nature

https://doi.org/10.1038/s41586-021-03402-9

Accelerated Article Preview

Emergence of a SARS-CoV-2 variant of concern with mutations in spike glycoprotein

Received: 21 December 2020

Accepted: 26 February 2021

Accelerated Article Preview Published online 9 March 2021

Cite this article as: Tegally, H. et al. Emergence of a SARS-CoV-2 variant of concern with mutations in spike glycoprotein. *Nature* https://doi.org/10.1038/s41586-021-03402-9 (2021).

Houriiyah Tegally, Eduan Wilkinson, Marta Giovanetti, Arash Iranzadeh, Vagner Fonseca, Jennifer Giandhari, Deelan Doolabh, Sureshnee Pillay, Emmanuel James San, Nokukhanya Msomi, Koleka Mlisana, Anne von Gottberg, Sibongile Walaza, Mushal Allam, Arshad Ismail, Thabo Mohale, Allison J. Glass, Susan Engelbrecht, Gert Van Zyl, Wolfgang Preiser, Francesco Petruccione, Alex Sigal, Diana Hardie, Gert Marais, Marvin Hsiao, Stephen Korsman, Mary-Ann Davies, Lynn Tyers, Innocent Mudau, Denis York, Caroline Maslo, Dominique Goedhals, Shareef Abrahams, Oluwakemi Laguda-Akingba, Arghavan Alisoltani-Dehkordi, Adam Godzik, Constantinos Kurt Wibmer, Bryan Trevor Sewell, José Lourenço, Luiz Carlos Junior Alcantara, Sergei L. Kosakovsky Pond, Steven Weaver, Darren Martin, Richard J. Lessells, Jinal N. Bhiman, Carolyn Williamson & Tulio de Oliveira

This is a PDF file of a peer-reviewed paper that has been accepted for publication. Although unedited, the content has been subjected to preliminary formatting. Nature is providing this early version of the typeset paper as a service to our authors and readers. The text and figures will undergo copyediting and a proof review before the paper is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers apply.

ACCELERATED ARTICLE PREVIEW

Article

Emergence of a SARS-CoV-2 variant of concern with mutations in spike glycoprotein

https://doi.org/10.1038/s41586-021-03402-9

Received: 21 December 2020

Accepted: 26 February 2021

Published online: 9 March 2021

Houriiyah Tegally[1,33], Eduan Wilkinson[1,33], Marta Giovanetti[2,3,33], Arash Iranzadeh[4,33], 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[9,10,34], Carolyn Williamson[5,8,19,34] & Tulio de Oliveira[1,8,32,34] ✉

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[1,2]. 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[3–5]. This lineage was identified 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[6–8].

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[9,10].

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

[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. [33]These authors contributed equally: Houriiyah Tegally, Eduan Wilkinson, Marta Giovanetti, Arash Iranzadeh. [34]These authors jointly supervised this work: Jinal N Bhiman, Carolyn Williamson, Tulio de Oliveira. ✉e-mail: tuliodna@gmail.com

ACCEPTED ARTICLE PREVIEW

this mutation have increased transmissibility[11–14]. 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[15].

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)[16]. 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[17,18]. 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.

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. 1a)[19]. 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. 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 (Extended Data Fig 1). 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. 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 (Extended Data Fig 2). The resurgence of the daily case counts at an exponential rate happened later for WC and KZN than for EC (Fig. 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.

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 (Extended Data Fig 3). 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. 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. 2b). Seven South African sequences are basal to the 501Y.V2 cluster (Fig. 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 has 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. 2a)[18]. 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. 2c, Extended Data Fig 4).

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. 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.

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. 2b, Fig. 3a, Extended Data Fig 5). 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 different laboratories across the network. This is a difficult to align region, and the deletion could potentially also be located at position 241-243, but the resulting sequence would be 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. 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. 3c). The main South African 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) and did not show the rapid accumulation of mutations as observed with 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 (Extended Data Fig 6). 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. 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 (Extended Data Fig 7).

Selection analysis

We examined patterns of nucleotide variations and fluctuations in mutant frequencies at eight polymorphic spike gene sites (Fig. 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

ACCELERATED ARTICLE PREVIEW

> 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 that this variant has a selective advantage, either from increased transmissibility, immune escape, or both. These data highlight the urgent need to refocus the public health response in South Africa on driving transmission down to low levels, not only to reduce hospitalisations and deaths but to limit the spread of this lineage and the further evolution of the virus.

We detected this new variant through intensified genomic surveillance in response to a rapid resurgence of cases in the Eastern Cape Province[20]. However, both before and after the detection of 501Y.V2, our genomic surveillance involved regular sequencing of a random selection of residual samples from routine diagnostic services. We show that 501Y.V2 was detected in samples from 197 different health facilities in multiple districts across four provinces. Therefore we are confident that whilst our sequencing coverage is relatively low, the sequences are representative of the circulating viruses in these provinces. Although the epidemic in the Eastern Cape was contracting in mid-July to mid-August (the estimated tMRCA), this was not really a period of low transmission. Incidence was still above 20 per 100 000 per week at this time and the test positivity rate remained above 10%, suggesting moderate-to-high levels of transmission. As there were many different lineages circulating at this time, the rapid expansion of 501Y.V2 and almost complete displacement of other lineages in multiple regions strongly suggests a selective advantage for this variant.

Preliminary modelling suggests that the 501Y.V2 could be approximately 50% more transmissible than the previously circulating variants, although this estimate assumes that natural immunity confers complete protection against re-infection[6]. Increased transmissibility is plausible given what we know about the spike mutations in 501Y.V2, and what we are learning about similar SARS-CoV-2 variants emerging in other locations. This new lineage has three mutations at key sites in the RBD (K417N, E484K and N501Y). The N501Y mutation was also recently identified in a new lineage that has spread rapidly in the United Kingdom (B.1.1.7)[21]. There is now good evidence that this lineage is associated with increased transmissibility[22]. The N501Y mutation has been shown through deep mutation scanning and in a mouse model to enhance binding affinity to hACE2[3,23]. There is some evidence that the E484K mutation may also increase binding affinity to hACE2[3]; and that the combination of N501Y and E484K particularly enhances binding affinity[24,25]. Further work is being conducted to understand the precise mechanisms underlying the increased transmissibility of these new variants.

The other reason for a selective advantage of 501Y.V2 could be immune escape, i.e. the capacity to cause re-infection. 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[26]. 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. The spike RBD is the main target of neutralizing antibodies (NAbs) elicited during SARS-CoV-2 infection[27]. NAbs to the RBD can be broadly divided into four main classes[28]. 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[27]. 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[4,5,29–31]. Aside from the RBD, the remaining neutralizing activity is targeted at the N-terminal domain, and some of the NTD mutations in 501Y.V2 are located within or near residues that form an antigenic supersite[32,33]. We now have preliminary evidence, from live virus and pseudovirus experiments, that 501Y.V2 shows substantial or complete escape from neutralizing antibodies in convalescent plasma[7,8]. We are currently investigating the frequency of re-infection in the second wave, and the clinical presentations of those with re-infection to understand better the clinical and epidemiological impact of any immune escape. We are also conducting neutralization assays on plasma from vaccine recipients, and await results of vaccine efficacy trials conducted in South Africa during the expansion of 501Y.V2.

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[34–36]. This hypothesis is supported by the long branch length connecting the lineage to the remaining sequences in our phylogenetic tree (Extended Data Fig 8). 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[34]. 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[37]. 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 emerged independently around the world (Extended Data Fig 9) and been found in circulating lineages together.

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.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information,

ACCELERATED ARTICLE PREVIEW

acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-021-03402-9.

1. Fontanet A, Autran B, Lina B, Kieny MP, Karim SSA, Sridhar D. SARS-CoV-2 variants and ending the COVID-19 pandemic. *Lancet*. 2021.
2. Mascola JR, Graham BS, Fauci AS. SARS-CoV-2 Viral Variants-Tackling a Moving Target. *JAMA*. 2021.
3. 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.
4. Greaney AJ, Loes AN, Crawford KHD, 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*. 2021.
5. 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*. 2021;**29**(1):44-57 e49.
6. Pearson CAB, Russell TW, Davies N, et al. Estimates of severity and transmissibility of novel South Africa SARS-CoV-2 variant 501Y.V2. 2021.
7. Cele S, Gazy I, Jackson L, et al. Escape of SARS-CoV-2 501Y.V2 variants from neutralization by convalescent plasma. *medRxiv*. 2021:2021.2001.2026.21250224.
8. Wibmer CK, Ayres F, Hermanus T, et al. SARS-CoV-2 501Y.V2 escapes neutralization by South African COVID-19 donor plasma. *bioRxiv*. 2021:2021.2001.2018.427166.
9. Hadfield J, Megill C, Bell SM, et al. Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics*. 2018;**34**(23):4121-4123.
10. Shu Y, McCauley J. GISAID: Global initiative on sharing all influenza data - from vision to reality. *Euro Surveill*. 2017;**22**(13):30494.
11. 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.
12. Plante JA, Liu Y, Liu J, et al. Spike mutation D614G alters SARS-CoV-2 fitness. *Nature*. 2020.
13. 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.
14. Volz E, Hill V, McCrone JT, et al. Evaluating the Effects of SARS-CoV-2 Spike Mutation D614G on Transmissibility and Pathogenicity. *Cell*. 2021;**184**(1):64-75 e11.
15. Thomson EC, Rosen LE, Shepherd JG, et al. Circulating SARS-CoV-2 spike N439K variants maintain fitness while evading antibody-mediated immunity. *Cell*. 2021.
16. Bradshaw D, Laubscher R, Dorrington R, Groenewald P, Moultrie T. *Report on Weekly Deaths in South Africa: 1 January - 8 December 2020 (Week 49)*. 2020.
17. 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*. 2021;**103**:234-241.
18. Tegally H, Wilkinson E, Lessells RJ, et al. Sixteen novel lineages of SARS-CoV-2 in South Africa. *Nat Med*. 2021.
19. 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.
20. 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.
21. 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.
22. Volz E, Mishra S, Chand M, et al. Transmission of SARS-CoV-2 Lineage B.1.1.7 in England: Insights from linking epidemiological and genetic data. *medRxiv*. 2021:2020.2012.2030.20249034.
23. 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.
24. Nelson G, Buzko O, Spilman P, Niazi K, Rabizadeh S, Soon-Shiong P. Molecular dynamic simulation reveals E484K mutation enhances spike RBD-ACE2 affinity and the combination of E484K, K417N and N501Y mutations (501Y.V2 variant) induces conformational change greater than N501Y mutant alone, potentially resulting in an escape mutant. *bioRxiv*. 2021:2021.2001.2013.426558.
25. Zahradník J, Marciano S, Shemesh M, et al. SARS-CoV-2 RBD in vitro evolution follows contagious mutation spread, yet generates an able infection inhibitor. *bioRxiv*. 2021:2021.2001.2006.425392.
26. 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.
27. 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.
28. Barnes CO, Jette CA, Abernathy ME, et al. SARS-CoV-2 neutralizing antibody structures inform therapeutic strategies. *Nature*. 2020;**588**(7839):682-687.
29. 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.
30. Liu Z, VanBlargan LA, Bloyet LM, et al. Identification of SARS-CoV-2 spike mutations that attenuate monoclonal and serum antibody neutralization. *Cell Host Microbe*. 2021.
31. Weisblum Y, Schmidt F, Zhang F, et al. Escape from neutralizing antibodies by SARS-CoV-2 spike protein variants. *Elife*. 2020;**9**.
32. Chi X, Yan R, Zhang J, et al. A neutralizing human antibody binds to the N-terminal domain of the Spike protein of SARS-CoV-2. *Science*. 2020;**369**(6504):650-655.
33. McCallum M, Marco A, Lempp F, et al. N-terminal domain antigenic mapping reveals a site of vulnerability for SARS-CoV-2. *bioRxiv*. 2021:2021.2001.2014.426475.
34. 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.
35. 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;**183**(7):1901-1912 e1909.
36. Kemp SA, Collier DA, Datir RP, et al. SARS-CoV-2 evolution during treatment of chronic infection. *Nature*. 2021.
37. 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.

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

© The Author(s), under exclusive licence to Springer Nature Limited 2021

ACCELERATED ARTICLE PREVIEW

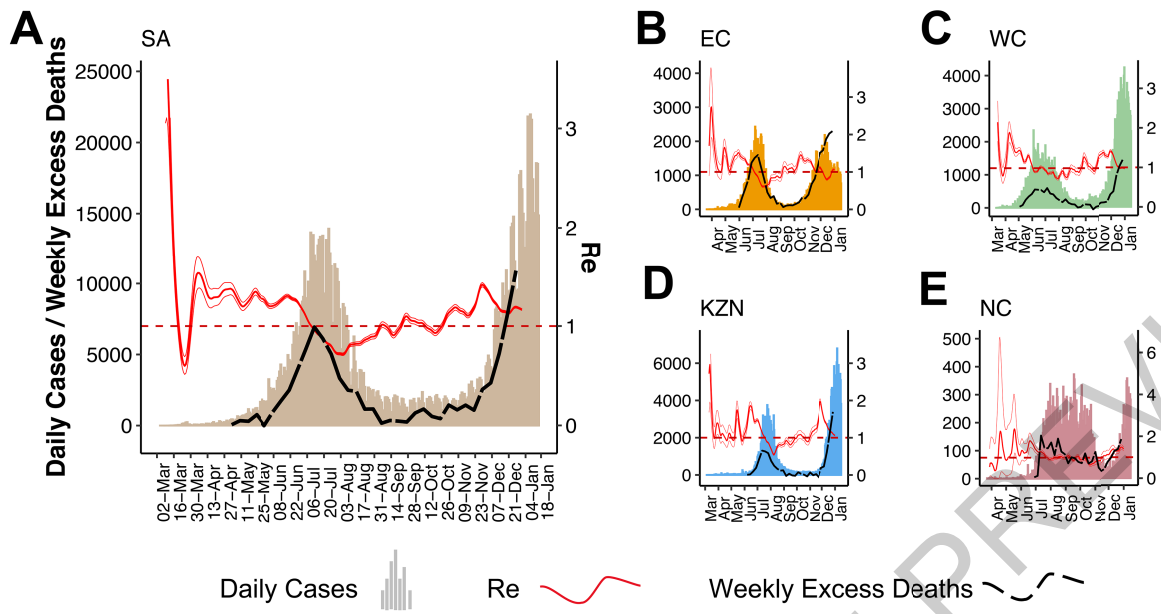


Fig. 1 | SARS-CoV-2 epidemiological dynamics in South-Africa. The histograms show the number of daily confirmed COVID-19 cases (mapped to left y-axis) in South Africa (A), and the four provinces under study, Eastern Cape (B), Western Cape (C), KwaZulu-Natal (D) and Northern Cape (E). Fluctuations to 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).

ACCELERATED ARTICLE PREVIEW

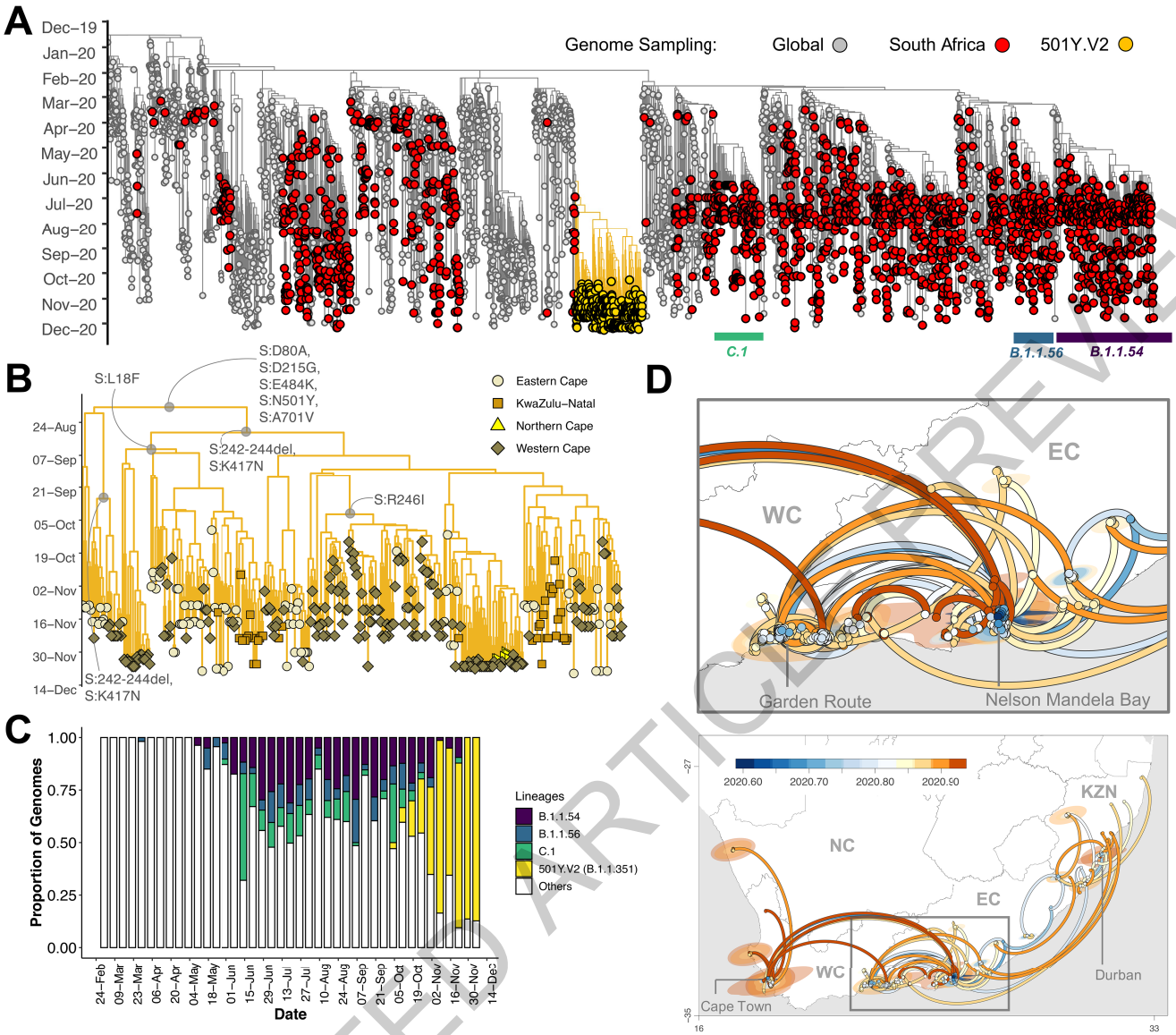


Fig. 2 | Evolution and spread of the 501Y.V2 cluster in South Africa. A) Time-resolved maximum clade credibility phylogeny of 5239 SARS-CoV-2 sequences, 2756 of which are from South Africa (red). The new SARS-CoV-2 cluster is highlighted in yellow. B) Time-resolved maximum clade credibility phylogeny 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).

ACCELERATED ARTICLE PREVIEW

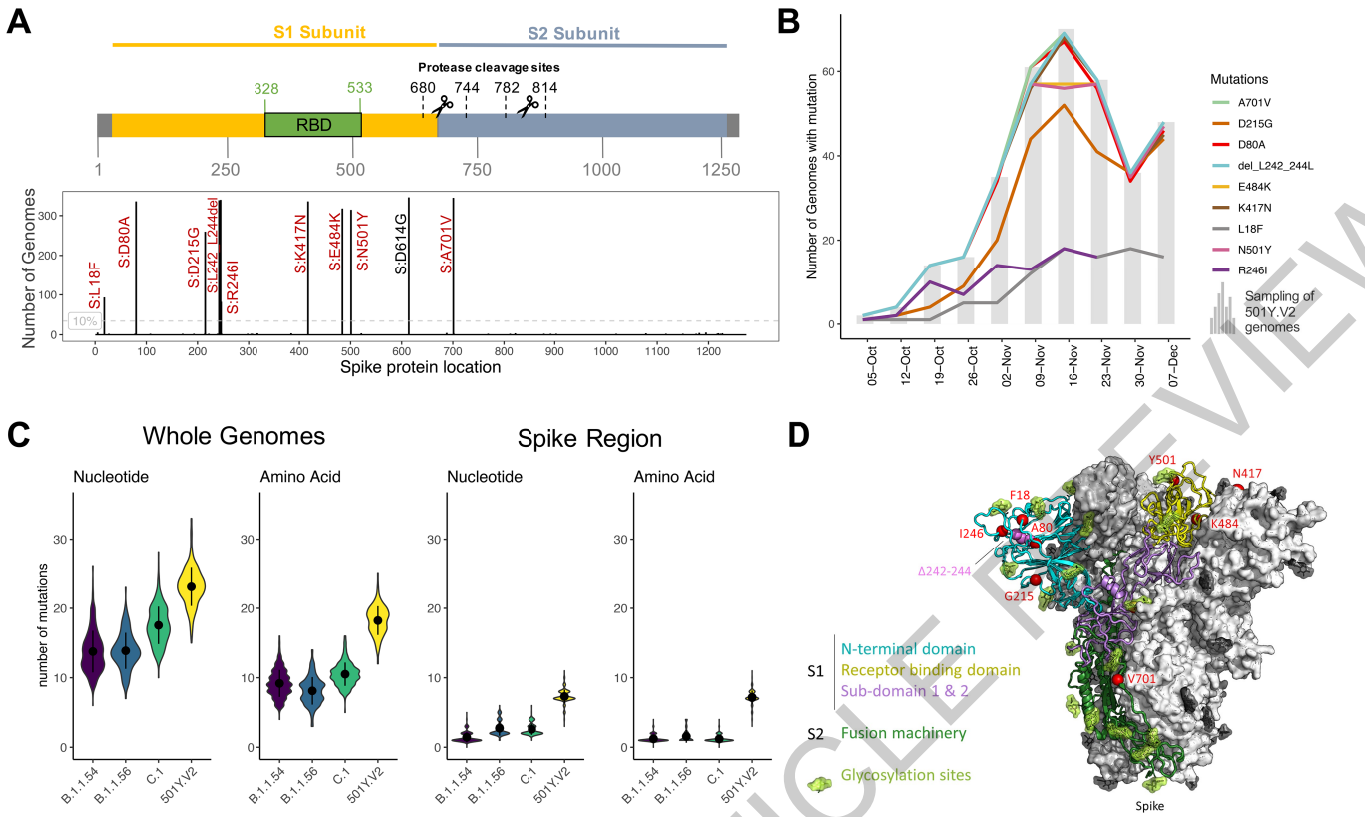


Fig. 3 | Mutational profile of the 501Y.V2 spike region. A) Amino acid changes in the spike region of the 341 501Y.V2 genomes in this study mapped to the spike protein sequences structure, indicating key regions, such as the RBD. Each spike protein variant is shown at their respective protein locations, 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 (n=341), compared to lineages B.1.1.54 (n=472), B.1.1.56 (n=179), and C.1 (n=271), three major lineages circulating in South Africa during the first wave. The dot and error bars inside each group denote the mean and range for two standard deviations respectively. 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.

ACCELERATED ARTICLE PREVIEW

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)[38,39]. 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[40]. 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[41].

Sampling of SARS-CoV-2

As part of the Network for Genomic Surveillance in South Africa (NGS-SA)[42], 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 (Extended Data Fig 10).

Ethical Statement

The project was 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). 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

cDNA synthesis was performed on the extracted RNA using random primers followed by gene specific multiplex PCR using the ARTIC V3 protocol[43]. 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. For nanopore sequencing, we adapted the nCoV-2019 sequencing LoCost protocol v3[43]. Briefly, PCR reactions were done in 12,5 µl volumes and no PCR product purification was done. After DNA repair (NEB, USA) and end-prep reactions (NEB, USA), up to 24 samples were barcoded by ligation (EXP-NBD104 / NBD114, Oxford Nanopore Technologies, Oxford, UK). Barcoded samples were pooled, bead-purified and ligated to sequence adapters. After the bead clean-up, the DNA concentration was determined with a Qubit 2.0 instrument (Thermo Fisher, USA). Up to 50 ng of the library in 75 µl were loaded on a prepared R9.4.1 flow-cell. A GridION X5 sequencing run was initiated using MinKNOW software with the high accuracy base-call setting. The NC045512 reference was used for alignment during base-calling and the barcodes were split into different folders. Fastq files were downloaded from the GridION X5 for assembly and further analysis.

For Illumina sequencing, 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[44,45].

We assembled paired-end and nanopore fastq reads using Genome Detective 1.132 (https://www.genomedetective.com) and the Coronavirus Typing Tool[46]. For short reads, 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[47] against the reference rather than the de novo consensus sequence, and subsequently final mutations and indels are called using GATK HaplotypeCaller[48], with low quality variants (with QD < 10) filtered using GATK VariantFiltration[48]. 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). For nanopore data, candidate reads are assigned to candidate reference sequences using NCBI blastn with sensitive settings and low gap costs. Candidate reads are then aligned using AGA (Annotated Genome Aligner), after which a draft majority consensus sequence is subsequently called, and iteratively improved by realignment of all reads against the draft consensus sequence and realignment of regions with a putative insert against the reference using global alignment (MAFFT). The resulting consensus sequence is further polished by considering and correcting indels of length one or two in homopolymer regions of length 4 or longer that break the open reading frame (likely sequencing errors). Mutations were confirmed visually with bam files using Geneious software V2020.1.2 (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 previously 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). In addition, we have implemented a Sanger sequencing method that covers the main RBD sites and this was used to confirm the K417N and other mutations (i.e. E484K and N501Y) in sequences in which we were not confident about the call from NGS data. The full sequence properties, mutation and spike mutations of the 501Y.V2 sequences are shown in Supplementary Tables S3 and S4.

LoFreq was used to detect minor viral variants to study the intra-host heterogeneity of viral variants (quasi-species)[49] (Extended Data Fig 5).

ACCELERATED ARTICLE PREVIEW

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 4 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 (https://github.com/nextstrain/ncov)[50]. The pipeline contains several python scripts that manage the analysis workflow. It performs alignment of genomes in MAFFT[51], phylogenetic tree inference in IQ-Tree V1.6.9[52], 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) (Extended Data Fig 6).

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[53]. 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[54]. Maximum clade credibility trees were summarised from the MCMC samples using TreeAnnotator after discarding 10% as burn-in. The phylogenetic trees were visualized using ggplot and ggtree[55,56].

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[57]. 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[58].

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)[59]. 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)[60], (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)[61], 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[62]. We used the Pymol program (The PyMOL Molecular Graphics System, Version 2.2.0, Schrödinger, LLC.) for visualization.

Reporting summary

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

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

All custom 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.

ACCELERATED ARTICLE PREVIEW

38. 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.
39. Marivate V, de Waal A, Combrink HM, et al. Coronavirus disease (COVID-19) case data - South Africa. https://zenodo.org/record/3732419#.YBXzFegzZPY. Published 2020. Accessed.
40. Bradshaw D, Laubscher R, Dorrington R, Groenewald P, Moultrie T. *Report on Weekly Deaths in South Africa: 1 January - 8 December 2020 (Week 49)*. 2020.
41. 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.
42. 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.
43. Quick J. nCoV-2019 sequencing protocol v3 (LoCost) V.3. https://www.protocols.io/view/ncov-2019-sequencing-protocol-v3-locost-bh42j8ye. Published 2020. Accessed.
44. Giandhari J, Pillay S, Tegally H, et al. NEBnext library construction and sequencing for SARS-CoV-2: Adapting COVID-19 ARTIC protocol. https://www.protocols.io/view/nebnext-library-construction-and-sequencing-for-sa-bhu2j6ye. Published 2020. Accessed.
45. 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).
46. 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.
47. Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*. 2018;**34**(18):3094-3100.
48. 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.
49. 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.
50. Hadfield J, Megill C, Bell SM, et al. Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics*. 2018;**34**(23):4121-4123.
51. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol*. 2013;**30**(4):772-780.
52. 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.
53. 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.
54. 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.
55. Wickham H. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York; 2016.
56. Yu G, Smith DK, Zhu H, Guan Y, Lam TTY. ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution*. 2016;**8**(1):28-36.
57. 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.
58. McLaren ZM, Ardington C, Leibbrandt M. Distance decay and persistent health care disparities in South Africa. *BMC Health Serv Res*. 2014;**14**(1):541.
59. Rambaut A, Holmes EC, O'Toole A, et al. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nat Microbiol*. 2020;**5**(11):1403-1407.
60. 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.
61. 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.
62. 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.

Acknowledgements We thank Koen Deforche, Wim Dumon and Sara Cleemput at Emweb for their invaluable help to improve genome assembly and variant calling for SARS-CoV-2 which was central to the achievement of this work. 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. Finally, we would like to thank 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 Supplementary Table S5.

Author contributions Produced SARS-CoV-2 genomic data: Jennifer Giandhari, Sureshnee Pillay, Susan Engelbrecht, Deelan Doolabh, Lynn Tyers, Innocent Madua 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, Stephen Korsman, Diana Hardie, Carolyn Williamson, Gert Marais, Deelan Doolabh, Innocent Madau, Lynn Tyers. 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, Carolyn Williamson and Tulio de Oliveira. Helped with study design and data interpretation: Carolyn Williamson, Mary-Ann Davies, 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 The authors declare no competing interests.

Additional information

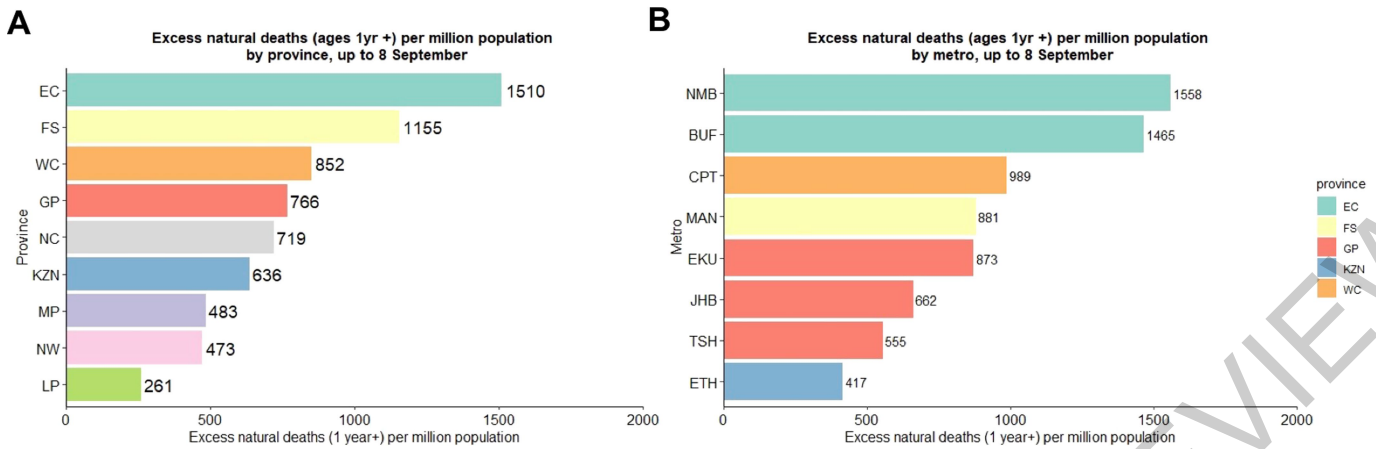
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-021-03402-9.

Correspondence and requests for materials should be addressed to T.d.O.

Peer review information *Nature* thanks Nathan Grubaugh, Tommy Tsan-Yuk Lam, Youchun Wang and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

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

ACCELERATED ARTICLE PREVIEW



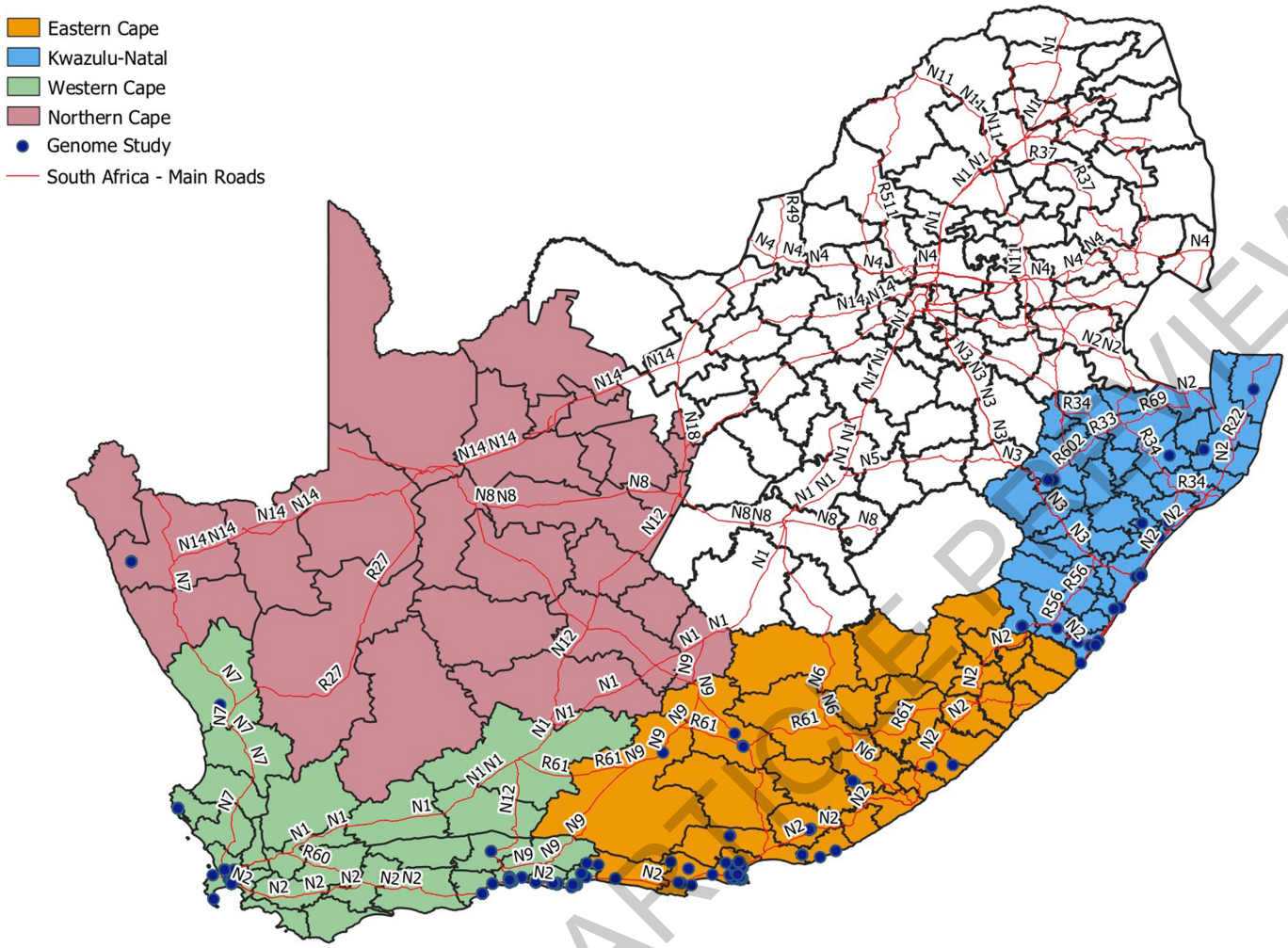
Extended Data Fig. 1 | Excess deaths per million population, by province and metropolitan municipalities of South Africa, up until the week ending 8 September 2020 (just past the first epidemic peak). These graphs indicate the disproportionate impact of the first wave of the epidemic in the province of the Eastern Cape (A) and its metro areas Nelson Mandela Bay (NMB) and Buffalo City (BUF) (B). [Key: EC=Eastern Cape, FS=Free State, WC=Western Cape, GP=Gauteng, 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].

ACCELERATED ARTICLE PREVIEW



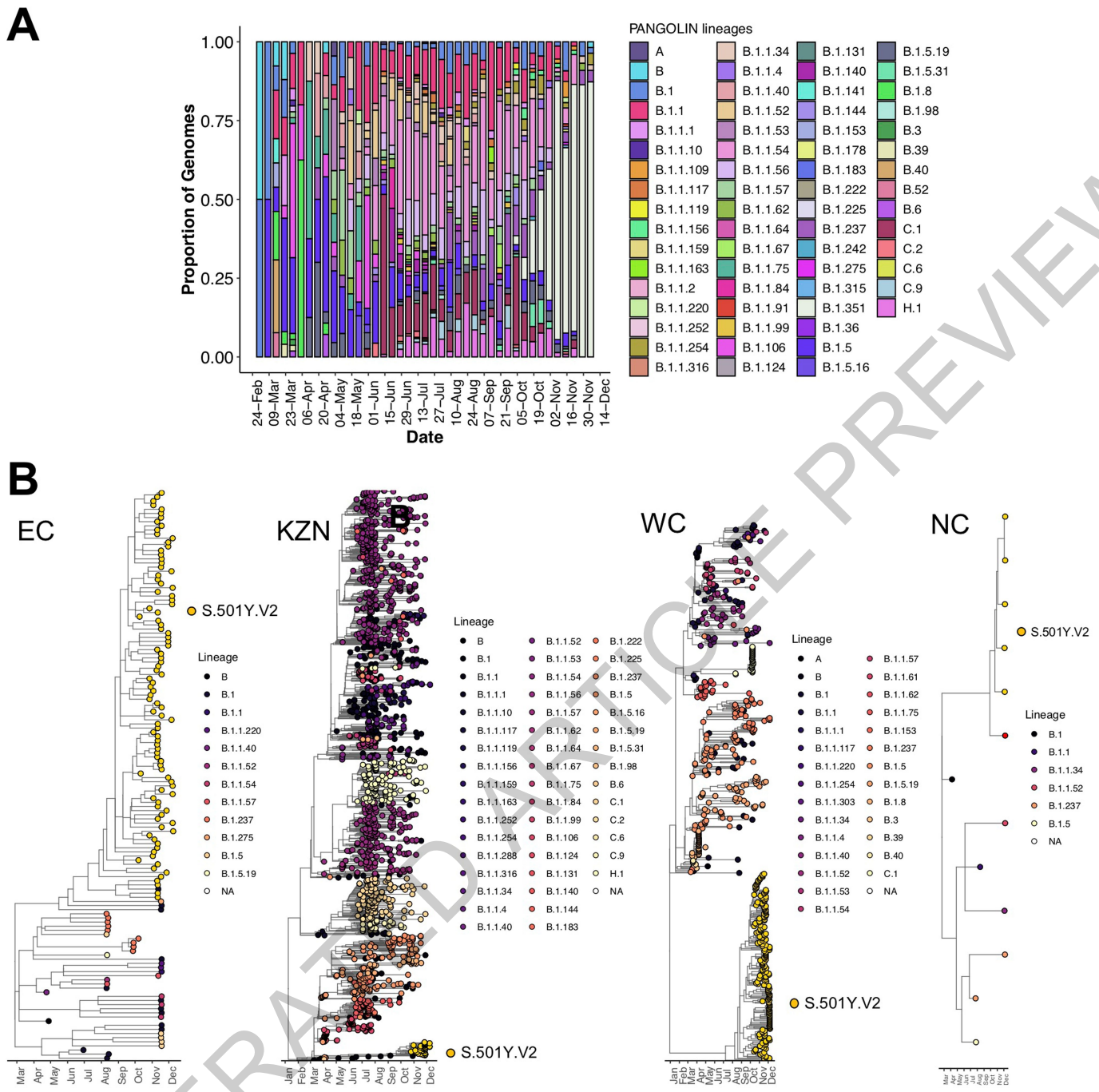
Extended Data Fig. 2 | Positivity test rates across 4 provinces of South Africa. Geographical maps of the Northern Cape, Western Cape, the Eastern Cape and KwaZulu-Natal (the four main provinces investigated in this study), showing a weekly progression of SARS-CoV-2 prevalence per district and colored by the rate of positive SARS-CoV-2 PCR-tests per district. Data for this figure was obtained from the NICD weekly testing report. The shapefile for this map was obtained from ArcGIS.

ACCELERATED ARTICLE PREVIEW



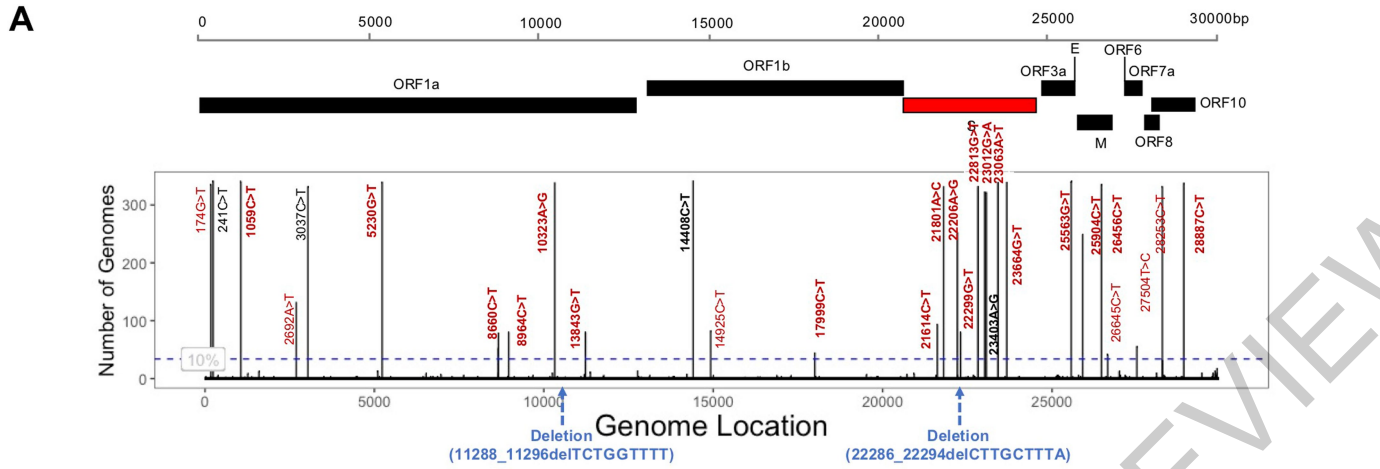
Extended Data Fig. 3 | Sampling location of 501Y.V2 genomes. A general map of South Africa, showing the sampling location of the 501Y.V2 genomes in this study (blue dots) in relation to the main road networks in the country, hinting to potential land transmission routes of this lineage along the coast. The shapefile for this map was obtained from ArcGIS.

ACCELERATED ARTICLE PREVIEW



Extended Data Fig. 4 | Replacement of other lineages by the new 501Y.V2 lineage. A) Progression of SARS-CoV-2 Pangolin lineages circulating in South Africa from March to December, showing over-representation of the new 501Y.V2 lineage from October onwards (B.1.351 in off-white). B) Independent regional phylogenetic trees for EC, KZN, WC and NC, showing a variety of circulating lineages prior to October and the dominance of 501Y.V2 (in yellow) in late October and November, especially in WC and EC.

ACCELERATED ARTICLE PREVIEW



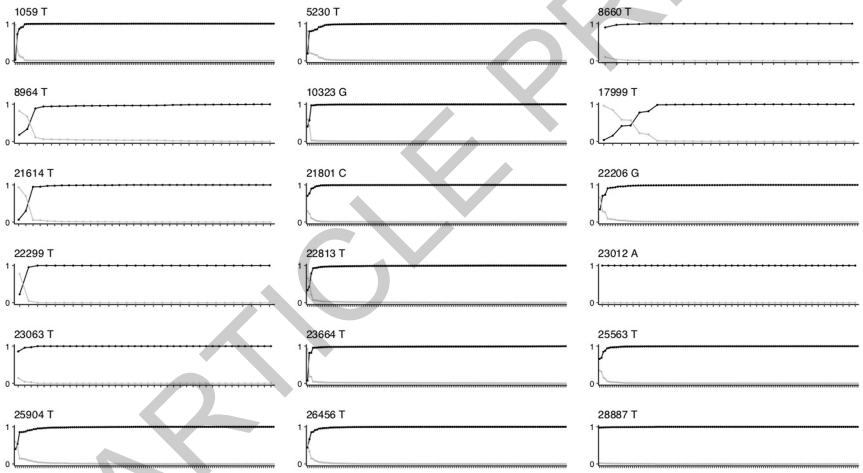
B

501Y.V2

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

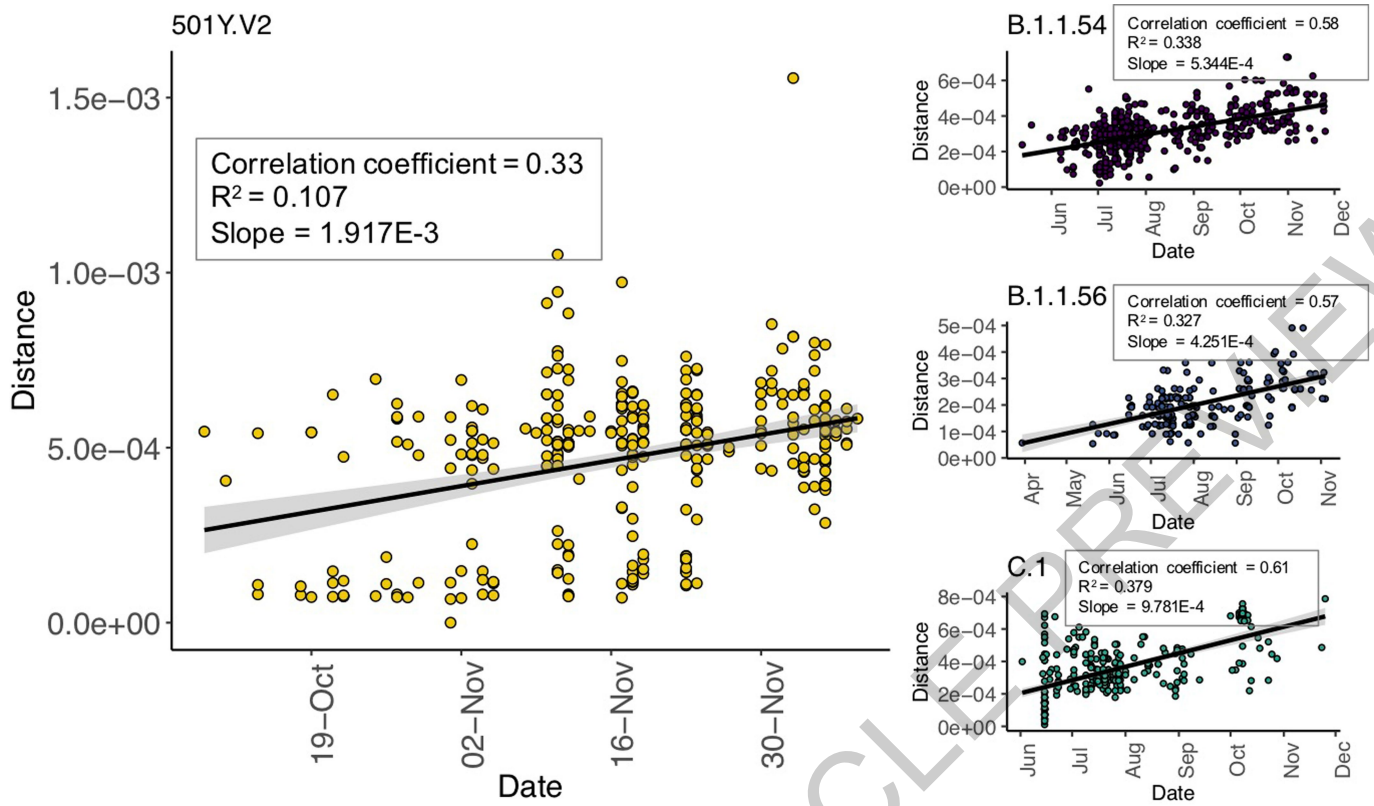
Summary:
19 lineage defining mutations (11 fixed)
8 in spike (4 fixed)
2 deletions (1 in spike) (both fixed)

C



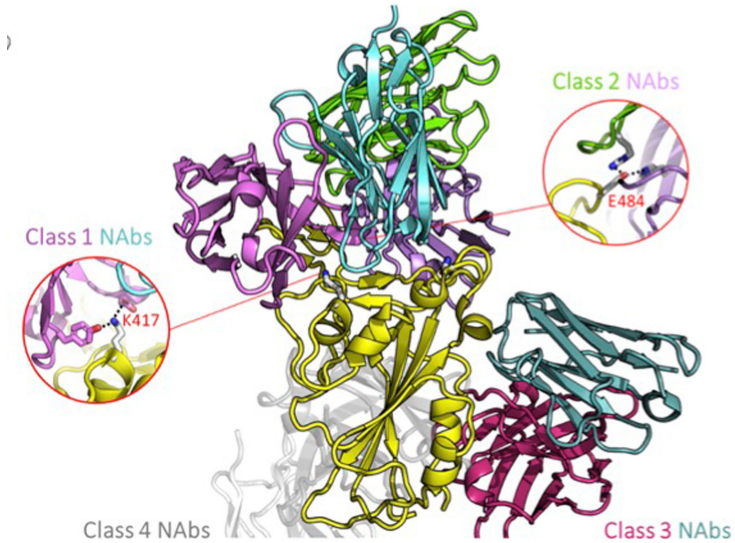
Extended Data Fig. 5 | Overview of mutations associated with the 501Y.V2 lineage. A) All nucleotide substitutions present in >10% of the genomes in the 501Y.V2 lineage mapped on the SARS-CoV-2 genomic structure. Mutations present in the parent lineage (B.1) are marked in black, while mutations specific to this lineage are marked in red or blue. All non-synonymous mutations are styled in bold, and also reported in panel B. The sites in blue specifically denote the location of deletions on the genomes of the 501Y.V2 lineage. It is important to note an unresolvable ambiguity in the representation of the exact location of the 22286-22294 nucleotide deletion; because of a hard-to-align repeat region 'CTTT', the deletion could be any 9-nucleotide segment between 22281-22289 and 22286-22294, which means that technically, the deletion could also be in amino acids 241-243, but the resulting amino acid sequence of all the possibilities are exactly the same (OTLH). B) A summary of all non-synonymous lineage-defining changes in relevant genes occurring in the 501Y.V2 lineage. C) Showing allele proportions at each 501Y.V2 lineage defining mutations sites, with the black line and dots showing the mutant allele proportion and the grey line and dots showing the reference allele proportion in individual samples in three sequencing runs.

ACCELERATED ARTICLE PREVIEW



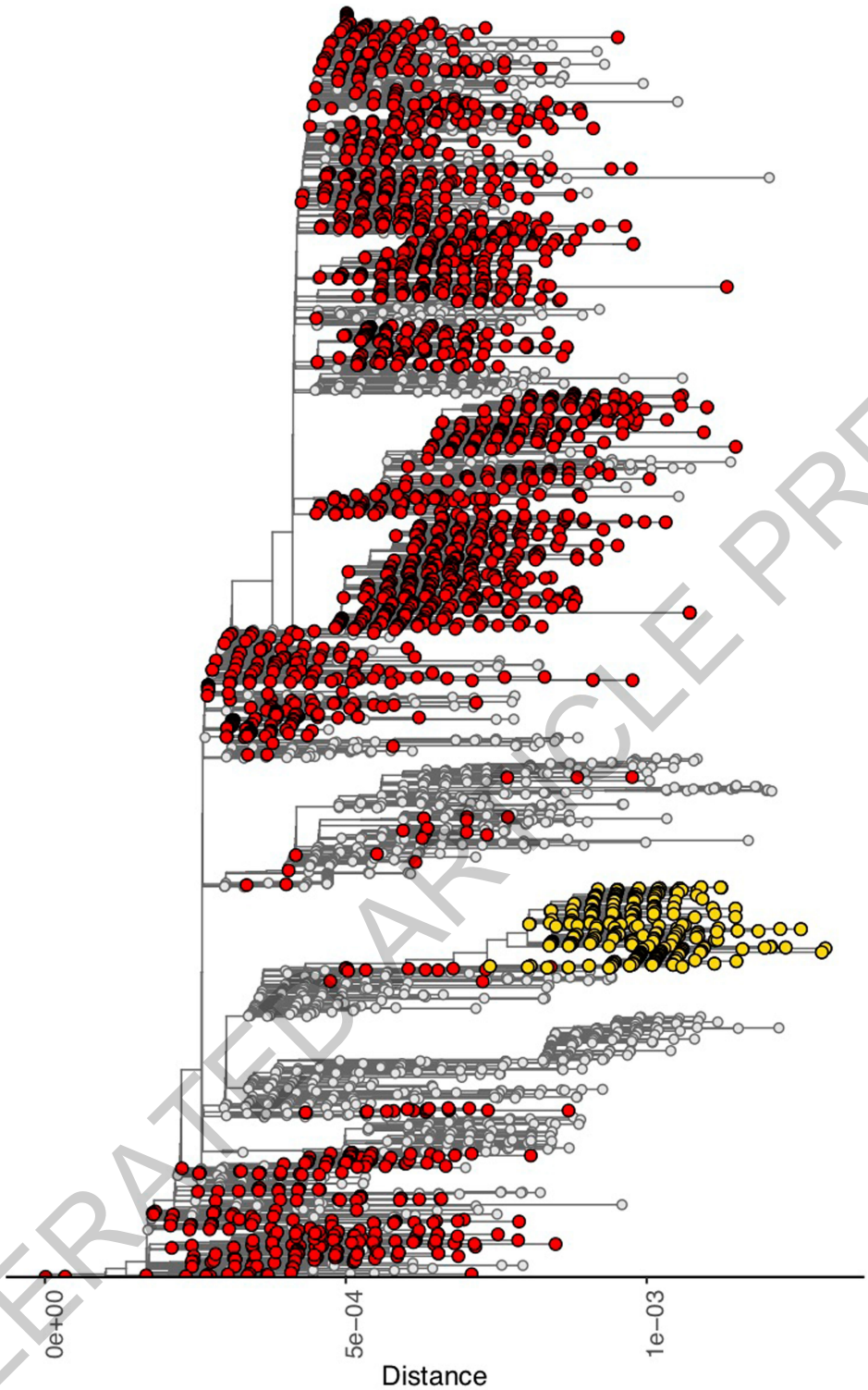
Extended Data Fig. 6 | Molecular clock signal of 4 main clusters spreading in South-Africa. Root-to-tip regression obtained from TempEst analysis for the 501Y.V2 lineage cluster (n=341), showing a relatively strong clock-like behavior (correlation coefficient = 0.33, R2 = 0.107) and a regression line slope, representing mean evolutionary rate, of 1.917E-3 nucleotide changes/site/year. A comparison is made 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 showing estimated mean evolutionary rates of 5.344E-4, 4.251E-4 and 9.781E-4 respectively. Regression lines are shown with error buffers (shaded area) representing 90% confidence intervals.

ACCELERATED ARTICLE PREVIEW



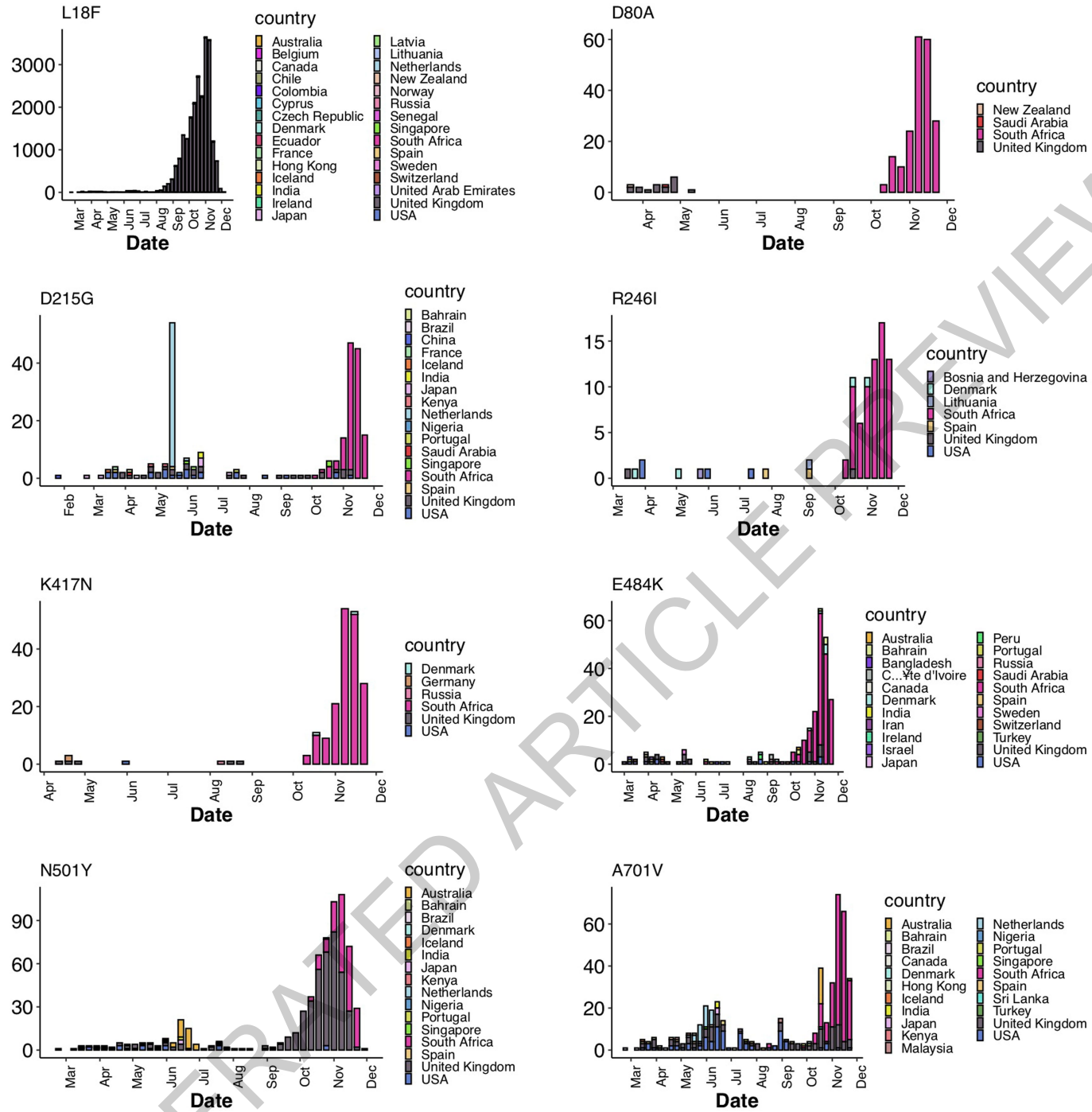
Extended Data Fig. 7 | SARS-CoV-2 receptor binding domain interactions with NAbs. A model of the SARS-CoV-2 receptor binding domain in cartoon view (yellow), showing representative Fab domains for neutralising antibodies from classes 1, 2, 3, and 4. Two zoomed insets show common, key interactions between RBD residue K417 and class 1 NAbs, or RBD residue E484 and class 2 NAbs.

ACCELERATED ARTICLE PREVIEW



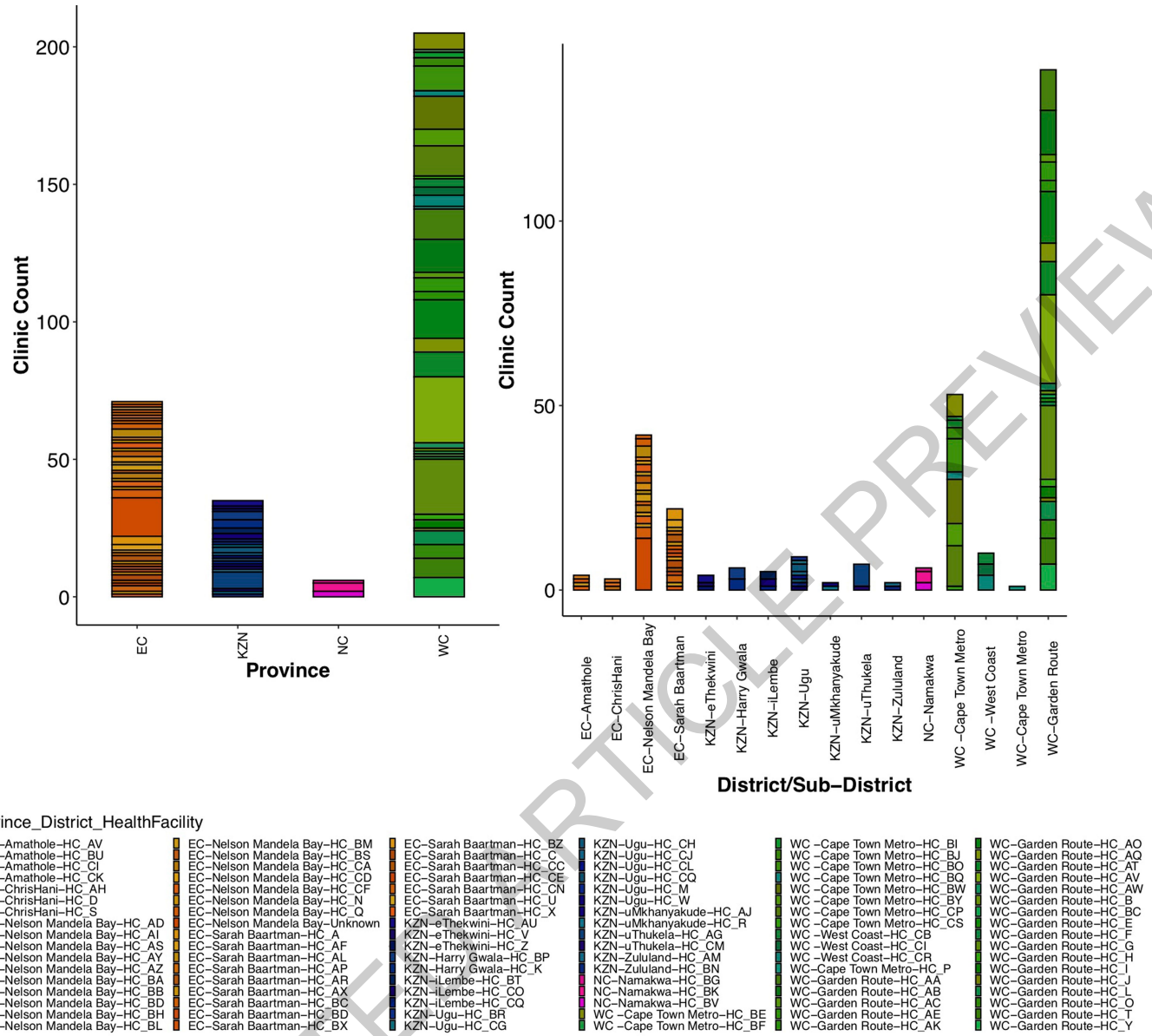
Extended Data Fig. 8 | A Maximum-Likelihood tree of 5332 SARS-CoV-2 genome of which 2759 are sampled from South Africa. The branch lengths represent the diversity of the genomes against the Wuhan reference. The 501Y. V2 lineages (in yellow) shows relatively longer branches compared to other South Africa viral genomes, forming other lineages circulating in the country, prior to the detection of this new lineage.

ACCELERATED ARTICLE PREVIEW



Extended Data Fig. 9 | Worldwide emergence of 8 spike mutations. Showing the prevalence of the 8 spike mutations around the world, indicating that several of them emerge independently in multiple regions.

ACCELERATED ARTICLE PREVIEW



Extended Data Fig. 10 | Random sampling of 501Y.V2 samples across health centers in 4 provinces. A representation of the number of health centers per province where the 501Y.V2 lineage was detected in sampled genomes in 4 provinces, showing a total number of 317 samples from 197 health centers. (There was no indication of health facility for the remaining 501Y.V2 samples presented in this study).

ACCELERATED ARTICLE PREVIEW

nature research

Corresponding author(s): Tulio de Oliveira

Last updated by author(s): Feb 9, 2021

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐ A description of all covariates tested
☒	☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software was used
Data analysis	Genome assembly was performed with Genome Detective online tool version 1.132 and validated with Geneious software v.2020.1.2. Phylogenetic analysis was performed using Nextstrain (https://github.com/nextstrain/ncov), iqtree v1.6.9, TempEst v1.5.3, MAFFT, BEASTv.1.10.4, and Tracer v.1.7.1. R packages used for data analysis included ggplot, ggtree, seraphim. Custom codes are all available at: https://github.com/krisp-kwazulu-natal/SARSCoV2_South_Africa_501Y_V2_B_1_351.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on 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/).

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	At the time of writing, 400 sequences of the 501Y.V2 SARS-CoV-2 variant had been produced by the NGS-SA (all fastq in SRA), and 341 genomes that passed quality control were used in this analysis. We believe this sample size was sufficient because the genomes come from >90 clinics across 4 provinces and numerous districts of South Africa.
Data exclusions	For phylogenetic analysis, genomes were excluded if they presented <90% coverage against the reference AND/OR have sequencing quality problem - e.g. gaps in key regions of the spike protein that causes spurious clustering.
Replication	Reproducibility were performed for maximum likelihood and bayesian MCMC phylogenetic tree reconstructions. We computed MCMC (Markov chain Monte Carlo) triplicate runs of 100 million states each, sampling every 10,000 steps for the 501Y.V2 dataset. All attempts at replication were successful and the MCC tree for the 501Y.V2 cluster was of high support.
Randomization	Samples for SARS-CoV-2 sequencing in South Africa were randomly selected. As part of the Network for Genomic Surveillance in South Africa (NGS-SA), five sequencing hubs receive randomly selected samples for sequencing every week according to approved protocols at each site. 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 (Suppl Fig. S1).
Blinding	Geographical blinding of data was not necessary for the study as it involves phylogeographical analysis, however the exact name of the health facilities associated with the genomic samples were anonymized. Data identification from the samples were also anonymized as this was not necessary for the analysis.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data

Methods

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

Human research participants

Policy information about studies involving human research participants

Population characteristics	We obtained samples consisting of 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. The 501Y.V2 genomes in this study came from patients of ages 6-84, from 192 female and 139 male patients, for which the 501Y.V2 genotype was confirmed by sequencing.
Recruitment	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. 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 (Suppl Fig. S1).
Ethics oversight	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.

Note that full information on the approval of the study protocol must also be provided in the manuscript.