12 Oct)	0.191 (0.151, 0.231)	3.6 (2.9, 4.5)
	1.1	24 Oct 2021 (19 Oct, 29 Oct)	0.286 (0.243, 0.329)	2.4 (2.1, 2.8)

95% HPD intervals in parentheses.

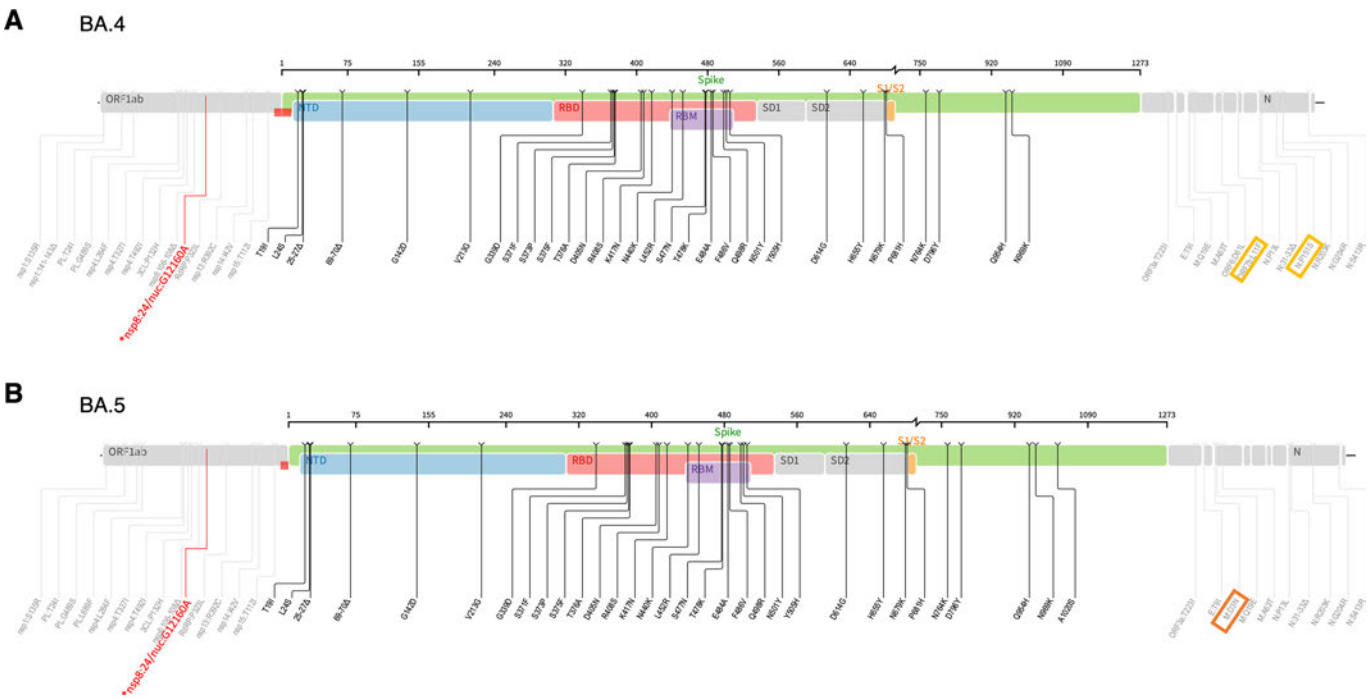
Extended Data Table 5 | Time of most recent common ancestor, exponential growth rate and doubling time estimates for the full South Africa and Botswana dataset and the reduced dataset of only Gauteng Province genomes under the 4-epoch BDSKY model in which the sampling proportion was allowed to change at 4 equidistantly spaced time points

	Fixed clock rate ($x10^{-3}$ substitutions/site/year)	Time of most recent common ancestor (TMRCA)	Exponential growth rate (per day)	Doubling time (days)
South Africa and Botswana (n = 522)	1.20	19 Oct 2021 (13 Oct, 25 Oct)	0.205 (0.186, 0.225)	3.4 (3.1, 3.7)
	0.75	11 Oct 2021 (2 Oct, 17 Oct)	0.179 (0.160, 0.197)	3.9 (3.5, 4.3)
Gauteng Province only (n = 277)	0.30	27 Sep 2021 (16 Sep, 7 Oct)	0.146 (0.114, 0.180)	4.8 (3.8, 5.9)
	1.1	23 Oct 2021 (17 Oct, 28 Oct)	0.261 (0.220, 0.302)	2.7 (2.3, 3.1)

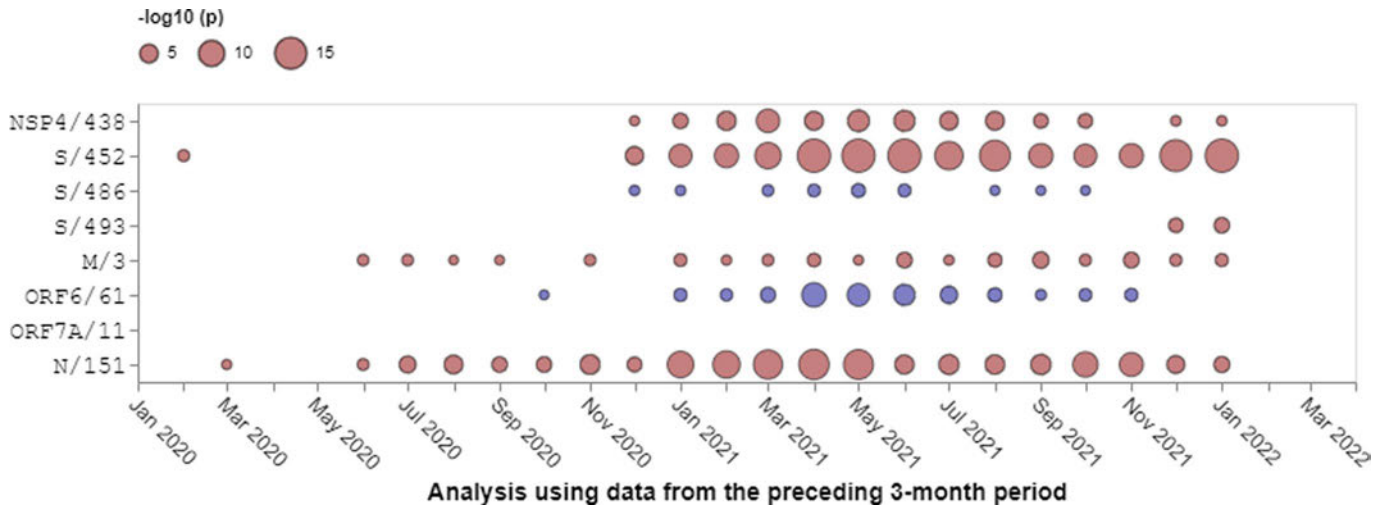
95% HPD intervals in parentheses.



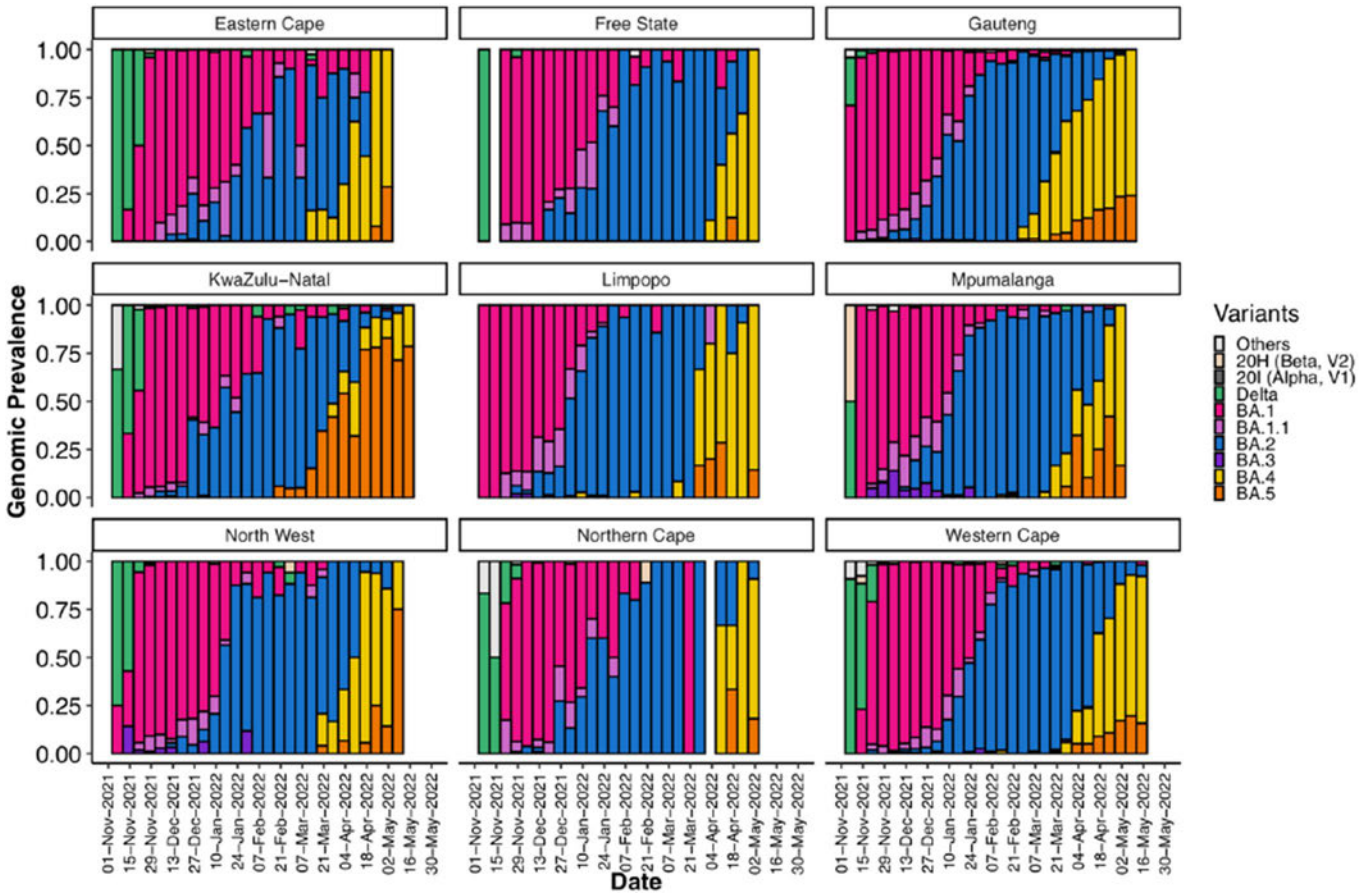
Extended Data Fig. 1 | Molecular clock signal of the dataset of BA.2, BA.4 and BA.5 lineages used in the Bayesian analysis. Root-to-tip regression obtained from TempEst analysis for the sampled cluster of BA.2, BA.4 and BA.5, showing a relatively strong clock-like behaviour (correlation coefficient = 0.6, R2 = 0.4) The regression line (representing the estimated mean evolutionary rate) is shown with error buffers (shaded area) that represent 90% confidence intervals.



Extended Data Fig. 2 | Whole genome mutations present in BA.4 and BA.5 lineages. Differences in BA.4 and BA.5 are highlighted with a rectangle. The synonymous mutations in nsp8 is indicated in red.



Extended Data Fig. 3 | Patterns of natural selection between January 2020 and January 2022 at codon sites differentiating BA.4 and BA.5 from BA.2. All SARS-CoV-2 sequences deposited in GISAID were analyzed with each time-point representing an analysis of all sequences sampled during the preceding three months. Red dots indicate evidence at positive selection and blue spots indicate evidence of negative selection. The sizes of the dots indicate degrees of statistical support for selection signals. Only sequences deposited in GISAID prior to the discovery of BA.4 and BA.5 are considered here.



Extended Data Fig. 4 | Progression of the weekly genomic prevalence of various variants and lineages in the nine provinces of South Africa from November 2021 to May 2022.

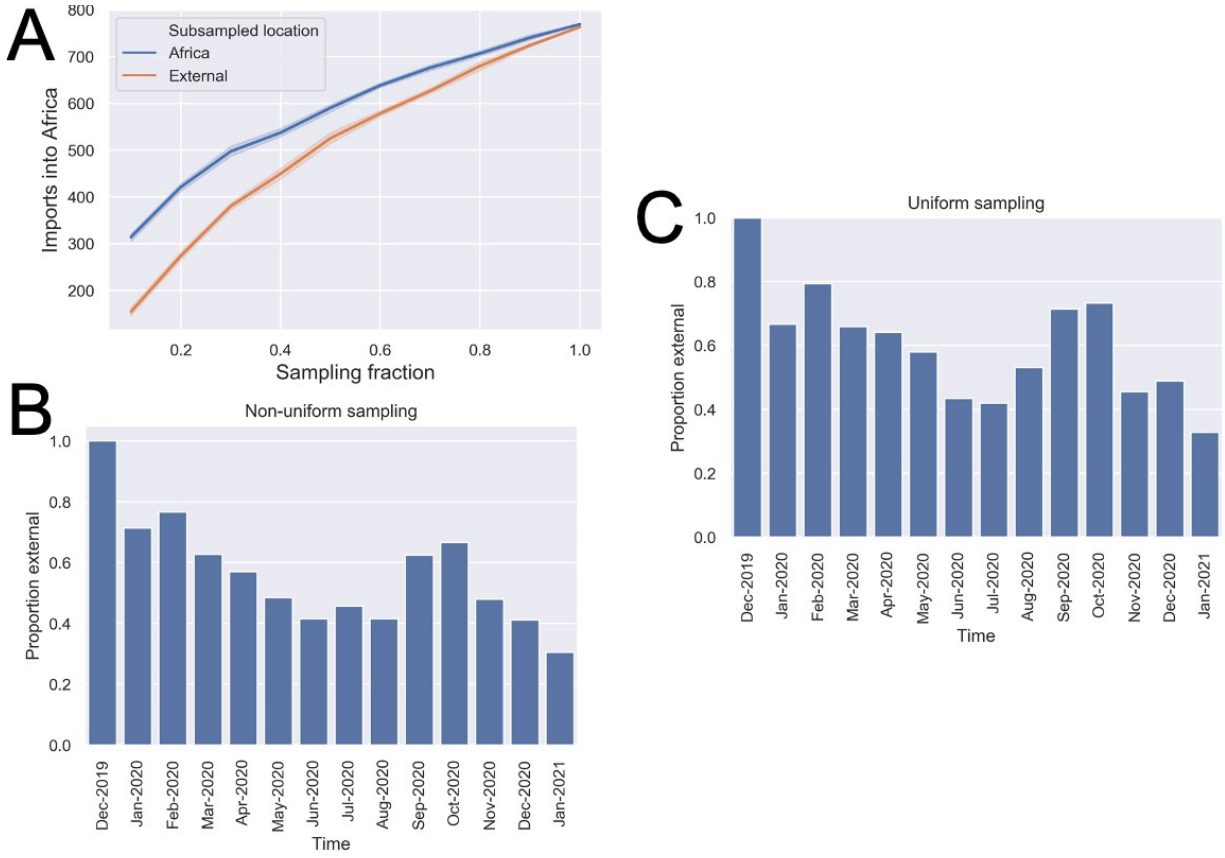
Extended Data Table 1 | S-gene target status (TaqPath COVID-19 qPCR assay) for 198 samples sequenced by the KRISP laboratory. *One BA.2 sequence had the 69–70 deletion, and the other BA.2 sequence had large gaps in coverage of the spike gene region

Omicron lineage	S-gene target failure	S-gene target positive
BA.1	9	0
BA.2	2*	120
BA.3	0	1
BA.4	26	0
BA.5	40	0
Total	77	121

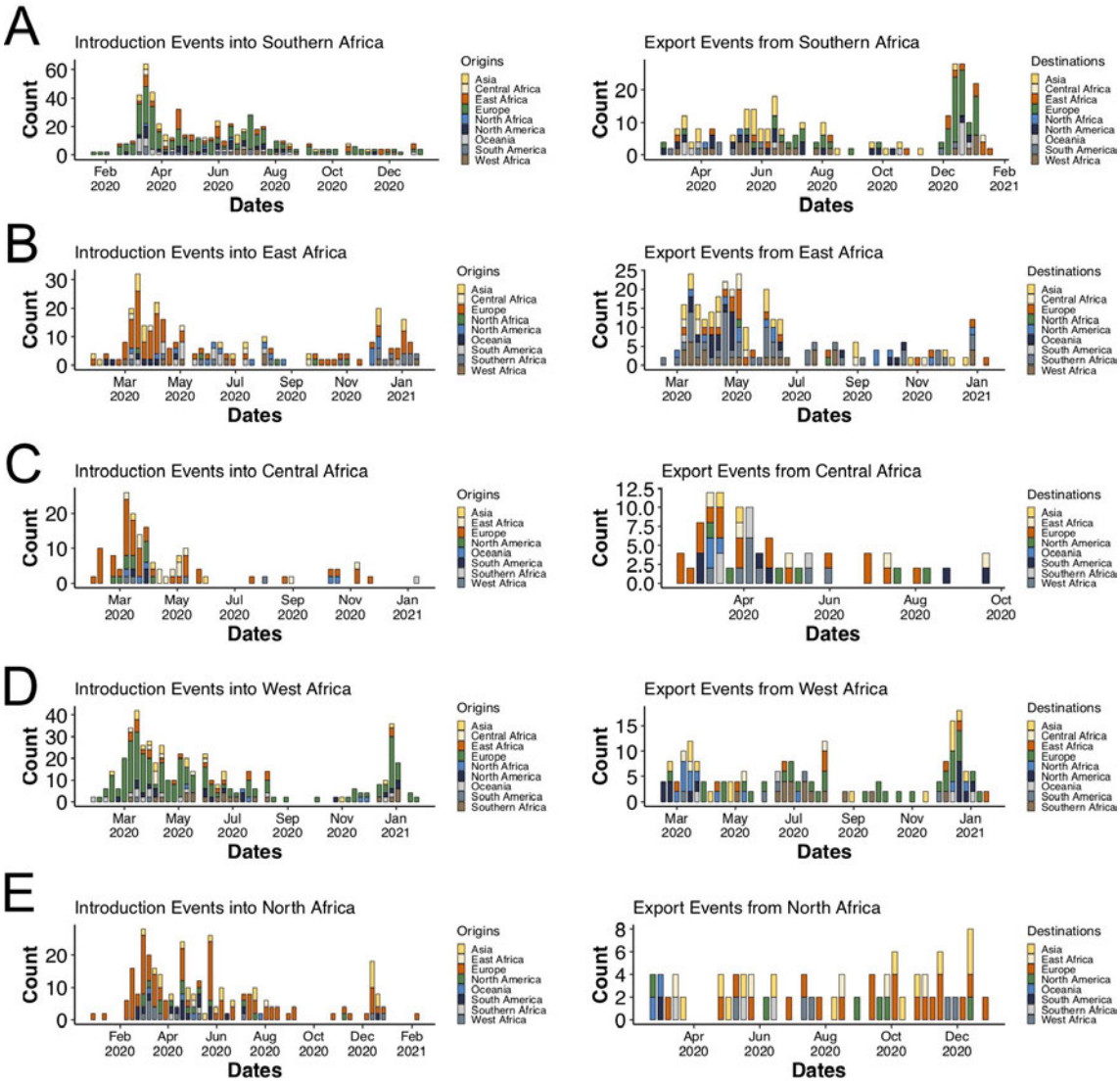
Extended Data Table 2 | Comparison of daily growth rates of all Omicron lineages and Delta. Rates were estimated with multinomial logistic regression models based on South African SARS-CoV-2 genomic data spanning the period of 1 November 2021 to 19 May 2022. Negative values indicate the comparative lineage to have a growth advantage over the reference lineage, whereas a positive value indicates the reference lineage to have a growth rate advantage over the lineage of comparison

Reference lineage	Comparative lineage	Growth rate per day	95% Confidence Intervals
BA.5	BA.1	0.164	0.154 – 0.175
	BA.2	0.096	0.086 – 0.106
	BA.3	0.154	0.138 – 0.170
	BA.4	-0.014	-0.023 – -0.005
	Delta	0.235	0.094 – 0.121
BA.4	BA.1	0.15	0.143 – 0.158
	BA.2	0.082	0.075 – 0.089
	BA.3	0.14	0.126 – 0.154
	Delta	0.221	0.204 – 0.239
BA.3	BA.1	-0.01	-0.022 – -0.002
	BA.2	-0.058	-0.070 – -0.046
	Delta	0.0814	0.062 – 0.101
BA.2	BA.1	0.068	0.065 – 0.072
	Delta	0.139	0.123 – 0.155
BA.1	Delta	0.071	0.056 – 0.087

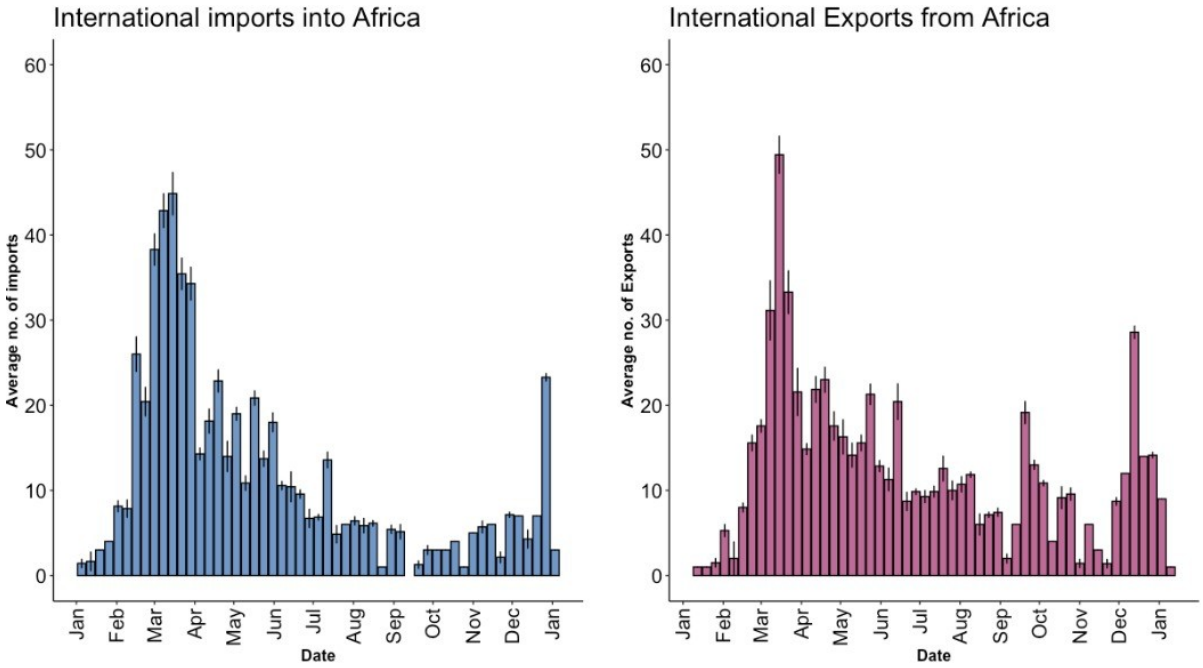
Supplementary Figures & Tables



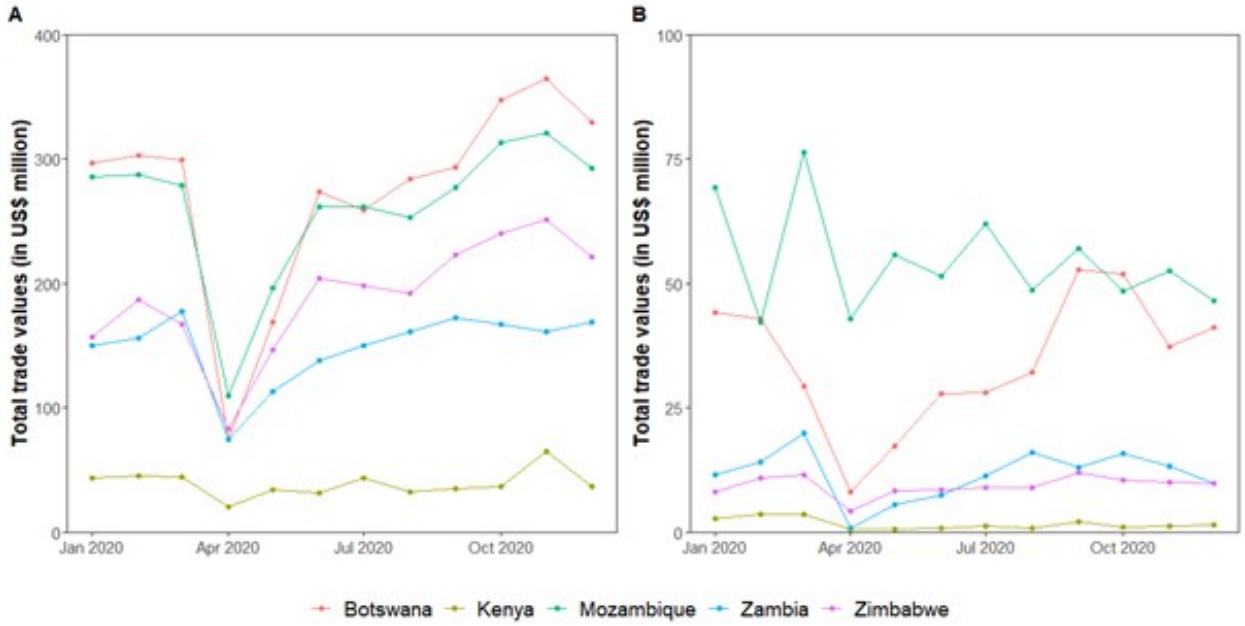
Supplementary Figure S1: *Sensitivity of the viral introduction analysis to geographic sampling biases. (A) A rarefaction analysis showing how the number of imports into Africa depends on the extent of sampling in Africa (blue) and the extent of external sampling in the rest of the world (orange). At each sampling fraction, a random set of samples was subsampled from the full dataset 10 times to create bootstrap replicates from which confidence intervals (shaded intervals) on the number of imports were computed. (B-C) Sensitivity analysis showing how the proportion of imports into African countries from external locations outside of Africa varied depending on the temporal distribution of samples in Africa. This analysis was performed twice with either non-uniform sampling through time using the same dataset as in Figure 2B-C of the main text (B) or uniform sampling through time in which we capped the number of samples from Africa at a maximum threshold of 400 each month.*



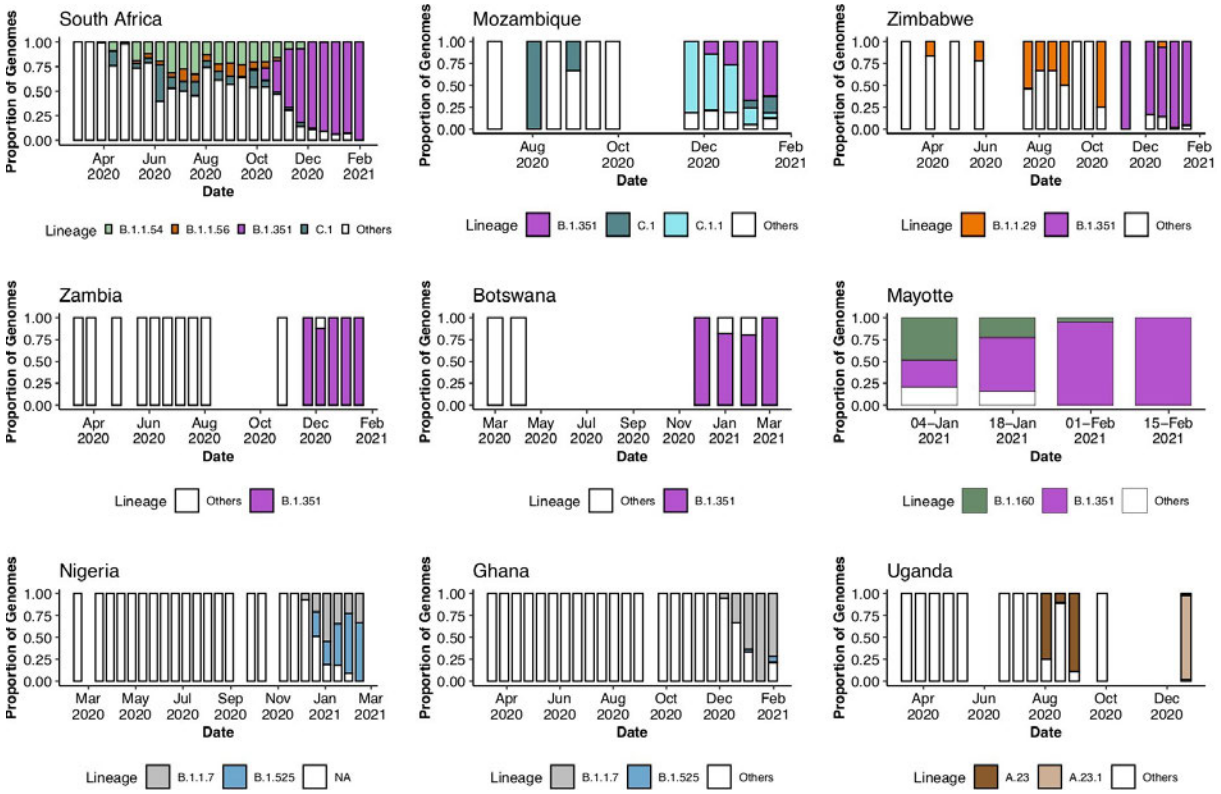
Supplementary Figure S2: *Number of importation and exportation events for various subregions on the African continent. African subregions are defined based on the African Union classification scheme.*



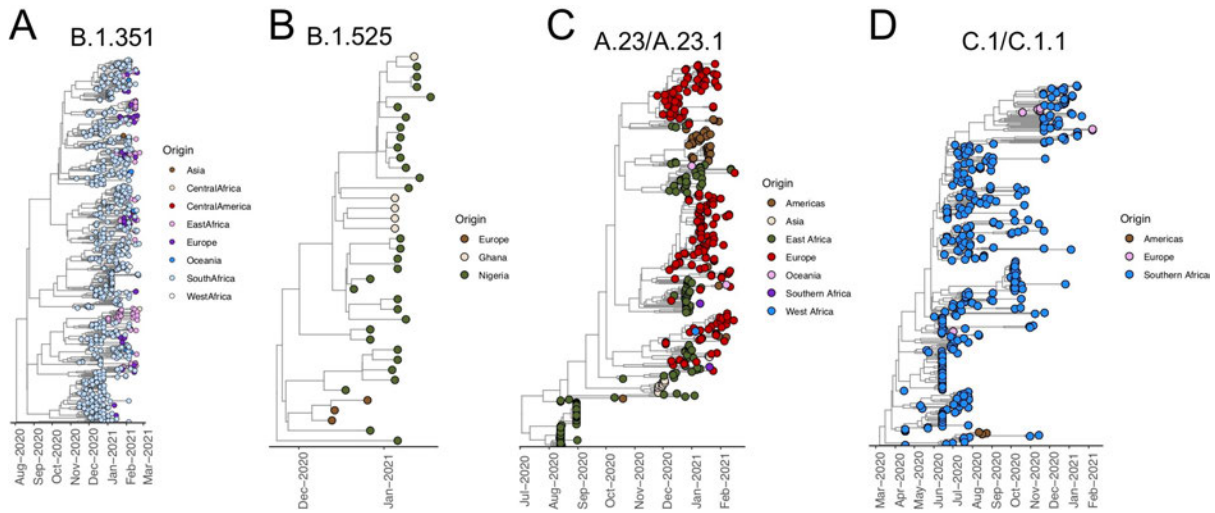
Supplementary Figure S3: *Numbers of importation and exportation events between Africa and the rest of the world over the first year of the SARS-CoV-2 pandemic.*



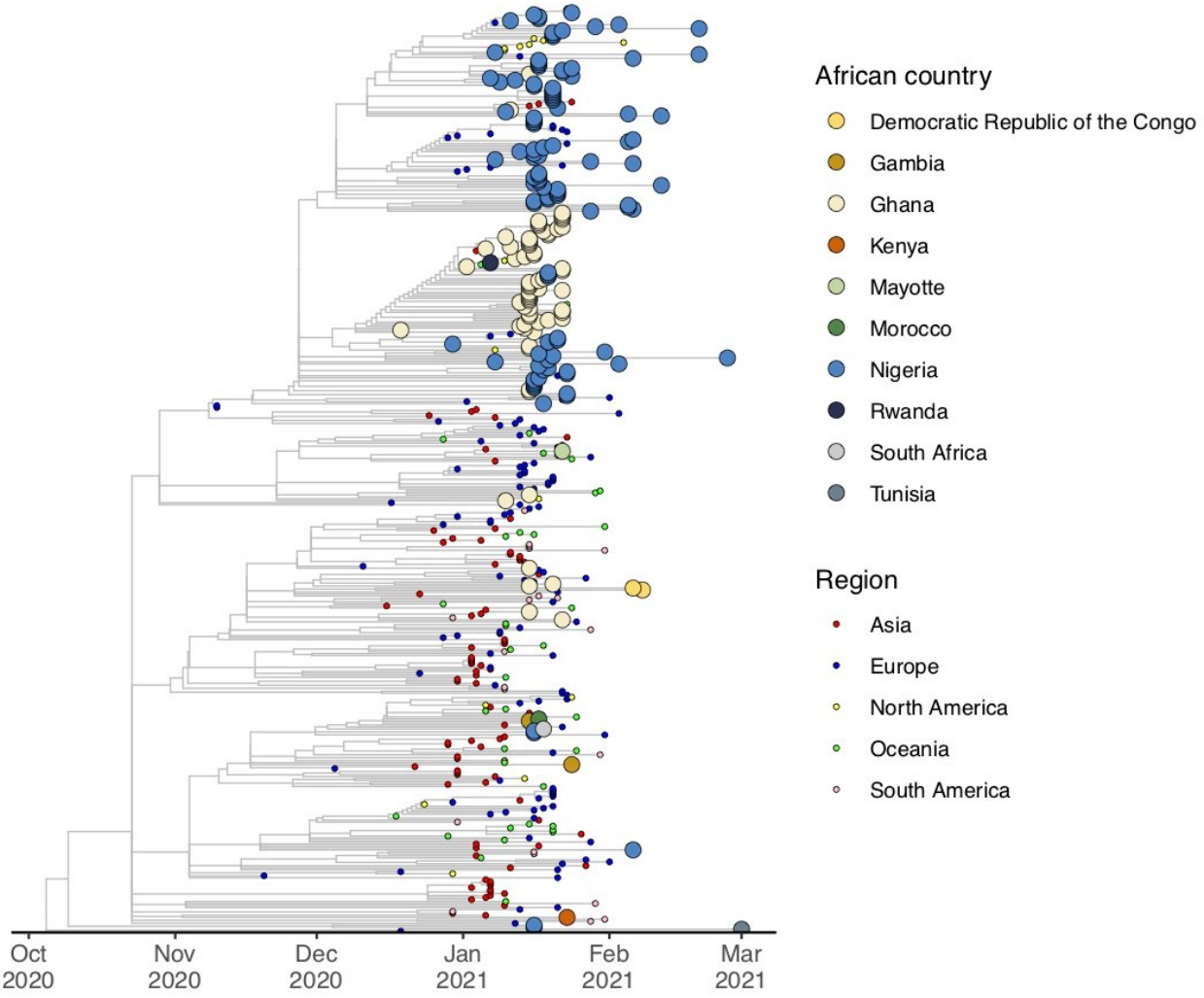
Supplementary Figure S4: *Total monthly international trade values in US million dollars in 2020 for A) exported goods from South Africa; and B) imported goods to South Africa with the following neighbouring countries: Botswana, Democratic Republic of the Congo, Eswatini, Lesotho, Malawi, Mozambique, Namibia, Zambia, and Zimbabwe. Source: UN Comtrade Database.*



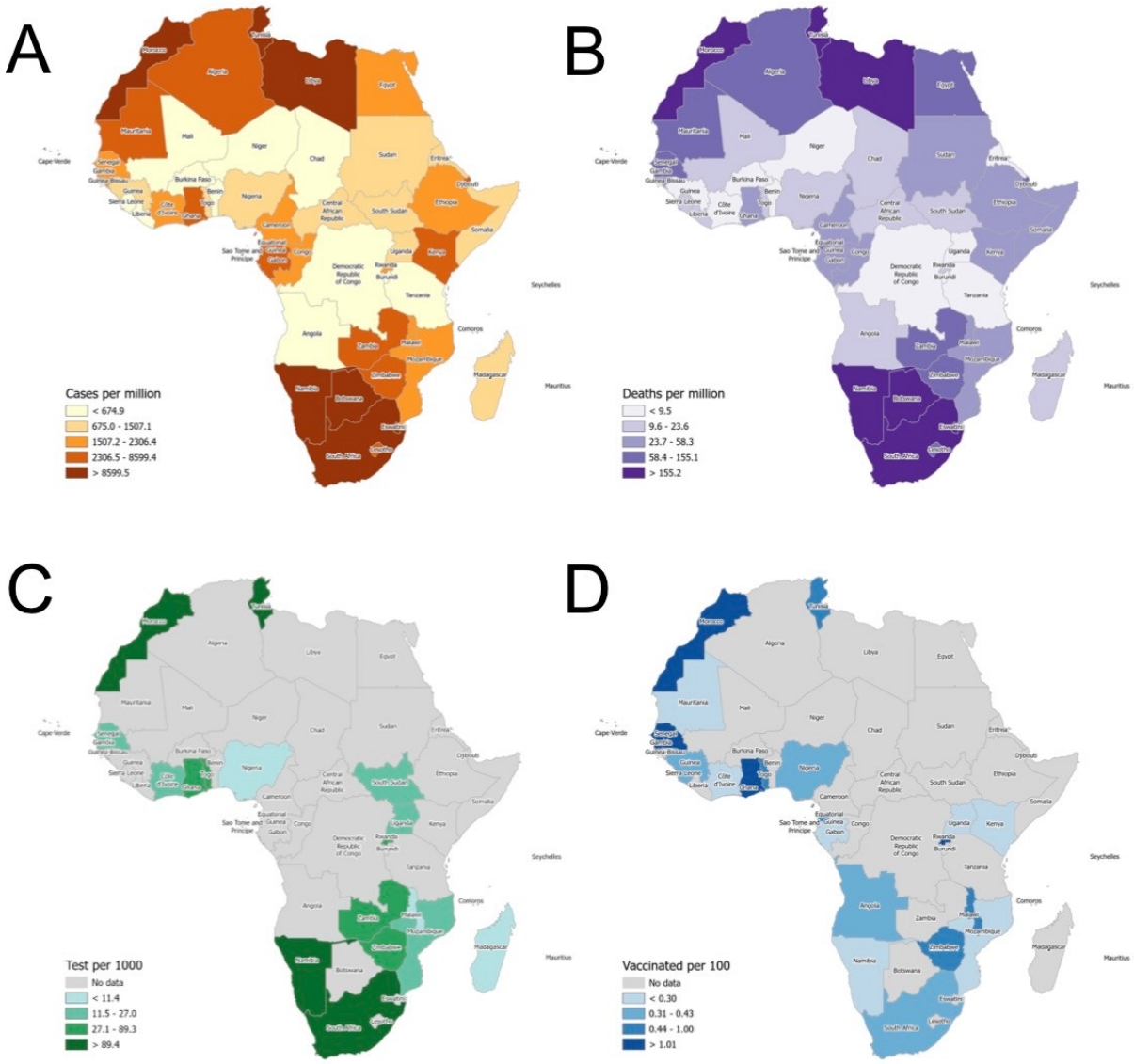
Supplementary Figure S5: *PANGO lineages through time for a select number of African countries.*



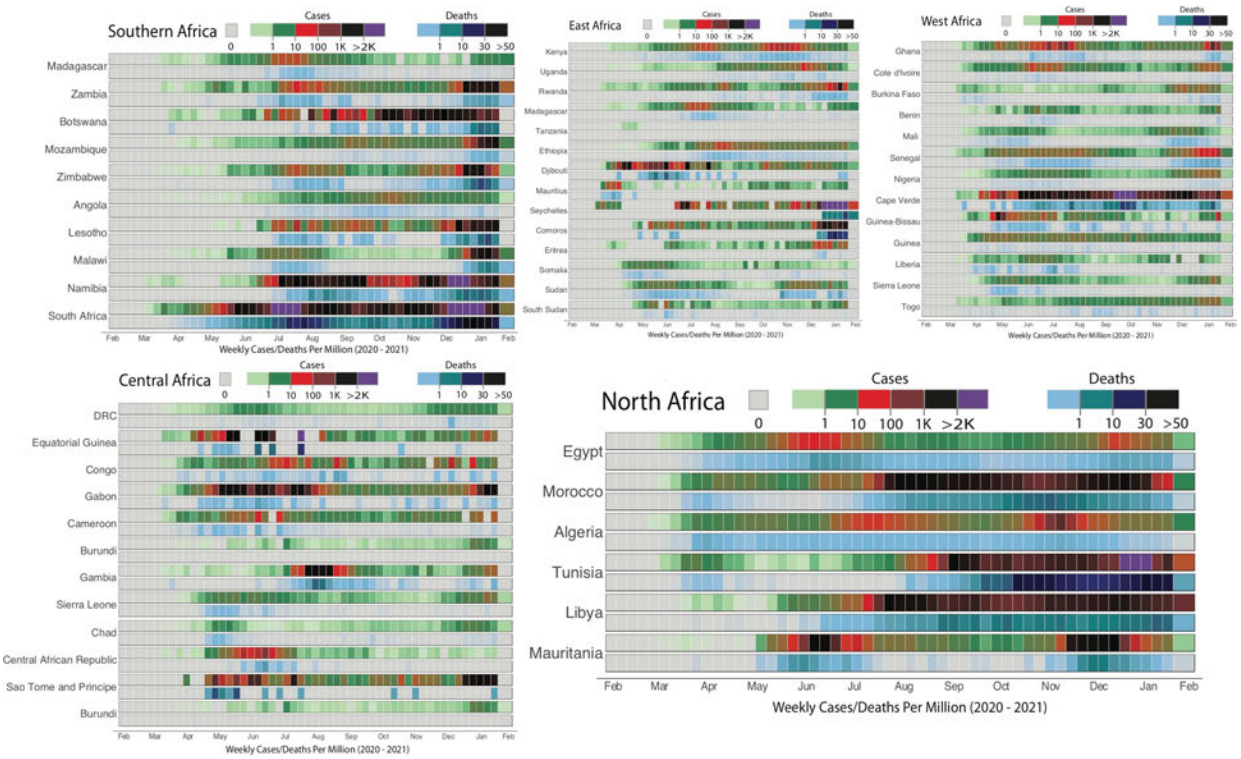
Supplementary Figure S6: *Maximum clade credibility phylogeographic trees including all global VOC or VOI samples. Branch colours represent most probable inferred locations of ancestral viruses. Numbers at internal nodes represent clade posterior probabilities.*



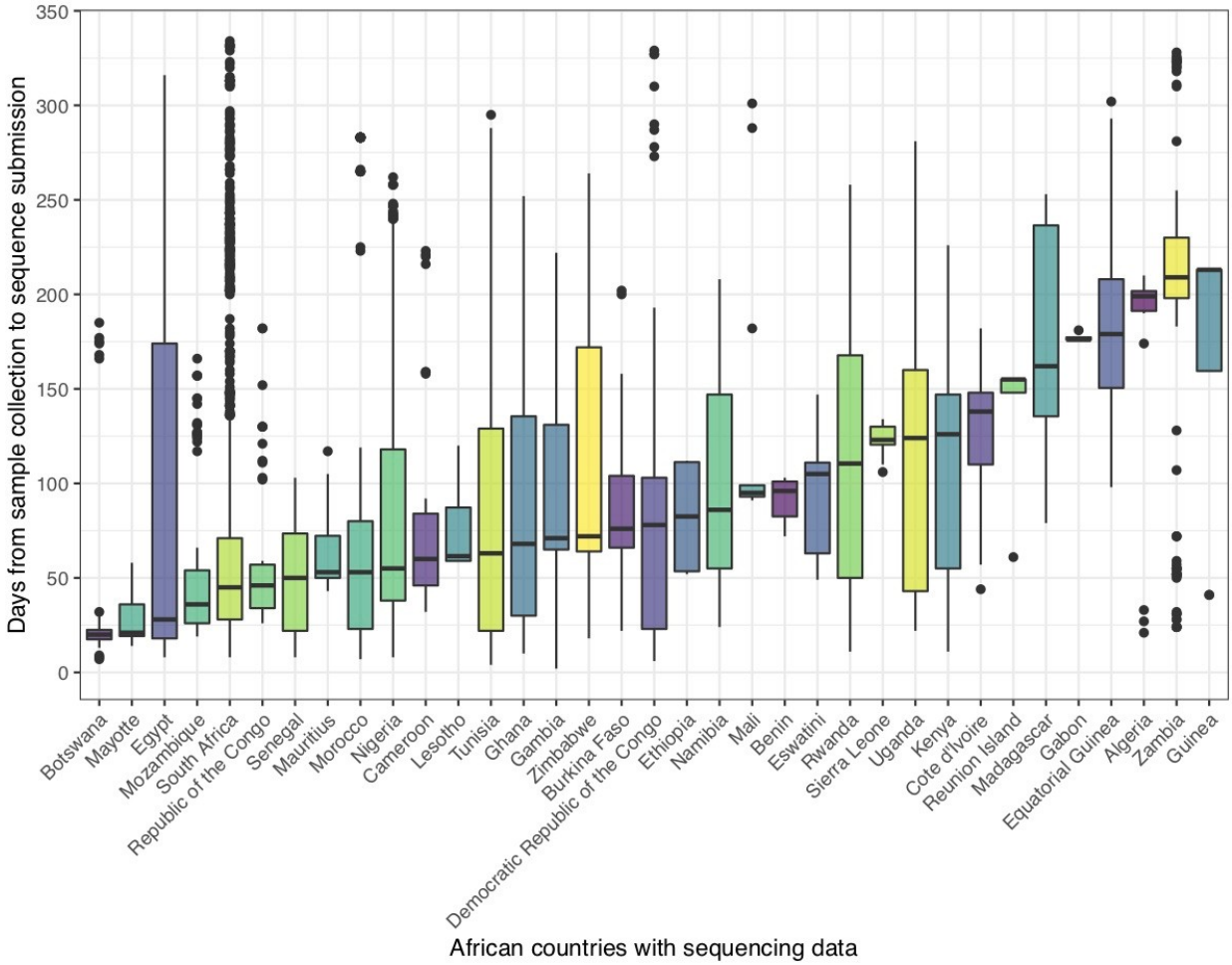
Supplementary Figure S7: *Time scaled phylogeny of the B.1.1.7 lineage. This phylogenetic cluster was extracted from the large dated phylogeny in Figure 2A. African sequences are highlighted by large circles, while non-African sequences appear as smaller dots. The branches are scaled in calendar time.*



Supplementary Figure S8: *Epidemiological metricises of COVID-19 on the African continent. Clockwise from top left: reported COVID-19 cases per million individuals; reported COVID-19 attributed mortalities per million individuals; numbers of COVID-19 tests performed per 1,000 individuals; and numbers vaccinated per 100 individuals.*



Supplementary Figure S9: *Epidemiological heatmaps of cases and deaths for various subregions on the African continent. African subregions are defined based on the African Union classification scheme.*



Supplementary Figure S10: *Graph of days from sampling to submission in various African countries.*

Supplementary Table S1. *Status and restrictions of land border posts in South Africa as of Feb 19, 2021.*

Country route	Number of land border posts		Restrictions
	Closed (n/N)	Open (n/N)	
South Africa - Botswana	13/17	4/17	· All passengers passing through the border posts are required to present a medical certificate with a negative COVID-19 test result issued within 72 hours or get tested upon arrival and subject to quarantine in a government holding facility. The entry to Zimbabwe requires a negative COVID-19 test result that is within 48 hours. · Rail, ocean, air and road transport is permitted for the movement of cargo to and from other countries, subject to national legislation and any directions. · All borders were closed on Jan 11, 2021 then reopened on February 15, 2021.
South Africa - eSwatini	6/11	5/11	
South Africa - Lesotho	7/13	6/13	
South Africa - Mozambique	2/4	2/4	
South Africa - Namibia	4/6	2/6	
South Africa - Zimbabwe	0/1	1/1	

Supplementary Table S2: *Variants of Concern/Note (VoC/Ns) in Africa.*

Variant Name	Lineage	Date Range	Spike Mutations of Biological Significance (all mutations)	Impact	Countries
N501Y.V2	B.1.351	Oct. 2020 – Feb. 2021	K417N, E484K, N501Y	Transmissibility, Escape Neutralization, ACE binding Affinity	South Africa, DRC, Mayotte, La Reunion, Zambia, Botswana, Congo, Kenya, Rwanda,
A.23, A.23.1	A.23.1	Dec. 2020 – Feb 2021	V367F, Q613H	Infectivity	Uganda, Rwanda, Ghana, South Africa, Zambia, Botswana
C.1.1	C.1.		S477N		Mozambique,
B.1.525	B.1.525	Dec. 2020 - Feb 2021	E484K, Q677H, F888L	Escape Neutralization, ACE binding Affinity	Nigeria, Ghana, Mayotte, Côte d'Ivoire/Bouaké Algeria
A.27/N501 Y.V4	A.27	Jan 2021 - Feb 2021	L18F, L452R, N501Y, A653V, H655Y, Q677H, D796Y, G1219V	under investigation (VUI not VOC)	Mayotte, Europe, Ghana, Côte d'Ivoire/Bouaké
N501Y.V3					Brazil
B.1.160	B.1.160		D614G, S477N	confirmed reinfection (**under investigation**)	**Tunisia** (reinfection), Large European lineage Ghana
N501Y	B.1.1.7	Jan - Mash2021	D614G, N501Y, del69-70,	Transmissibility	**Ghana, Morocco Algeria,** Côte d'Ivoire/Bouaké, DRC

Supplementary Table S3: *Sampling or surveillance strategies in various participating institutions.*

Country	Proportion of cases sequenced	Sampling strategies				Other (details)
		Regular surveillance (random sampling)	Cluster/outbreak investigations	Surveillance of imported cases (linked to border testing)	Investigation of re-infections	
South Africa	0.20%	Yes	Yes	No	Yes	Sequencing of infections in vaccine trials Sequencing for health facility-based and community-based research projects
Zambia	0.27% (0.42%)	Yes	Yes	Yes	Yes	Not all investigations are being performed at all times. When cases exceed a particular threshold cluster, random and imported case surveillance reduces or stops. Total cases 8/2/21 = 63.573, 8/3/21 = 82,421.

Democratic Republic of Congo (DRC)	1.4 % (2.87%)	Yes	No	Yes	No	Regular surveillance is based on samples availability; the surveillance of imported cases is based on samples of travellers coming in DRC. there are also "sequencing based on a research project focused on respiratory infections (Andemia)
South Africa (FS)		Yes	No	No	No	All samples with Cts lower that 30 are stored (with storage record). From 5 districts samples are selected randomly on a week basis (10 - 30) per district. From the ~15 000 stored samples no repeat testing has been identified within less than 90 days.
Ghana (Uhas)	0.36% (0.12%)	Yes	Yes	No	No	Random surveillance based on clusters of cases. During periods of

						suspected widespread infections, cases are randomly selected and sequenced.
Tunisia	0.04%	Yes	No	No	Yes	Random surveillance. Cases are randomly selected and sequenced. Some suspected reinfection cases are now tested in Sfax (Tunisia).
Morocco		Yes	Yes	Yes	Yes	Sequencing of 10% of Sample that are positif for S drop real time PCR test using (taqPath kit from thermo) . Sanger Sequencing of the entire S gene for the confirmation of mutation related to new varriants. WGS for the genomic surveillance over time et geographical localtion.
Equatorial Guinea	3.10%	Yes	YES	Yes	No	During the first wave from March to August, all positive samples were stored and a

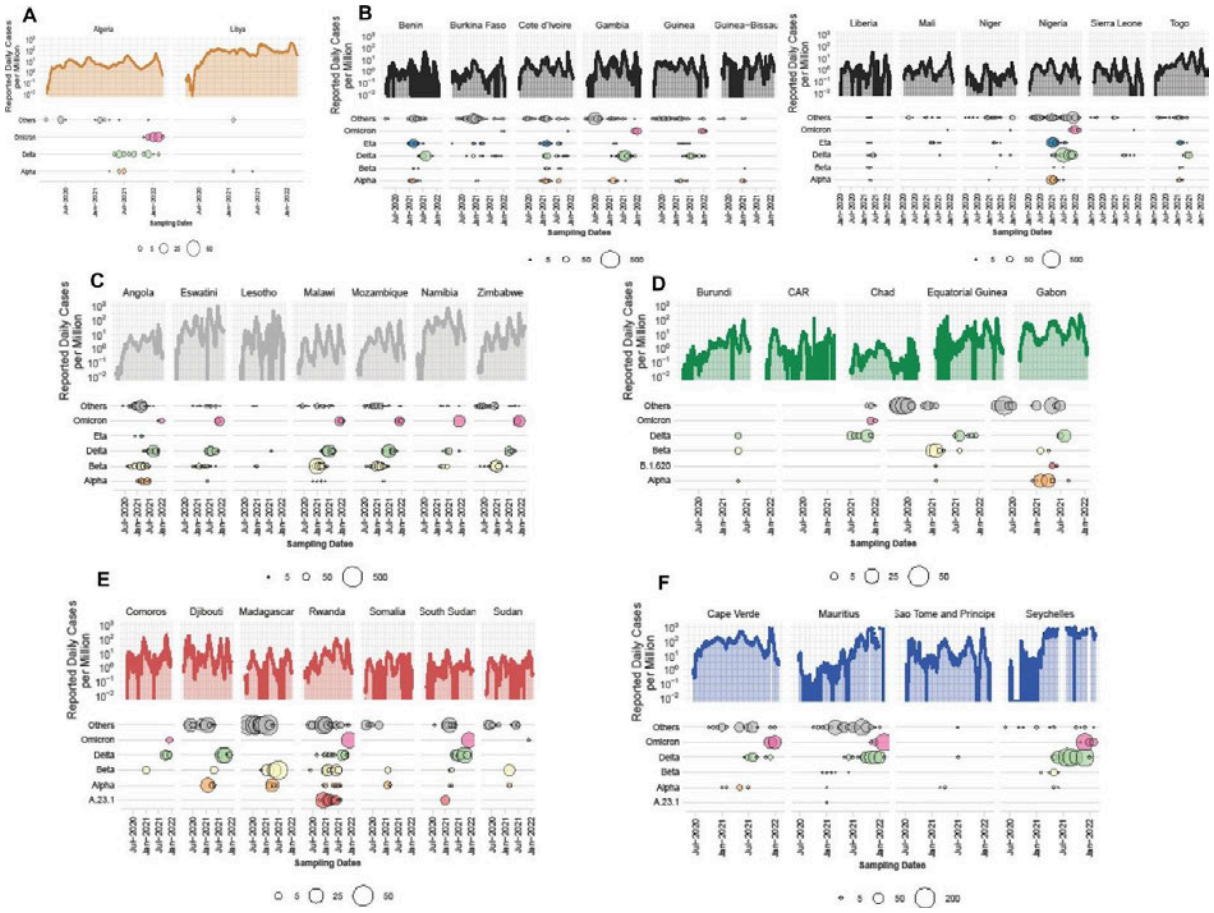
						random selection of these samples were sequenced.
Côte d'Ivoire (Bouaké)	24.30%	Yes	No	No	No	Data set includes all CoV-2 RT-PCR samples tested positive from surveillance in regions of Côte d'Ivoire other than Abidjan; testing at CHU Bouaké; sampling period May-November 2020. Currently generating sequences from samples collected between Dec 2020 and March 2021. Calculation of cases (collumn C): suspected cases: 1199; of those tested: 100%; of those tested positive: 268 (22.36%); of those sequenced: 65

Algeria	0,08%	Yes	Yes	Yes	No	Sequencing of Sample that are negatif for S by rRTPCR test using (taqPath kit from thermo) . Sanger Sequencing of the entire S gene for the confirmation of mutation related to new varriants. WGS for the genomic surveillance using MinION nanopore is in progress.
Mayotte		Yes	Yes	No	No	Random surveillance, with extra samples collections in case of

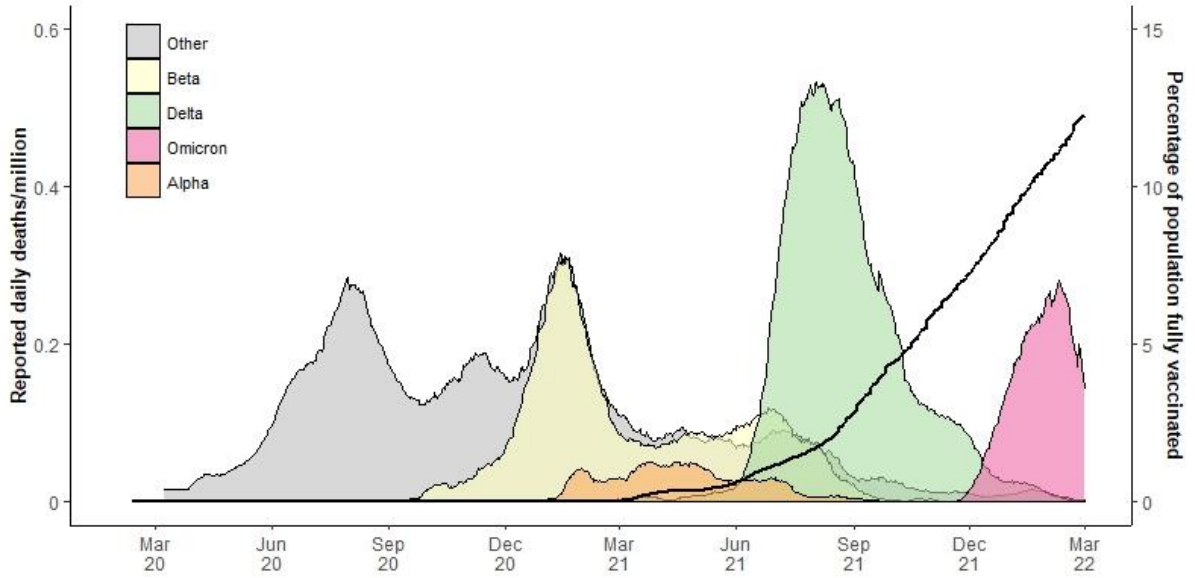
Supplementary Table S4: GISAID Acknowledgements Table supplied as an Excel attachment

Supplementary Materials

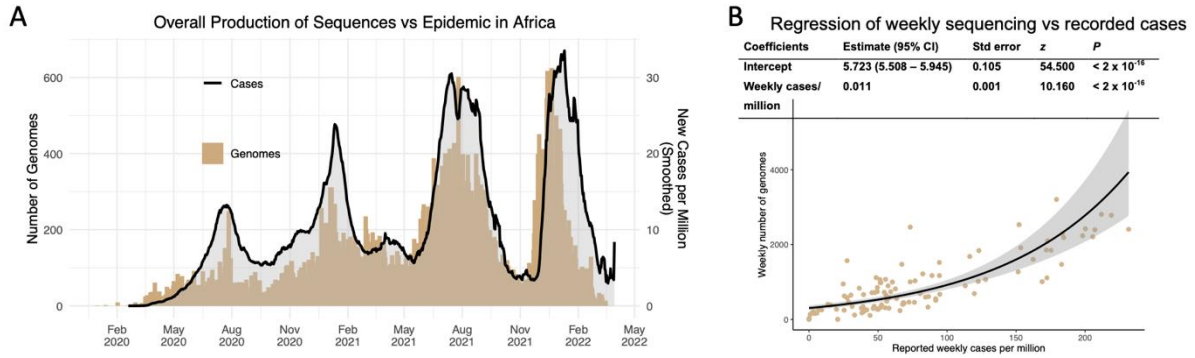
Supplementary Figures



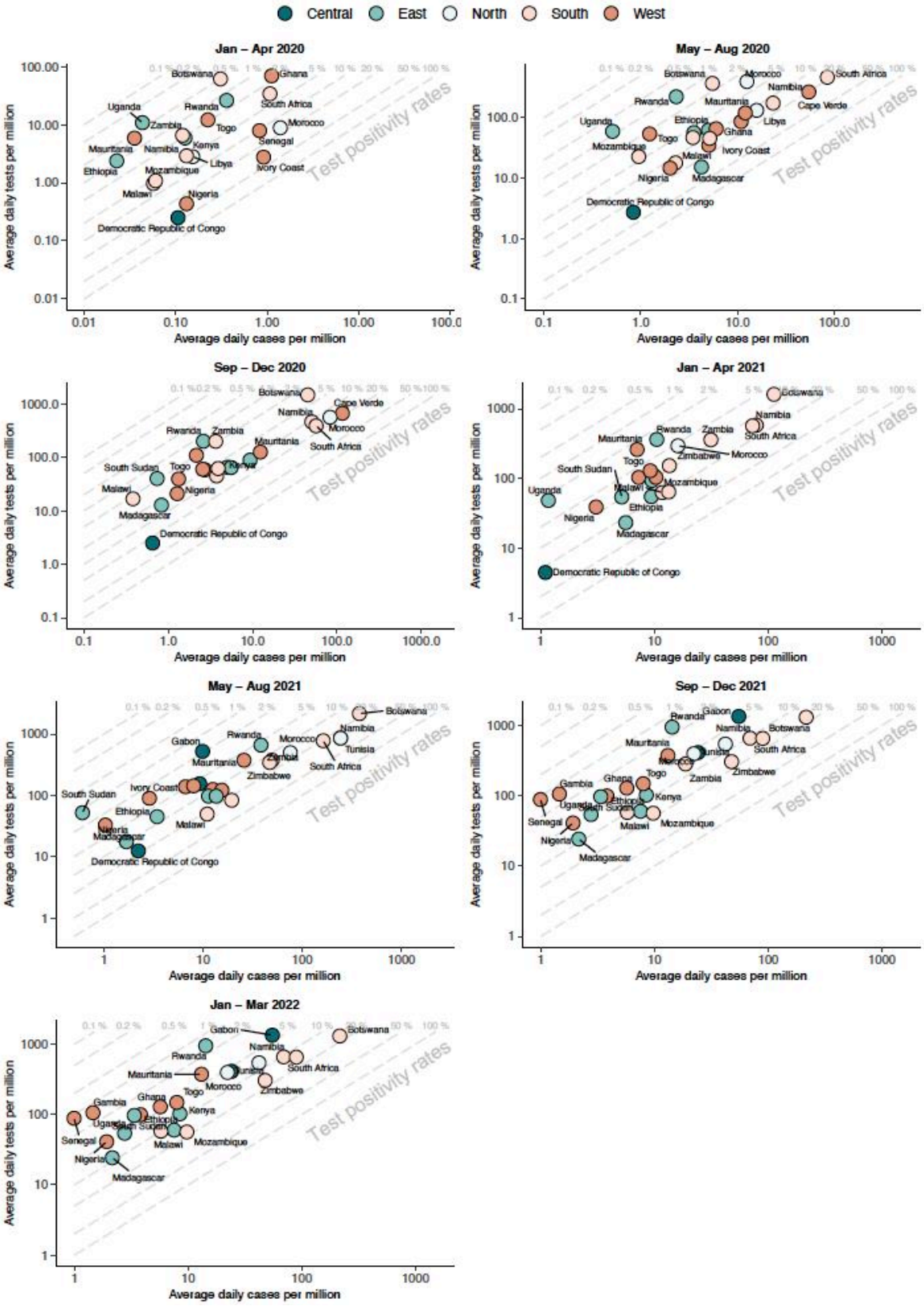
Supplementary Figure S1: Epidemiological progression of the COVID-19 pandemic in all African countries overlaid with the distribution of VOCs, the Eta VOI and other lineages through time (size of circles proportional to the number of genomes sampled per month for each category). The graphs show a breakdown of new cases per million and monthly sampling of VOCs, regional variant or lineage of interest and other lineages for all African countries not shown in Figure 1, grouped by region: A) North Africa, B) West Africa, C) Southern Africa, D) Central Africa, E) East Africa, F) Cape Verde, Mauritius, Sao Tome and Principe and Seychelles, from the beginning of the pandemic to February 2022.



Supplementary Figure S2: Daily reported deaths per million people attributed to each variant of concern with vaccination coverage across Africa. The daily reported deaths were calculated to be attributed to each variant of concern based on the proportion of the variants in genomic surveillance data available from GISAID. We applied an assumption of a 20 day time lag from infection to death (*77*). Alpha caused the lowest peak in reported mortality in Africa, while the largest is attributed to Delta which coincided with the beginning of vaccinations on the continent. The impact of Omicron on mortality is to a much lesser extent than Delta, the peak of Omicron mortality occurred when approximately 12% of the African population was fully vaccinated (completed the initial vaccination protocol of 2 doses for most vaccines, 1 or 3 for some manufacturers).

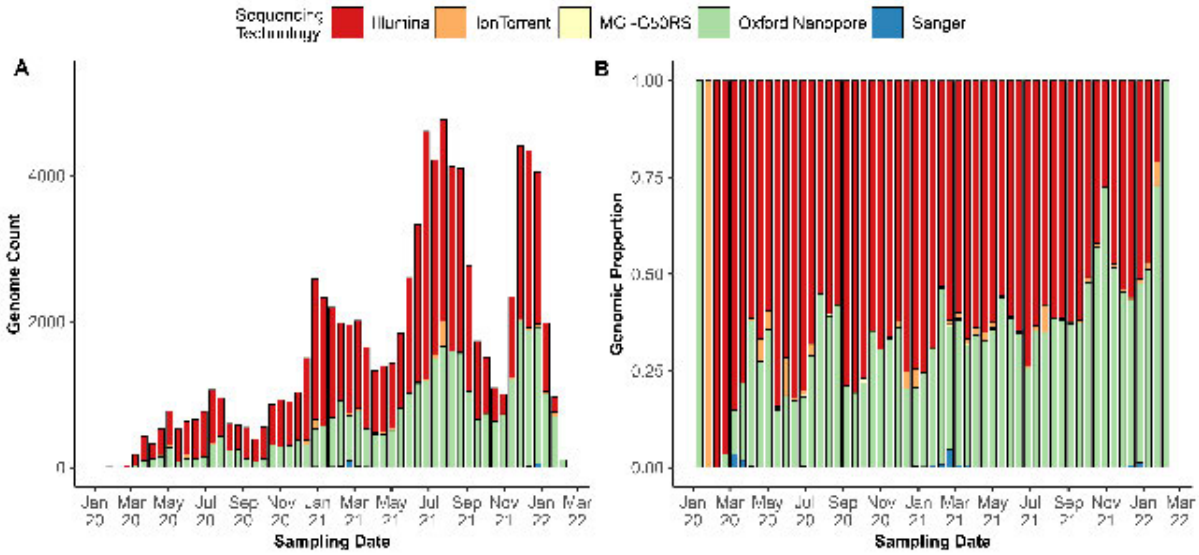


Supplementary Figure S3: Trends of genomic sequencing and epidemic size in Africa. A) Corresponding daily progression of genomic sequence production and epidemic size in Africa. B) Regression of weekly sequencing against recorded cases in Africa. We used a negative binomial regression, with a log link function and maximum likelihood estimation of theta, to investigate the relationship between the number of SARS-CoV-2 genomes produced per week and the weekly number of reported COVID-19 cases in Africa. The regression results indicated a significant positive effect of case numbers on the number of genomes produced, with each one-unit increase in reported weekly cases per million, the expected log count of genomes produced increased by 0.011 ($\theta = 2.23$).

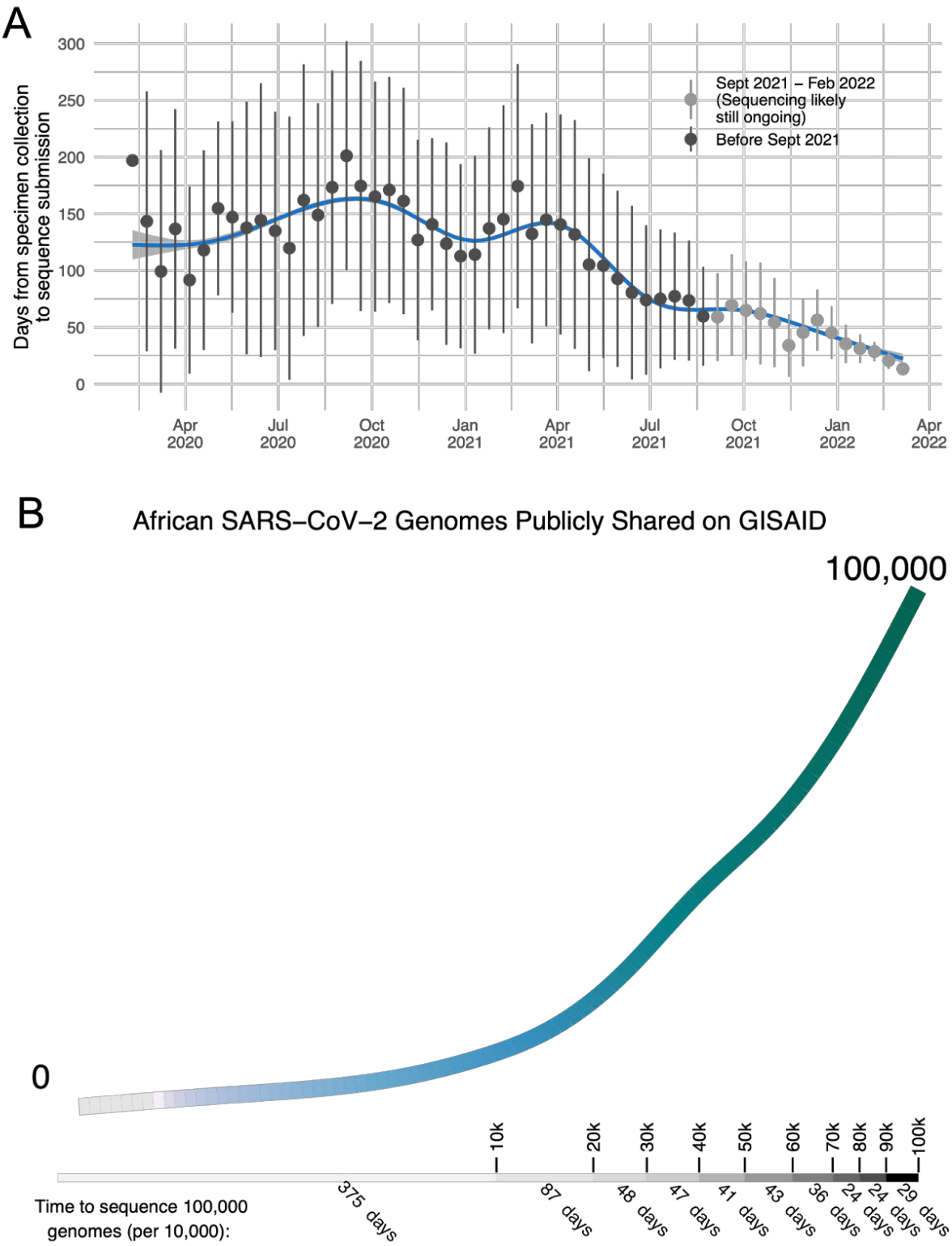


Supplementary Figure S4: COVID-19 testing rates in Africa against recorded cases per million for countries with available data. The extent of testing (average daily number of tests) is shown relative to the size of outbreaks (average daily number of reported cases) per million people per country in Africa. Data is obtained from Our World in Data (OWID). Panels span four-month intervals over the first two years of the epidemic. Test positive rates are displayed in gray segmented lines (0.02 – 100%). Testing efforts of most countries varied considerably

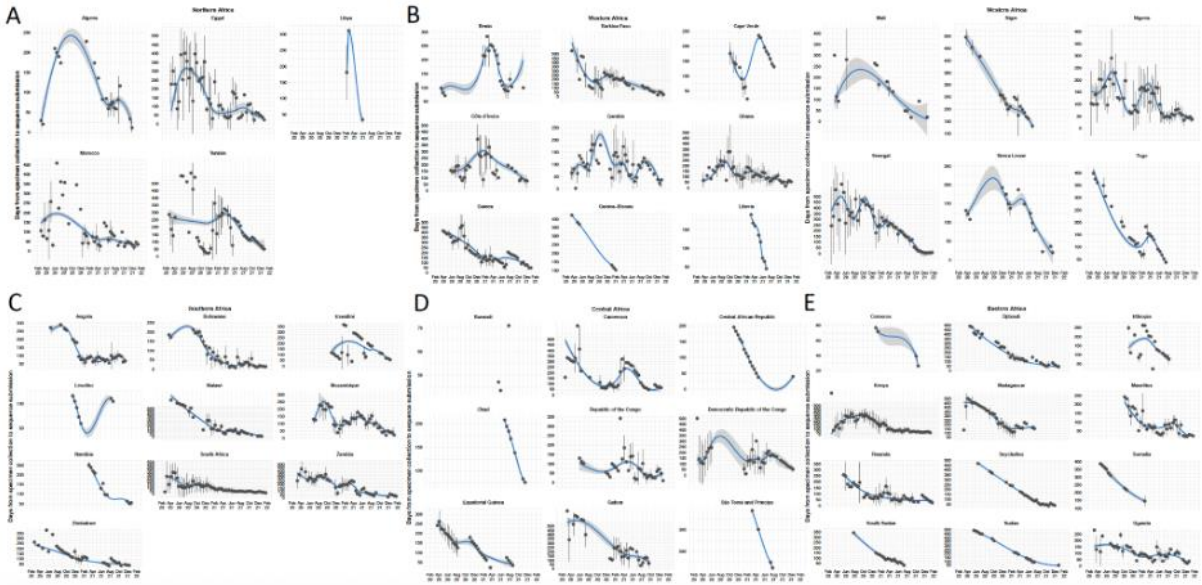
over the course of the epidemic with several southern African countries demonstrating generally the highest positive rate (highest testing rate per case) over the last year of the epidemic. Additionally, the spread of points mostly reduced to between 1 and 20% positive rate from January 2021 to March 2022. Points are shown for countries reporting the relevant data and with an average daily number of tests greater than zero for the relevant period. Axes are log scaled.



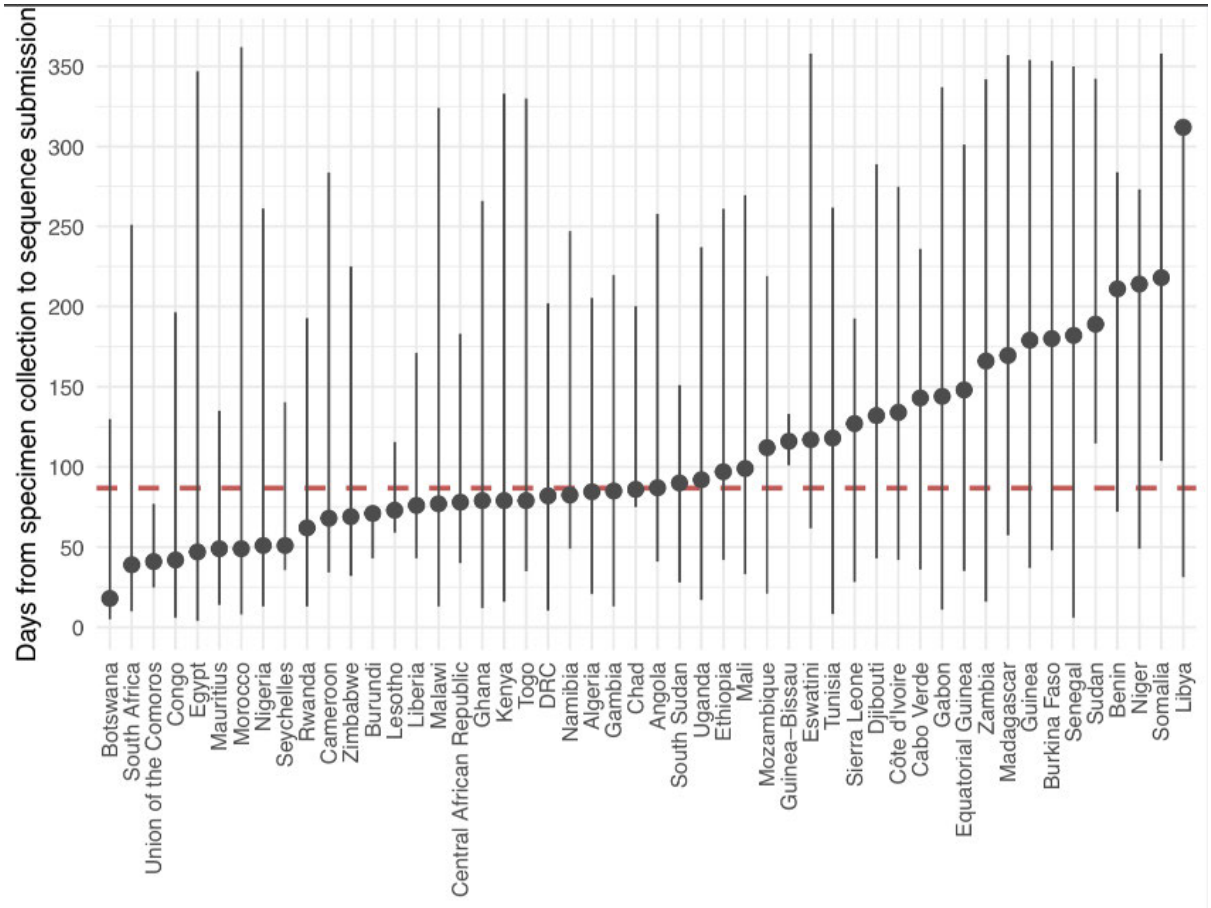
Supplementary Figure S5: The progression of sequencing technologies used for SARS-CoV-2 sequencing in Africa. Sequences are aggregated by associated sequencing technologies as reported on GISAID. Panel A represent the raw number of sequence generated using the five different technologies, while panel B represent the proportion of sequences generated by technology over time.



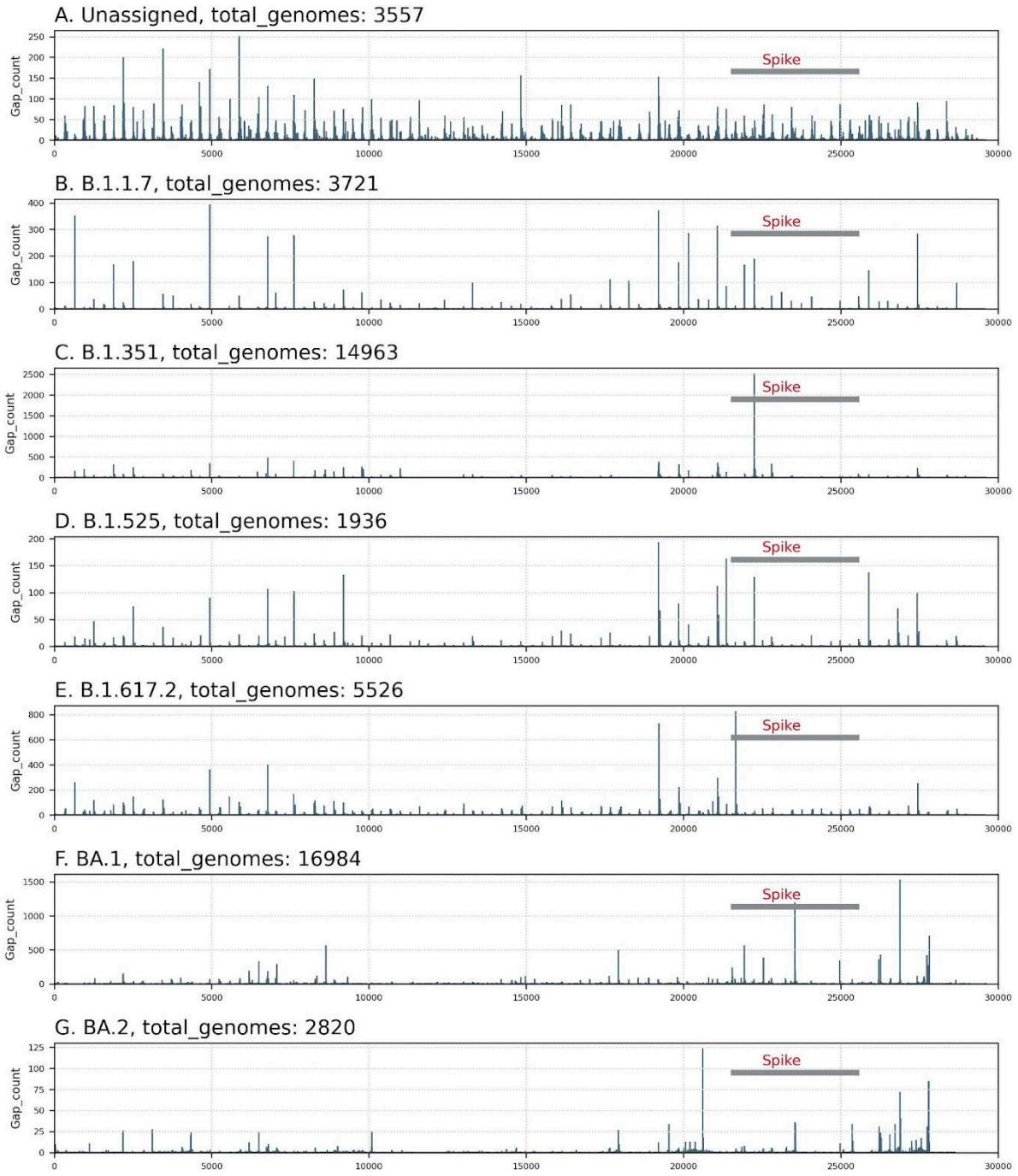
Supplementary Figure S6: Trends of sequencing turnaround time in Africa. A) Overall decreasing trend in sequencing turnaround time in Africa, with the caveat that sequencing for recent samples may not be completed. B) Number of days taken to sequence 10 000 sequences increment in Africa, until the 100 000 sequences milestone.



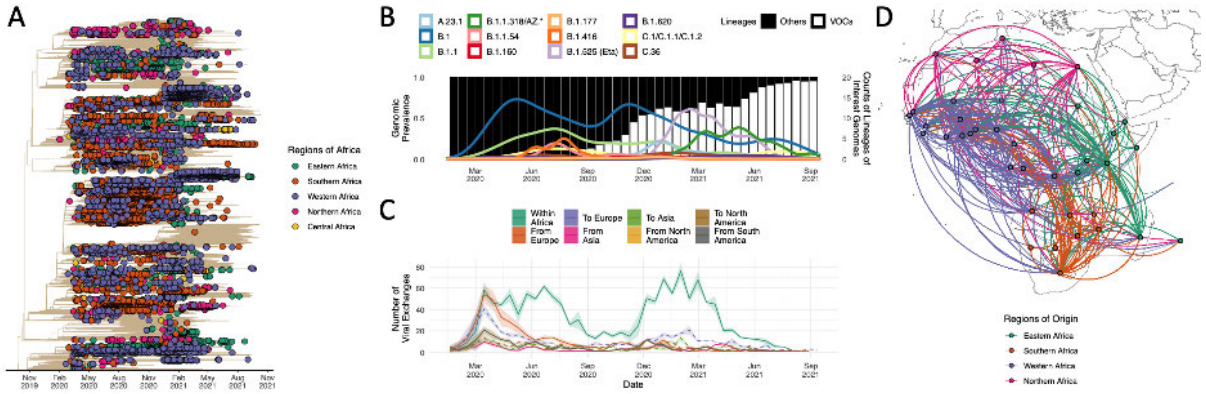
Supplementary Figure S7: Sequencing turnaround time progression per country in each region of Africa. Trends are shown for A) North Africa, B) West Africa, C) Southern Africa, D) Central Africa, and E) East Africa



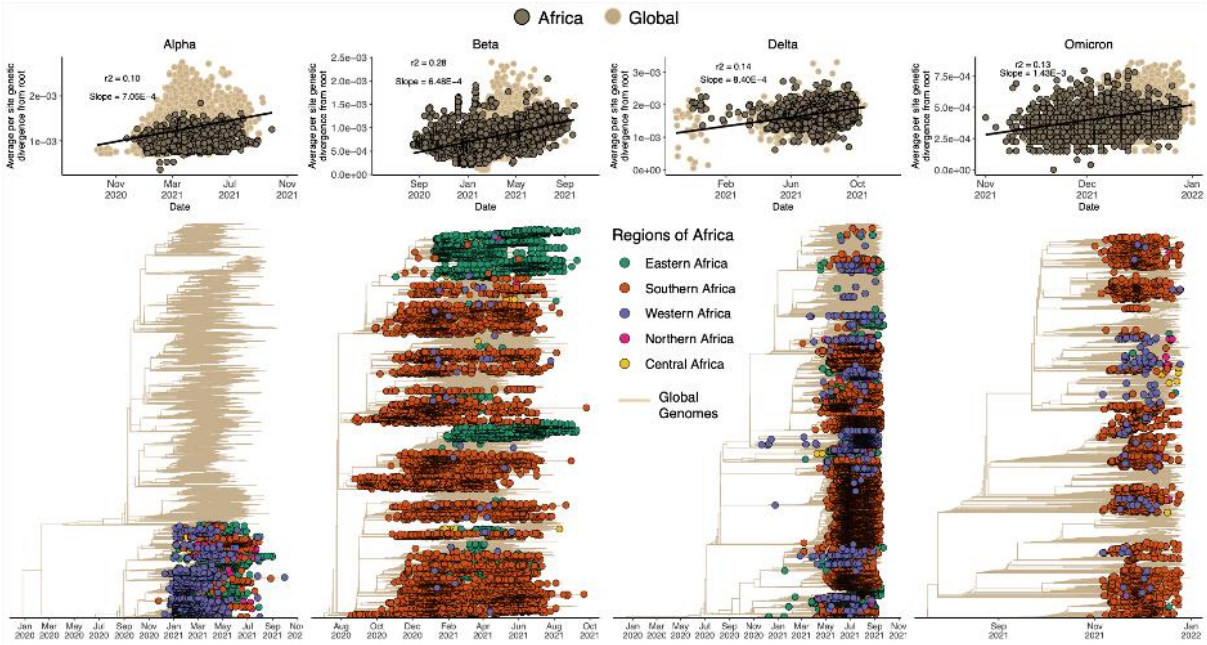
Supplementary Figure S8: Sequencing turnaround time average by country. The circles indicate the mean number of days between specimen collection and sequence submission to GISAID and the bars indicate the distribution of values for each country.



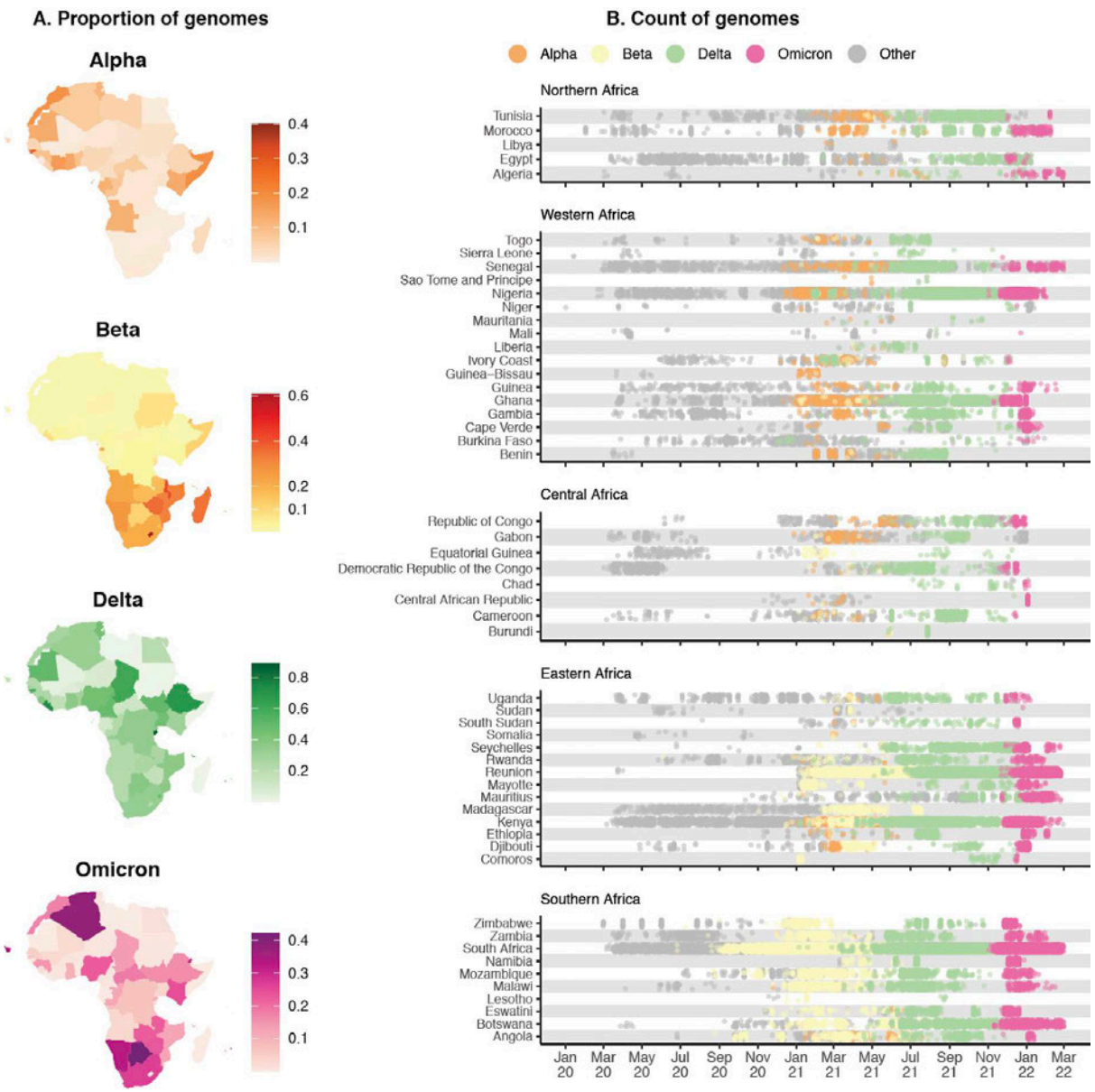
Supplementary Figure S9. Gapped positions in genomes by lineage. The positions of 200 nt N gaps in the 6 major SARS-CoV-2 lineages (plus unassigned entries) were plotted to document lineage-specific patterns in the genome entries. For each lineage (or unassigned entries) the number of 200nt N motifs was plotted by position. The lineage and the total number of genomes classified in the lineage are indicated above each panel. The position of the Spike coding region is indicated with a grey bar. A) Unassigned, B) B.1.1.7 (Alpha), C) B.1.351 (Beta), D) B.1.525 (Eta), E) B.1.617.2 (Delta), F) BA.1 (Omicron), G) BA.2 (Omicron).



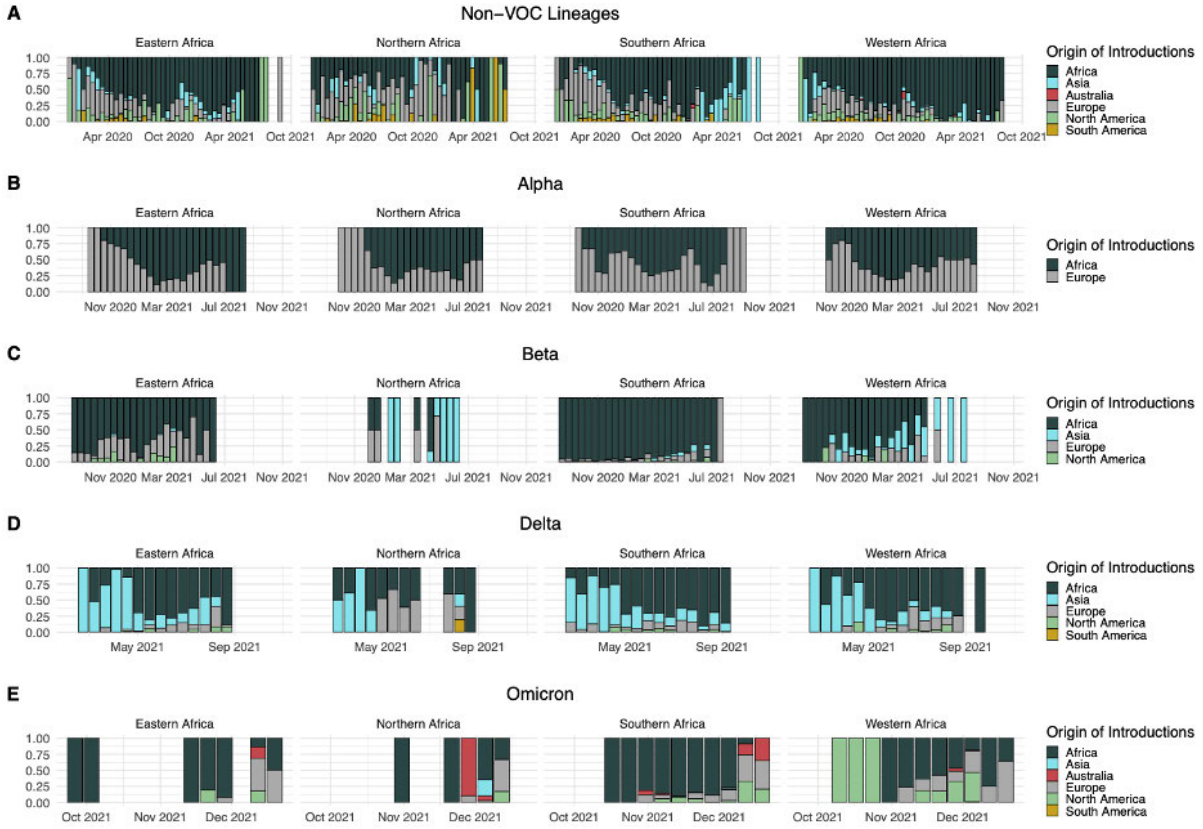
Supplementary Figure S10: Phylogenetic inference of non-VOC lineages in Africa. A) Maximum-Likelihood timetrees of non-VOC genomes in Africa from the beginning of the pandemic till March 2022 against a global reference with African genomes denoted by tippoint circles coloured by regions of Africa. B) Genomic prevalence of non-VOC (black) vs VOC (white) lineages in Africa overlaid by frequency progressions of some lineages of interest in Africa. C) *Inferred viral dissemination patterns of non-VOC lineages to, from and within the Africa continent from the beginning of the pandemic to October 2021. Introductions and viral transitions within Africa are shown in solid lines and exports from Africa are shown in dotted lines and these are coloured by continent. The shaded areas around the lines represent uncertainty of this analysis from ten replicates. D) Dissemination patterns of the non-VOC lineages within Africa, from inferred ancestral state reconstructions, annotated and coloured by region in Africa. The countries of origin of viral exchange routes are also shown with dots and the curves go from country of origin to destination country in an anti-clockwise direction.*



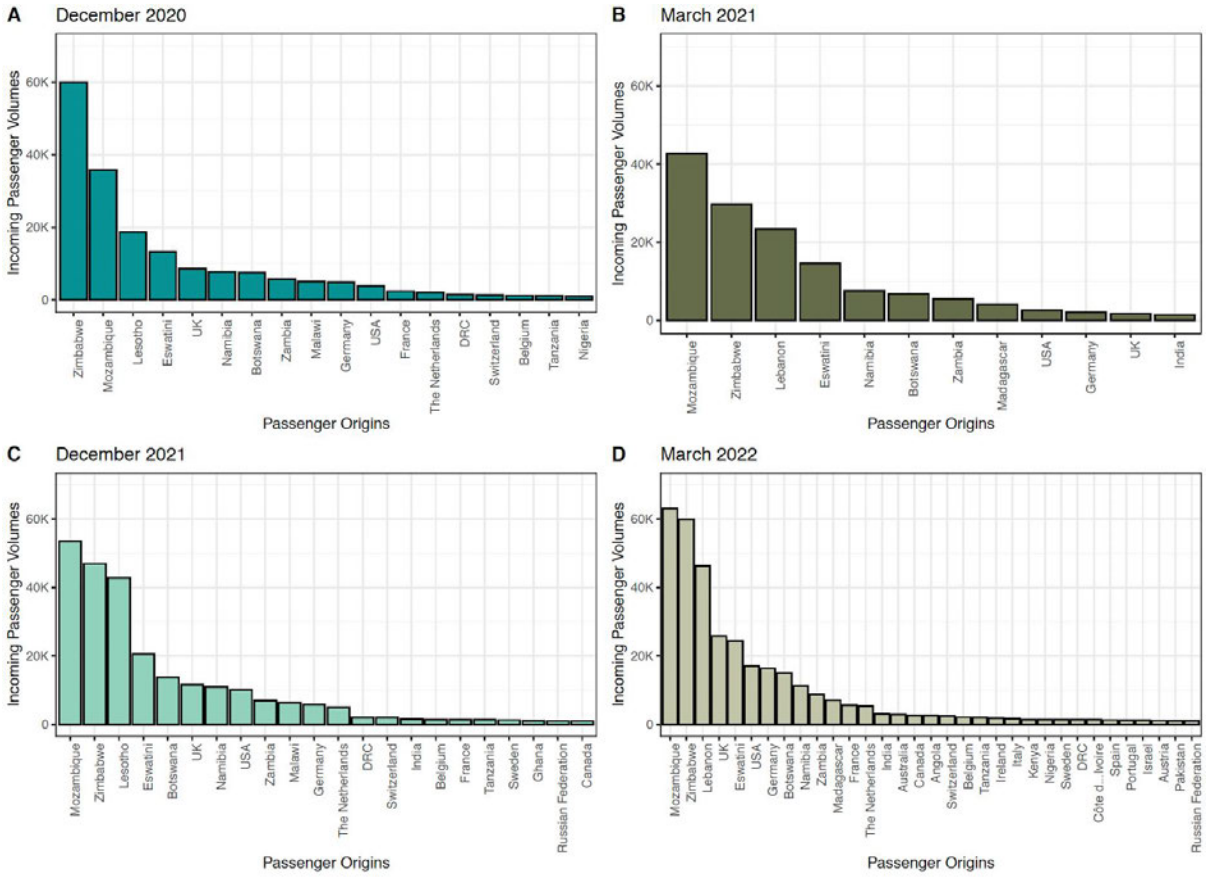
Supplementary Figure S11: Phylogenetic inference of VOCs in Africa using Africa-focused sampling. Top - Molecular clock evolution. Bottom - VOC-specific Maximum-Likelihood timetrees with African genomes denoted by tippoints coloured by regions of Africa.



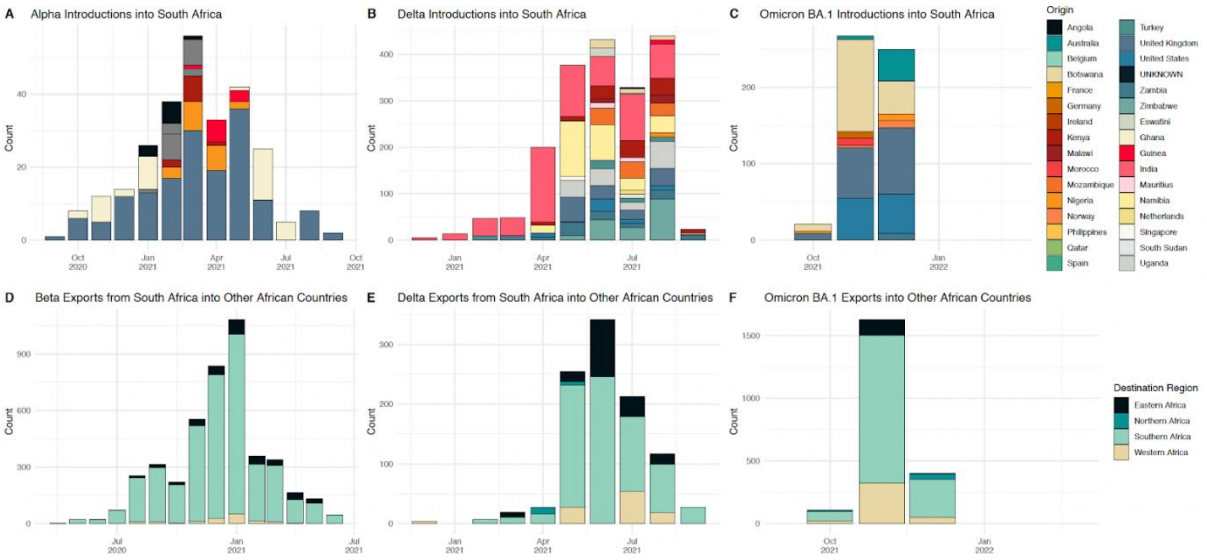
Supplementary Figure S12: Spatial and temporal circulation of four variants of concern (VOC) across Africa. A) The proportion of genomes of each VOC (of all genomes produced) sequenced per country. B) Scatterplot displaying the total number of genomes sequenced per VOC per country from January 2020 to March 2022. Alpha and Beta variants circulated within Africa at similar times in the epidemic however, a greater number of Alpha genomes were sequenced in Northern and Western African countries whereas Beta was more prevalent in Eastern and Southern African countries. Certain Central and Eastern African countries, such as Cameroon and Uganda, display a similar level of co-circulation of both variants. Thereafter, Delta and then Omicron, mostly dominated the landscape in isolation.



Supplementary Figure S13: Patterns of viral importations into different regions of Africa under an African-focused sampling strategy. Proportions of introductions into East, Northern, southern and West Africa attributed to specified origins for A) non-VOC lineages, B) Alpha VOC, C) Beta VOC, D) Delta VOC, and E) Omicron VOC during the relevant time periods are shown.

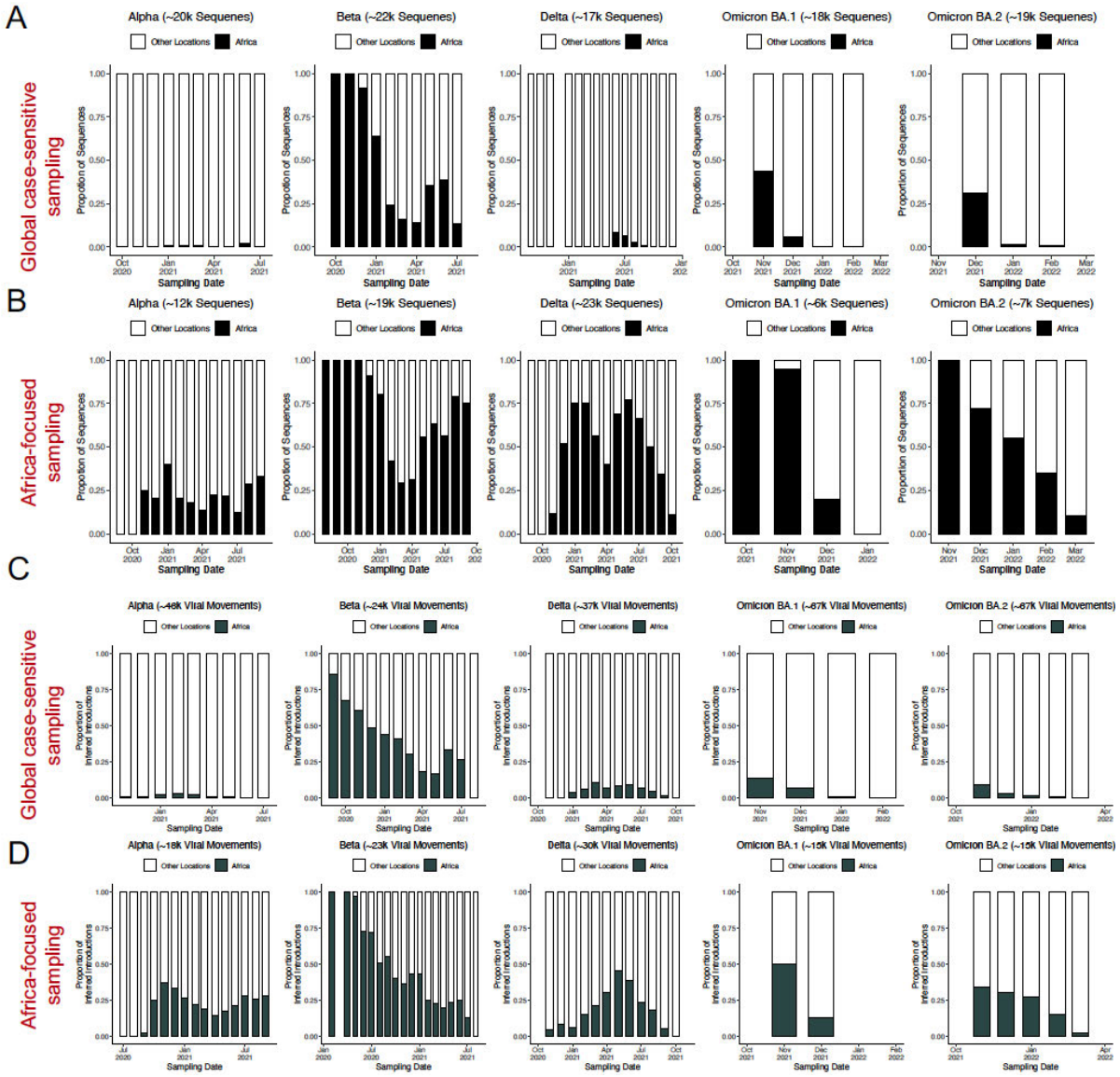


Supplementary Figure S14: Statistics of incoming passenger volume into South Africa by country of origin. Periods shown are December 2020 (A), March 2021 (B), December 2021 (C), and March 2022 (D). Data is obtained from Statistics South Africa Statistical Release P0351 - Tourism and Migration.



Supplementary Figure S15: Inferred introductions into and out of South Africa specified by country of inferred origin. (A) Alpha introductions into South Africa from Africa-focused phylogeography, (B) Delta introductions into South Africa from global case-sensitive phylogeography, (C) Omicron BA.1 introductions into South Africa from global case-sensitive

phylogeography, (D-E) Beta, Delta and Omicron BA.1 exports from South Africa into regions of Africa from Africa-focused phylogeography.



Supplementary Figure S16: Sensitivity analysis for phylogeography from ancestral state reconstruction. (A-B) Sampling proportions of African sequences in two strategies. (C-D) Proportion of inferred introductions into Africa in corresponding two strategies

Supplementary Table S1: NextStrain builds of the five major geographical regions in Africa. Each build focuses on a specific region within Africa and includes sequences from outside the region and the continent to place the regional sequences into context of the global pandemic. All the data used in the builds are publicly available on GISAID and are maintained and updated by the Africa CDC in collaboration with the NextStrain team on a weekly basis.

Region	Regional Builds
Central African Region	https://nextstrain.org/groups/africa-cdc/ncov/central-africa
Eastern African Region	https://nextstrain.org/groups/africa-cdc/ncov/eastern-africa
Northern African Region	https://nextstrain.org/groups/africa-cdc/ncov/northern-africa
Southern African Region	https://nextstrain.org/groups/africa-cdc/ncov/southern-africa
Western African Region	https://nextstrain.org/groups/africa-cdc/ncov/western-africa

Supplementary Table S2: Individual country specific NextStrain builds from the African continent. Each build focuses on a specific country within Africa and includes sequences from the rest of the world in order to place the country's sequences into context of the global pandemic. All the data used in the builds are publicly available on GISAID and are maintained and updated by the Africa CDC in collaboration with the NextStrain team on a weekly basis.

Region	Country	Build
Central African Region	Burundi	https://nextstrain.org/groups/africa-cdc/ncov/burundi
	Cameroon	https://nextstrain.org/groups/africa-cdc/ncov/cameroon
	Central African Republic	https://nextstrain.org/groups/africa-cdc/ncov/central-african-republic
	Chad	https://nextstrain.org/groups/africa-cdc/ncov/chad
	Republic of Congo	https://nextstrain.org/groups/africa-cdc/ncov/republic-of-the-congo
	Democratic Republic of Congo	https://nextstrain.org/groups/africa-cdc/ncov/democratic-republic-of-the-congo
	Equatorial Guinea	https://nextstrain.org/groups/africa-cdc/ncov/equatorial-guinea
	Gabon	https://nextstrain.org/groups/africa-cdc/ncov/gabon
Eastern African Region	Comoros	https://nextstrain.org/groups/africa-cdc/ncov/comoros
	Djibouti	https://nextstrain.org/groups/africa-cdc/ncov/djibouti
	Ethiopia	https://nextstrain.org/groups/africa-cdc/ncov/ethiopia
	Kenya	https://nextstrain.org/groups/africa-cdc/ncov/kenya
	Madagascar	https://nextstrain.org/groups/africa-cdc/ncov/madagascar
	Mauritius	https://nextstrain.org/groups/africa-cdc/ncov/mauritius
	Rwanda	https://nextstrain.org/groups/africa-cdc/ncov/rwanda

	Seychelles	https://nextstrain.org/groups/africa-cdc/ncov/seychelles
	Somalia	https://nextstrain.org/groups/africa-cdc/ncov/somalia
	South Sudan	https://nextstrain.org/groups/africa-cdc/ncov/south-sudan
	Sudan	https://nextstrain.org/groups/africa-cdc/ncov/sudan
	Uganda	https://nextstrain.org/groups/africa-cdc/ncov/uganda
Northern African Region	Algeria	https://nextstrain.org/groups/africa-cdc/ncov/algeria
	Egypt	https://nextstrain.org/groups/africa-cdc/ncov/egypt
	Libya	https://nextstrain.org/groups/africa-cdc/ncov/libya
	Morocco	https://nextstrain.org/groups/africa-cdc/ncov/morocco
	Tunisia	https://nextstrain.org/groups/africa-cdc/ncov/tunisia
Southern African Region	Angola	https://nextstrain.org/groups/africa-cdc/ncov/angola
	Botswana	https://nextstrain.org/groups/africa-cdc/ncov/botswana
	Eswatini	https://nextstrain.org/groups/africa-cdc/ncov/eswatini
	Lesotho	https://nextstrain.org/groups/africa-cdc/ncov/lesotho
	Malawi	https://nextstrain.org/groups/africa-cdc/ncov/malawi
	Mozambique	https://nextstrain.org/groups/africa-cdc/ncov/mozambique
	Namibia	https://nextstrain.org/groups/africa-cdc/ncov/namibia
	South Africa	https://nextstrain.org/groups/africa-cdc/ncov/south-africa
	Zambia	https://nextstrain.org/groups/africa-cdc/ncov/zambia

	Zimbabwe	https://nextstrain.org/groups/africa-cdc/ncov/zimbabwe
Western African Region	Benin	https://nextstrain.org/groups/africa-cdc/ncov/benin
	Burkina Faso	https://nextstrain.org/groups/africa-cdc/ncov/burkina-faso
	Cabo Verde	https://nextstrain.org/groups/africa-cdc/ncov/cabo-verde
	Côte d'Ivoire	https://nextstrain.org/groups/africa-cdc/ncov/cote-divoire
	Gambia	https://nextstrain.org/groups/africa-cdc/ncov/gambia
	Ghana	https://nextstrain.org/groups/africa-cdc/ncov/ghana
	Guinea	https://nextstrain.org/groups/africa-cdc/ncov/guinea
	Guinea-Bissau	https://nextstrain.org/groups/africa-cdc/ncov/guinea-bissau
	Liberia	https://nextstrain.org/groups/africa-cdc/ncov/liberia
	Mali	https://nextstrain.org/groups/africa-cdc/ncov/mali
	Niger	https://nextstrain.org/groups/africa-cdc/ncov/niger
	Nigeria	https://nextstrain.org/groups/africa-cdc/ncov/nigeria
	Senegal	https://nextstrain.org/groups/africa-cdc/ncov/senegal
	Sierra Leone	https://nextstrain.org/groups/africa-cdc/ncov/sierra-leone
	Togo	https://nextstrain.org/groups/africa-cdc/ncov/togo

Supplementary Table S3 (excel file): Sequencing and epidemiological reporting survey results. S3.1) Aggregate results of the survey with responses from 25 countries across the continent. S3.2) frequency of epidemiological reporting (i.e. of new cases and deaths) in the different countries. S3.3) Sequencing strategies employed by different countries as the pandemic progressed. S3.4) The proportion of sequences from the major administrative regions of each country (i.e. provinces, districts or regions). The table can also be found at the github repository (https://github.com/CERI-KRISP/SARS-CoV-2-epidemic-in-Africa).

Supplementary Table S4 (excel file): GISAID acknowledgment table. The table can also be found at the github repository (https://github.com/CERI-KRISP/SARS-CoV-2-epidemic-in-Africa).