Article

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

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

Received: 21 December 2020

Accepted: 26 February 2021

Published online: 9 March 2021

Check for updates

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], Nei-yuan 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 SARS-CoV-2 in many parts of the world is creating conditions for substantial evolutionary changes to the virus[1,2]. Here we describe a newly arisen lineage of SARS-CoV-2 (designated 501Y.V2; also known as B.1.351 or 20H) that is defined by eight mutations in the spike protein, including three substitutions (K417N, E484K and N501Y) at residues in its receptor-binding domain that may have functional importance[3–5]. This lineage was identified in South Africa after the first wave of the epidemic in a severely affected metropolitan area (Nelson Mandela Bay) that is located on the coast of the Eastern Cape province. This lineage spread rapidly, and became dominant in Eastern Cape, Western Cape and KwaZulu–Natal provinces within weeks. Although the full import of the mutations is yet to be determined, the genomic data—which show rapid expansion and displacement of other lineages in several regions—suggest that this lineage is associated with a selection advantage that most plausibly results from increased transmissibility or immune escape[6–8].

SARS-CoV-2 emerged in 2019 and has spread rapidly around the world, causing over 80 million recorded cases of COVID-19 and over 1.7 million deaths attributable to this disease 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 sharing of genomic surveillance data and collaborative online platforms have enabled the real-time tracking of the emergence and spread of these lineages[9,10].

To date, there has been relatively limited evidence for SARS-CoV-2 mutations that have had a substantial functional effect on the virus. A mutation resulting in a substitution 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 that carry this mutation have increased transmissibility[11–14]. Later in the epidemic, several lineages with a N439K substitution in the receptor-binding domain (RBD) of the spike protein emerged independently, probably in a range of European countries and the USA. This mutation is associated with escape from neutralization mediated by monoclonal antibodies or polyclonal serum[15].

South Africa is the most severely affected country in Africa, with over 80,000 excess natural deaths having occurred by the end of 2020 (approximately 1,400 per million individuals)[16]. The introduction and spread of several SARS-CoV-2 lineages to South Africa have previously been described, as has the identification of lineages unique to South Africa during the early phase of the epidemic[17,18]. Here we describe the emergence and spread of a SARS-CoV-2 lineage that contains several nonsynonymous spike mutations, including mutations that affect key sites in the RBD (resulting in K417N, E484K and N501Y substitutions) that may have functional importance. We demonstrate that this lineage is likely to have emerged after the first wave of the epidemic in the worst-affected metropolitan area within the Eastern Cape 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 wave of the SARS-CoV-2 epidemic in South Africa began around October 2020, weeks after a trough in daily recorded cases following the first peak[19] (Fig. 1a). The country-wide estimated effective reproduction number (R_e) increased to above 1 at the end of October (indicating a growing epidemic), which coincided with a steady rise in daily cases. At the peak of the national epidemic in the middle of July, there were over 13,000

A list of affiliations appears at the end of the paper.

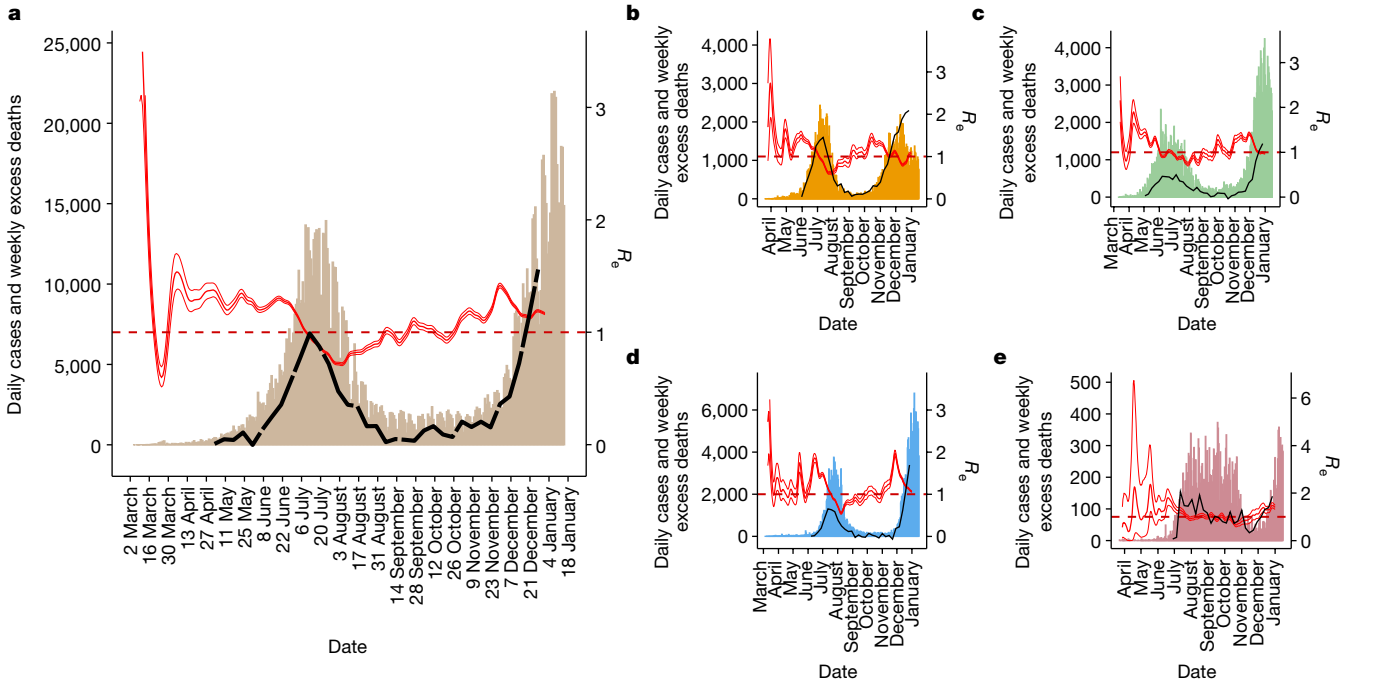


Fig. 1 | SARS-CoV-2 epidemiological dynamics in South Africa. **a**–**e**, Histograms show the number of daily confirmed cases of COVID-19 (mapped to the left *y* axis) from March 2020 to January 2021 in South Africa (**a**) and in the four provinces under study: Eastern Cape (**b**), Western Cape (**c**), KwaZulu–Natal (**d**) and Northern Cape (**e**). Fluctuations in the daily estimates of R_e are shown in red (mapped to the right *y* axis); the mean estimated median R_e with upper and lower bounds of the 95% confidence interval are shown, along with a cut-off for $R = 1$ (broken red line). Weekly excess deaths in South Africa and in each region are shown as black broken lines (mapped to the left *y* axis).

confirmed cases per day and almost 7,000 excess deaths per week. The epidemiological profile in the three provinces that are the focus of this analysis (the Eastern Cape, Western Cape and KwaZulu–Natal) were broadly similar, although the Western Cape had an earlier and flatter peak in the first wave (Fig. 1b–d). At the end of the first wave of the epidemic in early September, there had been over 10,000 excess deaths in the Eastern Cape (1,510 per million individuals)—the highest for any province (Extended Data Fig. 1). Although there was a plateau in cases after the first wave, this was noticeably short in the Eastern Cape; by early October, there was a second phase of exponential growth that was associated with an increase in deaths at a rate similar to that of the first wave (Fig. 1b). The rate of positive PCR tests at a local-municipality level shows very high levels of infection (>20%) in Nelson Mandela Bay from the middle of October, followed by rapidly rising levels in the surrounding areas through October and November (Extended Data Fig. 2). The resurgence of the daily case counts at an exponential rate happened later for the Western Cape and KwaZulu–Natal than for the Eastern Cape (Fig. 1c, d). By early December, all three provinces were experiencing a second wave and new cases in the Western Cape 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 Eastern Cape and Western Cape prompted the intensification of genomic surveillance by the Network for Genomic Surveillance in South Africa (NGS-SA), including sampling in and around Nelson Mandela Bay in the Eastern Cape and in the neighbouring Garden Route district of the Western Cape (Extended Data Fig. 3). We analysed 2,882 whole genomes of SARS-CoV-2 from South Africa, which were collected between 5 March and 10 December 2020. We estimated preliminary maximum-likelihood and molecular clock phylogenies for a dataset containing an additional 2,573 global reference genomes (Fig. 2a). We identified a previously unrecognized monophyletic cluster (501Y.V2) that contained 341 sequences, from samples collected between 8 October and 10 December in KwaZulu–Natal, Eastern Cape, Western Cape and Northern Cape (Fig. 2b). Seven South African sequences that are basal to the 501Y.V2 cluster (Fig. 2a) were sampled in the Eastern Cape, Western Cape, Gauteng and KwaZulu–Natal provinces between late June and early September. Although these sequences do not have any of the defining mutations of the 501Y.V2 variant, they are basal to the B.1.351 lineage and indicate that the precursor to the new variant was probably circulating throughout the country before the emergence of 501Y.V2.

The 501Y.V2 cluster is phylogenetically distinct from the three main lineages (B.1.1.54, B.1.1.56 and C.1) that were circulating widely in South Africa (>42% of samples sequenced before October 2020) during the first wave of infections[18] (Fig. 2a). These three lineages had been circulating in the KwaZulu–Natal, Western Cape, Gauteng, Free State, Limpopo and North-West provinces. By the middle of November, the 501Y.V2 lineage had superseded the B.1.1.54, B.1.1.56 and C.1 lineages, and it rapidly became the dominant lineage in samples from the Eastern Cape, KwaZulu–Natal and Western Cape (Fig. 2c, Extended Data Fig. 4).

Our spatiotemporal phylogeographic analysis suggests that the 501Y.V2 lineage emerged in early August (95% highest posterior density ranging from the middle of July to the end of August 2020) in Nelson Mandela Bay. Its initial spread to the Garden Route district of the Western Cape was followed by a more-diffuse spread from both of these areas to other regions of the Eastern Cape, and more recently to the City of Cape Town municipality and several locations in KwaZulu–Natal (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 province.

Mutational profile

At the point of first sampling on the 15 October, this lineage had—in addition to D614G—five nonsynonymous mutations resulting in substitutions in the spike protein: D80A, D215G, E484K, N501Y and A701V

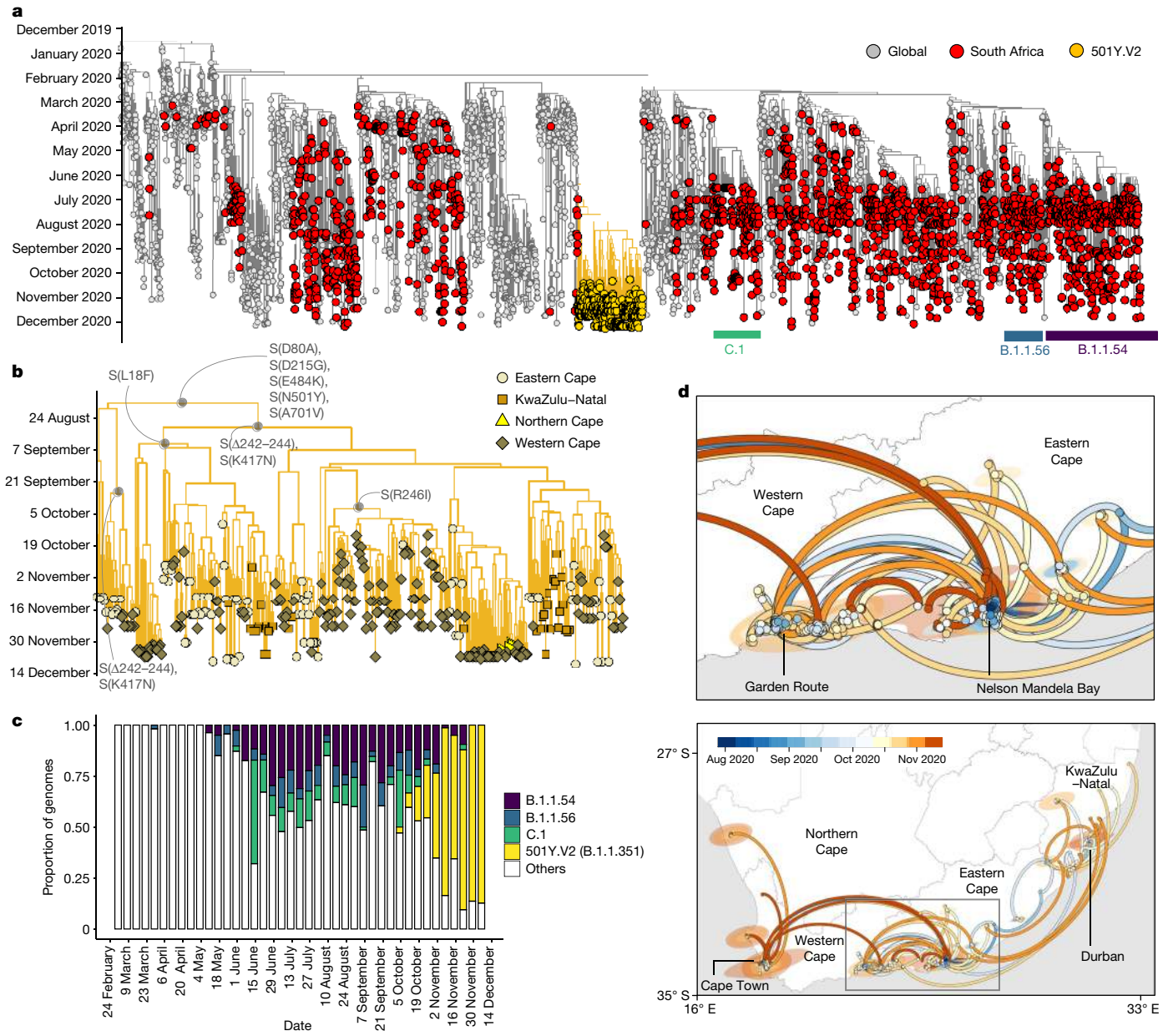


Fig. 2 | Evolution and spread of the 501Y.V2 cluster in South Africa. **a**, Time-resolved maximum clade credibility phylogeny of 5,329 SARS-CoV-2 sequences; 2,756 of these are from South Africa (red). The newly identified SARS-CoV-2 cluster (501Y.V2) is highlighted in yellow. **b**, Time-resolved maximum clade credibility phylogeny of the 501Y.V2 cluster ($n = 341$), with province indicated. Mutations that characterize the cluster are highlighted at the branch at which each 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 wave of the epidemic. Circles represent nodes of the maximum clade credibility phylogeny, coloured according to their inferred time of occurrence (scale in bottom 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.

(Figs. 2b, 3a, Extended Data Fig. 5). A further three mutations that lead to substitutions in the spike protein had emerged by the end of November: L18F, R246I and K417N. We also observe a deletion of three amino acids at positions 242 to 244, which was seen in samples extracted and generated in different laboratories across the NGS-SA. This region is difficult to align; the deletion could potentially also be located at positions 241 to 243, but the resulting sequence would be exactly the same. Although the variants appeared in a varying proportion of the sampled genomes and showed changing levels of frequency with time, the mutations in RBD seem to become fixed in our sampling set, are present in almost all of the samples and are consistently high in frequency across time (Fig. 3a, b). Compared to the previous three largest lineages circulating in South Africa, 501Y.V2 shows marked hypermutation both in the whole genomes and the spike regions—including nonsynonymous mutations that lead to amino acid changes (Fig. 3c). The main lineages identified in South Africa during first wave (B.1.1.54, B.1.1.56 and C.1) contained only the single nonsynonymous spike mutation (D614G) and did not show the rapid accumulation of mutations, as is observed with 501Y.V2. We estimate that substitutions on the 501Y.V2 lineage are happening at 1.917×10^{-3} nucleotide changes per site per year, compared to 5.344×10^{-4}, 4.251×10^{-4} and 9.781×10^{-4} nucleotide changes per site per year for B.1.1.54, B.1.1.56 and C.1, respectively (Extended Data Fig. 6).We performed structural modelling of the spike trimer with these mutations, which revealed that three of the substitutions (N501Y, E484K and

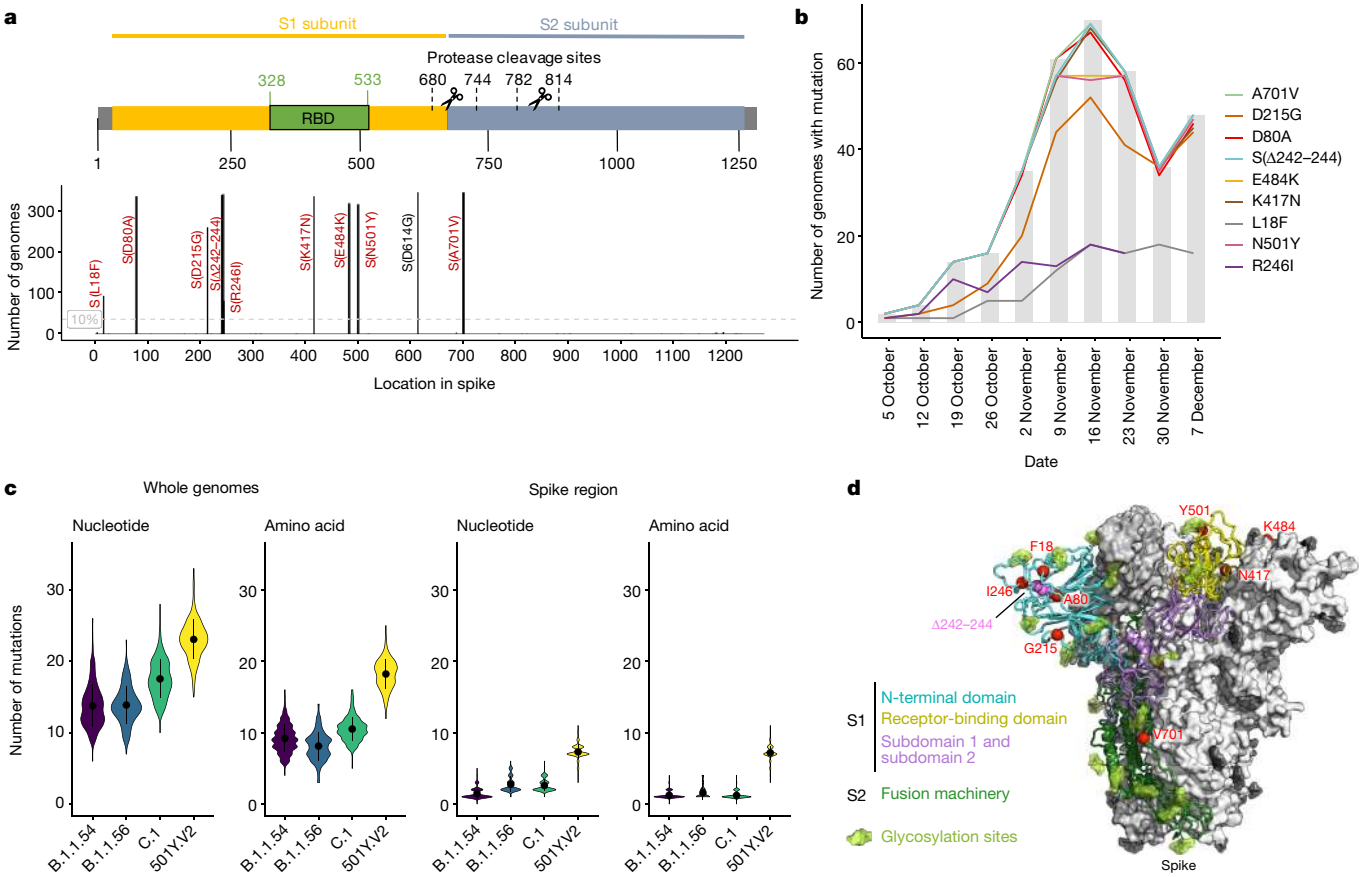


Fig. 3 | Mutational profile of the spike region of the 501Y.V2 lineage. **a**, Amino acid changes in the spike region of the 501Y.V2 genomes in this study ($n = 341$) mapped to the spike-protein sequence structure, indicating key regions (such as the RBD). Each spike protein variant is shown at its respective protein location; bar lengths represent the number of genomes that contain the specific mutations. Only mutations that appear in >10% (grey dotted line) of sequences are shown. The D614G substitution (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; coloured lines show the change in the number of sequences that contain each variant at the respective time points. **c**, Violin plots showing the number 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 ($n = 271$). The dot and error bars inside each group denote the mean and range for two s.d., respectively. **d**, A complete model of the SARS-CoV-2 spike trimer is shown, with domains of a single protomer shown in cartoon view and coloured cyan (N-terminal domain), yellow (C-terminal domain and receptor binding domain), purple (subdomain 1 and 2), and dark green; *N*-acetylglucosamine moieties are coloured in light green. The adjacent protomers are shown in surface view and coloured shades of grey. Eight nonsynonymous mutants (red) and a deletion of three amino acids (pink) that together define the spike of the 501Y.V2 lineage are shown as spheres.

K417N) are at key residues in the RBD; three (L18F, D80A and D215G) are in the N-terminal domain; and one (A701V) is in loop 2 (Fig. 3d). The deletion of three amino acid (242 to 244) also lies in the N-terminal domain. In particular, two of the RBD sites (at positions 417 and 484) are key regions for the binding of neutralizing antibodies (Extended Data Fig. 7).

Selection analysis

We examined patterns of nucleotide variation and fluctuations in mutant frequencies at eight polymorphic sites in the spike gene (Fig. 3a) to determine whether any of the observed polymorphisms might contribute to changes in viral fitness worldwide. For this analysis, we used 142,037 high-quality sequences from the Global Initiative On Sharing All Influenza Data (GISAID) sampled between the 24 December 2019 and 14 November 2020, which represented 5,964 unique spike haplotypes. The analysis indicated that two of the three sites in the 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 the 15 October shows indications of positive selection on five global-tree internal branches; codon 501 of the spike gene displays a significant excess of nonsynonymous substitutions globally ($dN/dS > 1$ on internal branches, $P = 0.0011$ by the fixed-effects likelihood method), and mutant viruses that encode Y at this site have rapidly increased in frequency in both the UK and South Africa (z score = 11, trend Jonckheere Terpstra non-parametric trend test). Similarly, at codon 484, there is an indication of positive selection on seven global-tree internal branches, with an overall significant excess of nonsynonymous substitutions globally ($P = 0.015$). Outside the RBD, codons 18 ($P < 0.001$), 80 ($P = 0.0014$) and 215 ($P < 0.001$) show evidence of positive diversifying selection globally, and the L18F mutation has also increased in frequency in the regions in which it has occurred (z score = 17). Up until the 14 November 2020, there was no statistical evidence of positive selection at codons 417, 246 and 701.

Discussion

We describe and characterize a newly identified SARS-CoV-2 lineage with several spike mutations that is likely to have emerged in a major metropolitan area in South Africa after the first wave of the epidemic, and then to have spread to multiple locations within two neighbouring

provinces. We show that this lineage has rapidly expanded and become dominant in three provinces, at the same time as there has been a rapid resurgence in infections. Although the full import of the mutations is not yet clear, the genomic and epidemiological data suggest that this variant has a selective advantage—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 hospitalizations and deaths but also to limit the spread of this lineage and the further evolution of the virus.

We detected this variant through intensified genomic surveillance that was enacted 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 the regular sequencing of a random selection of residual samples from routine diagnostic services. We show that 501Y.V2 was detected in samples from 197 health facilities in multiple districts across four provinces. We are therefore confident that, although 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 from the middle of July to the middle of August (the estimated time to the most-recent common ancestor), this was not a period of low transmission: incidence was above 20 case per 100,000 people per week at this time and the positive testing rate remained above 10%, which suggests moderate-to-high levels of transmission. As there were many lineages circulating at this time, the rapid expansion of 501Y.V2 and the almost complete displacement of other lineages in multiple regions strongly suggest 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 reinfection[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 that are emerging in other locations. The 501Y.V2 lineage has three substitutions that affect key sites in the RBD (K417N, E484K and N501Y). The N501Y substitution has also recently been identified in a lineage that has spread rapidly in the UK (designated B.1.1.7)[21]. There is now good evidence that this lineage is associated with increased transmissibility[22]. The N501Y substitution has previously been shown through deep mutation scanning, and in a mouse model, to enhance binding affinity to human ACE2[3,23]. There is some evidence that the E484K substitution may also increase binding affinity to human ACE2[3]; and that the combination of N501Y and E484K enhances binding affinity still further[24,25]. Additional work is being conducted to understand the precise mechanisms that underlie the increased transmissibility of these new variants.

The other reason for a selective advantage of 501Y.V2 could be immune escape (that is, the capacity to cause reinfection). We have very limited SARS-CoV-2 seroprevalence data from South Africa to help us to understand the true extent of the epidemic. In studies that used 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 and August (toward the end of the first wave of the epidemic in this area)[26]. We have shown that the Eastern Cape—and Nelson Mandela Bay, in particular—were worse-affected than City of Cape Town in the first wave, and we therefore believe that population immunity could have been sufficiently high in this region to contribute to population-level selection. The RBD of the spike protein is the main target of neutralizing antibodies that are elicited during SARS-CoV-2 infection[27]. Neutralizing antibodies to the RBD can be broadly divided into four main classes[28]. Of these, class 1 and class 2 antibodies appear to be elicited most frequently during SARS-CoV-2 infection, and their epitopes directly overlap the human ACE2 binding site[27]. Class 2 antibodies bind to E484, and the E484K substitution has previously been shown to confer resistance to neutralizing antibodies in this class and to panels of convalescent sera, which suggests 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 N-terminal domain mutations in 501Y.V2 affect residues that form an antigenic supersite or are close to this site[32,33]. Preliminary evidence from live virus and pseudovirus experiments indicates that 501Y.V2 shows substantial or complete escape from neutralizing antibodies in convalescent plasma[7,8]. We are currently investigating the frequency of reinfection in the second wave, as well as the clinical presentations of individuals with reinfection to better understand the clinical and epidemiological effects of any immune escape. We are also conducting neutralization assays on plasma from recipients of vaccines, 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 intrahost evolution[34–36]. This hypothesis is supported by the long branch length that connects the lineage to the remaining sequences in our phylogenetic tree (Extended Data Fig. 8). The mutation leading to the N501Y substitution is one of several spike mutations that emerged in an immunocompromised individual in the USA who had prolonged viral replication for over 20 weeks[34]. In South Africa (which has the largest HIV epidemic in the world), one concern has been the possibility of prolonged viral replication and intrahost 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]. However, 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 characterize 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 several individuals. Furthermore, antigenic evolution—even within individuals who are not immunosuppressed—could offer an alternative explanation, as several of the individual sites in the spike protein appear to be under selective pressure worldwide and several of the identified mutations have emerged independently around the world (Extended Data Fig. 9) and been found in circulating lineages together.

Although the full implications of the 501Y.V2 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 in enabling the early detection and characterization of new lineages and informing the global response to the COVID-19 pandemic.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, 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. et al. SARS-CoV-2 variants and ending the COVID-19 pandemic. *Lancet* **397**, 952–954 (2021).
2. Mascola, J. R., Graham, B. S. & Fauci, A. S. SARS-CoV-2 viral variants—tackling a moving target. *J. Am. Med. Assoc.* https://doi.org/10.1001/jama.2021.2088 (2021).
3. 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 (2020).
4. 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).
5. 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 (2021).
6. Pearson, C. A.B. et al. Estimates of severity and transmissibility of novel South Africa SARS-CoV-2 variant 501Y.V2. Preprint at https://cmmid.github.io/topics/covid19/sa-novel-variant.html (2021).

7. Cele, S. et al. Escape of SARS-CoV-2 501Y.V2 from neutralization by convalescent plasma. Nature https://doi.org/10.1038/s41586-021-03471-w (2021).
8. Wibmer, C. K. et al. SARS-CoV-2 501Y.V2 escapes neutralization by South African COVID-19 donor plasma. *Nat. Med*. https://doi.org/10.1038/s41591-021-01285-x (2021).
9. Hadfield, J. et al. Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics* **34**, 4121–4123 (2018).
10. Shu, Y. & McCauley, J. GISAID: global initiative on sharing all influenza data – from vision to reality. *Euro Surveill*. **22**, 30494 (2017).
11. 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 (2020).
12. Plante, J. A. et al. Spike mutation D614G alters SARS-CoV-2 fitness. *Nature* https://doi.org/10.1038/s41586-020-2895-3 (2020).
13. Yurkovetskiy, L. et al. Structural and functional analysis of the D614G SARS-CoV-2 spike protein variant. *Cell* **183**, 739–751 (2020).
14. Volz, E. et al. Evaluating the effects of SARS-CoV-2 spike mutation D614G on transmissibility and pathogenicity. *Cell* **184**, 64–75 (2021).
15. Thomson, E. C. et al. Circulating SARS-CoV-2 spike N439K variants maintain fitness while evading antibody-mediated immunity. *Cell* **184**, 1171–1187.e20 (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)* (Burden of Disease Research Unit, South African Medical Research Council, 2020).
17. 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).
18. Tegally, H. et al. Sixteen novel lineages of SARS-CoV-2 in South Africa. *Nat. Med*. **27**, 440–446 (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 (2020).
20. Msomi, N., Mlisana, K. & de Oliveira, T. A genomics network established to respond rapidly to public health threats in South Africa. *Lancet Microbe* **1**, e229–e230 (2020).
21. Rambaut, A. et al. Preliminary genomic characterisation of an emergent SARS-CoV-2 lineage in the UK defined by a novel set of spike mutations. *Virological* 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 (2020).
22. Volz, E. et al. Transmission of SARS-CoV-2 lineage B.1.1.7 in England: insights from linking epidemiological and genetic data. Preprint at https://doi.org/10.1101/2020.12.30.20249034 (2021).
23. Gu, H. et al. Adaptation of SARS-CoV-2 in BALB/c mice for testing vaccine efficacy. *Science* **369**, 1603–1607 (2020).
24. Nelson, G. et al. 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. Preprint at https://doi.org/10.1101/2021.01.13.426558 (2021).
25. Zahradník J. et al. SARS-CoV-2 RBD in vitro evolution follows contagious mutation spread, yet generates an able infection inhibitor. Preprint at https://doi.org/10.1101/2021.01.06.425392 (2021).
26. Hsiao, M. et al. SARS-CoV-2 seroprevalence in the Cape Town metropolitan sub-districts after the peak of infections. *NICD COVID-19 Special Public Health Surveill. Bull*. **18**, 1–9 (2020).
27. Piccoli, L. et al. Mapping neutralizing and immunodominant sites on the SARS-CoV-2 spike receptor-binding domain by structure-guided high-resolution serology. *Cell* **183**, 1024–1042 (2020).
28. Barnes, C. O. et al. SARS-CoV-2 neutralizing antibody structures inform therapeutic strategies. *Nature* **588**, 682–687 (2020).
29. Baum, A. et al. Antibody cocktail to SARS-CoV-2 spike protein prevents rapid mutational escape seen with individual antibodies. *Science* **369**, 1014–1018 (2020).
30. Liu, Z. et al. Identification of SARS-CoV-2 spike mutations that attenuate monoclonal and serum antibody neutralization. *Cell Host Microbe* **29**, 477–488.e4 (2021).
31. Weisblum, Y. et al. Escape from neutralizing antibodies by SARS-CoV-2 spike protein variants. *eLife* **9**, e61312 (2020).
32. Chi, X. et al. A neutralizing human antibody binds to the N-terminal domain of the spike protein of SARS-CoV-2. *Science* **369**, 650–655 (2020).
33. McCallum, M. et al. N-terminal domain antigenic mapping reveals a site of vulnerability for SARS-CoV-2. *Cell* https://doi.org/10.1016/j.cell.2021.03.028 (2021).
34. Choi, B. et al. Persistence and evolution of SARS-CoV-2 in an immunocompromised host. *N. Engl. J. Med*. **383**, 2291–2293 (2020).
35. Avanzato, V. A. et al. Case study: prolonged infectious SARS-CoV-2 shedding from an asymptomatic immunocompromised individual with cancer. *Cell* **183**, 1901–1912 (2020).
36. Kemp, S. A. et al. SARS-CoV-2 evolution during treatment of chronic infection. *Nature* https://doi.org/10.1038/s41586-021-03291-y (2021).
37. Karim, F. et al. HIV infection alters SARS-CoV-2 responsive immune parameters but not clinical outcomes in COVID-19 disease. Preprint at https://doi.org/10.1101/2020.11.23.20236828 (2020).

[1]KwaZulu–Natal Research Innovation and Sequencing Platform (KRISP), School of Laboratory Medicine and 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, Brazil. [4]Computational Biology Division, Department of Integrative Biomedical Sciences, University of Cape Town, Cape Town, 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, School of Laboratory Medicine and Medical Sciences and National Health Laboratory Service (NHLS), University of KwaZulu–Natal, Durban, South Africa. [7]NHLS, Johannesburg, South Africa. [8]Centre for the AIDS Programme of Research in South Africa (CAPRISA), Durban, South Africa. [9]National Institute for Communicable Diseases, NHLS, 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, Faculty of Medicine and Health Sciences, Stellenbosch University and NHLS Tygerberg Hospital, Cape Town, South Africa. [14]Centre for Quantum Technology, University of KwaZulu–Natal, Durban, South Africa. [15]National Institute for Theoretical Physics (NITheP), University of KwaZulu–Natal, Durban, 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, 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, NHLS Universitas Academic Laboratories, University of The Free State, Bloemfontein, South Africa. [25]NHLS, 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, CA, USA. [29]Structural Biology Research Unit, Department of Integrative Biomedical Sciences, University of Cape Town, Cape Town, South Africa. [30]Department of Zoology, University of Oxford, Oxford, UK. [31]Institute for Genomics and Evolutionary Medicine, Temple University, Philadelphia, PA, USA. [32]Department of Global Health, University of Washington, Seattle, WA, 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

Methods

No statistical methods were used to predetermine sample size. The experiments were not randomized, and investigators were not blinded to allocation during experiments and outcome assessment.

Epidemiological dynamics

We analysed daily cases of SARS-CoV-2 in South Africa up the 16 January 2020 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)[38,39]. The National Department of Health 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 onto general epidemiological data to determine the extent of potential underreporting 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 these data from the Report on Weekly Deaths from the South Africa Medical Research Council Burden of Disease Research Unit[16]. We generated estimates for the R_e of SARS-CoV-2 in South Africa from the 'covid-19-Re' data repository (https://github.com/covid-19-Re/dailyRe-Data) as of the 14 December 2020[40].

Sampling of SARS-CoV-2

As part of the NGS-SA[20], 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 the Eastern Cape and the Garden Route district of the Western Cape 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 Eastern Cape and Western Cape (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[41]. In brief, extracted RNA was converted to cDNA using the Superscript IV First Strand synthesis system (Life Technologies) 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 400-bp amplicons with an overlap of 70 bp that covers the 30-kb SARS-CoV-2 genome. For nanopore sequencing, we adapted the nCoV-2019 sequencing LoCost protocol v3[41]. In brief, PCR reactions were done in 12.5 µl volumes and no PCR product purification was done. After DNA repair (NEB) and end-prep reactions (NEB), up to 24 samples were barcoded by ligation (EXP-NBD104/NBD114, Oxford Nanopore Technologies). 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). 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) and quantified using the Qubit dsDNA High Sensitivity assay on the Qubit 4.0 instrument (Life Technologies).

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 4 nM, pooled and denatured with 0.2 N sodium acetate. A 12 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). Full details of the amplification and sequencing protocol have previously been published[42,43].

We assembled paired-end and nanopore .fastq reads using Genome Detective 1.132 (https://www.genomedetective.com) and the Coronavirus Typing Tool[44]. For short reads, to accurately call mutations and short insertions and deletions (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[45] against the reference rather than the de novo consensus sequence, and subsequently final mutations and indels are called using GATK HaplotypeCaller[46], with low-quality variants (with QD < 10) filtered using GATK VariantFiltration[46]. 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 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 four or longer that break the open reading frame (probably sequencing errors). 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). All of the sequences were deposited in GISAID (https://www.gisaid.org/), and the GISAID accession identifiers are included as part of Supplementary Table 2. Raw reads for our sequences have also been deposited at the NCBI Sequence Read Archive (BioProject accession PRJNA694014).

In some samples, the K417N substitution was previously not called. To avoid an assembly concern, these samples were also analysed using the ARTIC Illumina pipeline (https://github.com/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 1). 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 substitutions (that is, E484K and N501Y) in sequences in which we were not confident about the call from next-generation sequencing data. The full sequence

properties, mutation and spike mutations of the 501Y.V2 sequences are shown in Supplementary Tables 3, 4.

LoFreq was used to detect minor viral variants to study the intrahost heterogeneity of viral variants (quasi-species)[47] (Extended Data Fig. 5). Variants were called with at minimum coverage of 10% and conservative false discovery rate *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 that fall below the *P* value threshold.

Quality control of genomic sequences from South Africa

We retrieved all SARS-CoV-2 genomes from South Africa from the GISAID database as of the 4 January 2021 (n = 2,882). Before phylogenetic reconstruction, we removed low-quality sequences from this dataset. We filtered out genomes that did not pass standard quality assessment parameters used in NextClade (https://clades.nextstrain.org). We filtered out 105 genomes from South Africa owing to low coverage, and a further 18 owing to poor sequence quality. Poor sequence quality was defined as sequences with clustered single-nucleotide polymorphisms and ambiguous bases at >10% of sites, and low-coverage genomes were anything with <90% genome coverage against the reference. We therefore analysed a total of 2,756 South African genomes. We also retrieved a global reference dataset (n = 2,573). This was selected from the NextStrain global reference dataset, plus the five most similar sequences to each of the sequences from South Africa as defined by a local BLAST search.

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)[9]. The pipeline contains several Python scripts that manage the analysis workflow. It performs an alignment of genomes in MAFFT[48], phylogenetic tree inference in IQ-Tree V1.6.9[49], 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 enabled us to identify a large cluster of sequences (n = 341) with multiple spike mutations. We extracted this cluster and constructed a preliminary maximum-likelihood tree in IQ-tree, together with seven basal sequences from the region that were sampled from June to September 2020. We inspected this maximum-likelihood tree in TempEst v.1.5.3 for the presence of a temporal (that is, 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 = 0.33, R^2 = 0.11) (Extended Data Fig. 6).

We then estimated time-calibrated phylogenies using the Bayesian software package BEAST v.1.10.4. For this analysis, we used the strict molecular clock model, the HKY+I, nucleotide substitution model and the exponential growth coalescent model[50]. 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[51]. 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[52,53].

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[54]. 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 do not have access to residential geolocations within the genomic surveillance, the location of the health facility serves as a reasonable proxy, especially as two-thirds of the population live within 2 km of their nearest health facility[55].

As described in 'Phylogenetic analysis', MCMC chains were run in duplicate for 100 million generations and sampled every 10,000 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 to extract and map spatiotemporal information embedded in posterior trees.

Lineage classification

We used a previously proposed[56] dynamic lineage classification method from 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, 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. For the 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)[57], (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)[58] and (iii) whether these polymorphisms have increased in frequency in the regions of the world in which they have occurred.

Structural modelling

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[59]. We used the Pymol program (The PyMOL Molecular Graphics System, version 2.2.0) for visualization.

Reporting summary

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

Data availability

All of the SARS-CoV-2 501Y.V2 genomes generated and presented in this Article are publicly accessible through the GISAID platform (https://www.gisaid.org/), along with all other SARS-CoV-2 genomes generated by the NGS-SA. The GISAID accession identifiers of the 501Y.V2 sequences analysed in this study are provided as part of Supplementary Table 2, which also contains the metadata for the sequences. The raw reads for the 501Y.V2 have been deposited at the NCBI Sequence Read Archive (BioProject accession PRJNA694014). Other raw data for this study are provided as a supplementary dataset at 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 Article are available at https://github.com/krisp-kwazulu-natal/SARSCoV2_South_Africa_501Y_V2_B_1_351.

38. 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**, 19 (2020).
39. Marivate, V. et al. Coronavirus disease (COVID-19) case data – South Africa https://zenodo.org/record/3732419#.YBXzFegzZPY (2020).
40. Huisman, J. S. et al. Estimation and worldwide monitoring of the effective reproductive number of SARS-CoV-2. Preprint at https://doi.org/10.1101/2020.11.26.20239368 (2020).
41. Quick, J. nCoV-2019 sequencing protocol v3 (LoCost) V.3. *protocols.io* https://www.protocols.io/view/ncov-2019-sequencing-protocol-v3-locost-bh42j8ye (2020).
42. Giandhari, J. et al. NEBnext library construction and sequencing for SARS-CoV-2: adapting COVID-19 ARTIC protocol. *protocol.io* https://www.protocols.io/view/nebnext-library-construction-and-sequencing-for-sa-bhu2j6ye (2020).
43. Pillay, S. et al. Whole genome sequencing of SARS-CoV-2: adapting Illumina protocols for quick and accurate outbreak investigation during a pandemic. *Genes* **11**, 949 (2020).
44. Cleemput, S. et al. Genome detective coronavirus typing tool for rapid identification and characterization of novel coronavirus genomes. *Bioinformatics* **36**, 3552–3555 (2020).
45. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
46. McKenna, A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
47. 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).
48. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
49. 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).
50. Griffiths, R. C. & Tavaré, S. Sampling theory for neutral alleles in a varying environment. *Phil. Trans. R. Soc. Lond. B* **344**, 403–410 (1994).
51. 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).
52. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis* (Springer 2016).
53. Yu, G., Smith, D. K., Zhu, H., Guan, Y. & Lam, T. T. Y. ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol. Evol.* **8**, 28–36 (2017).
54. 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).
55. McLaren, Z. M., Ardington, C. & Leibbrandt, M. Distance decay and persistent health care disparities in South Africa. *BMC Health Serv. Res.* **14**, 541 (2014).
56. Rambaut, A. et al. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nat. Microbiol.* **5**, 1403–1407 (2020).
57. Murrell, B. et al. Detecting individual sites subject to episodic diversifying selection. *PLoS Genet.* **8**, e1002764 (2012).
58. Kosakovsky Pond, S. L. & Frost, S. D. Not so different after all: a comparison of methods for detecting amino acid sites under selection. *Mol. Biol. Evol.* **22**, 1208–1222 (2005).
59. 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).

Acknowledgements We thank K. Deforche, W. Dumon and S. Cleemput at Emweb for their invaluable help in improving genome assembly and variant calling for SARS-CoV-2, which was central to the achievement of this work; and 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 5). 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 J.G., S.P., S.E., D.D., L.T., I.M. and A. Ismail produced SARS-CoV-2 genomic data. N.M., K.M., N.-y.H., D.Y., D.G., A.v.G., S.W., A.J.G., A.S., G.V.Z., W.P., S.K., D.H., C.W., G.M., D.D., I.M., L.T., C.M., O.L.-A., T.M. and S.A. collected samples and curated metadata. H.T., E.W., M.G., A. Iranzadeh, R.J.L., S.L.K.P., S. Weaver, D.M., J.G., S.P., E.J.S., S.E., F.P., A. Ismail, J.N.B., V.F., J.L., L.C.J.A., A.A.-D., A.G., B.T.S., C.K.W., M.A., C.W. and T.d.O. analysed the data. C.W., M.-A.D., D.H., N.-y.H., D.M., D.G., E.J.S., M.G., J.L., L.C.J.A. and T.d.O. helped with study design and data interpretation. R.J.L., H.T., E.W., M.G., D.M. and T.d.O. wrote the initial manuscript, which was reviewed by all authors.

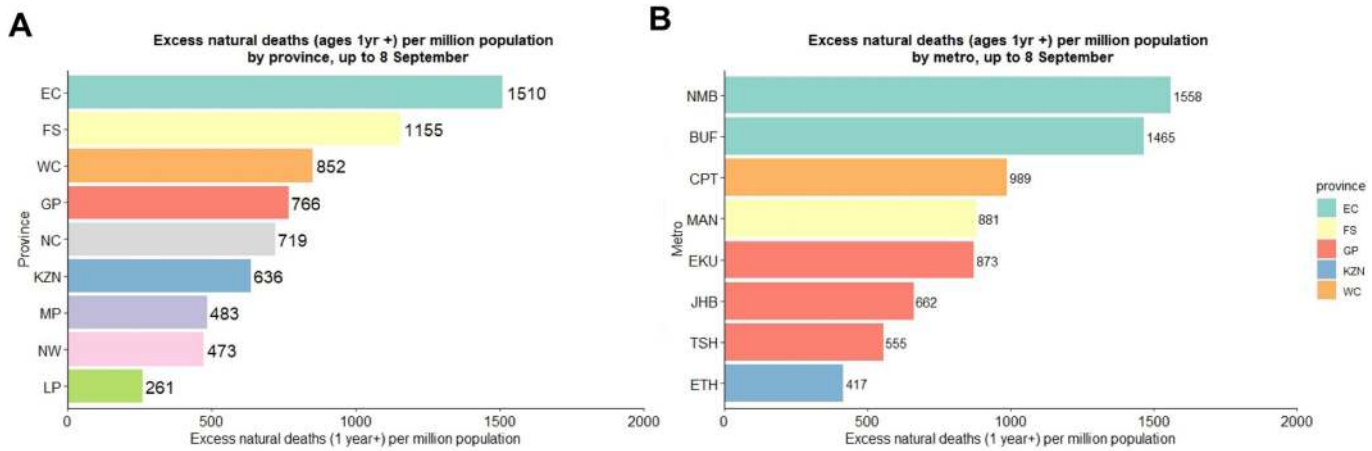
Competing interests The authors declare no competing interests.

Additional information

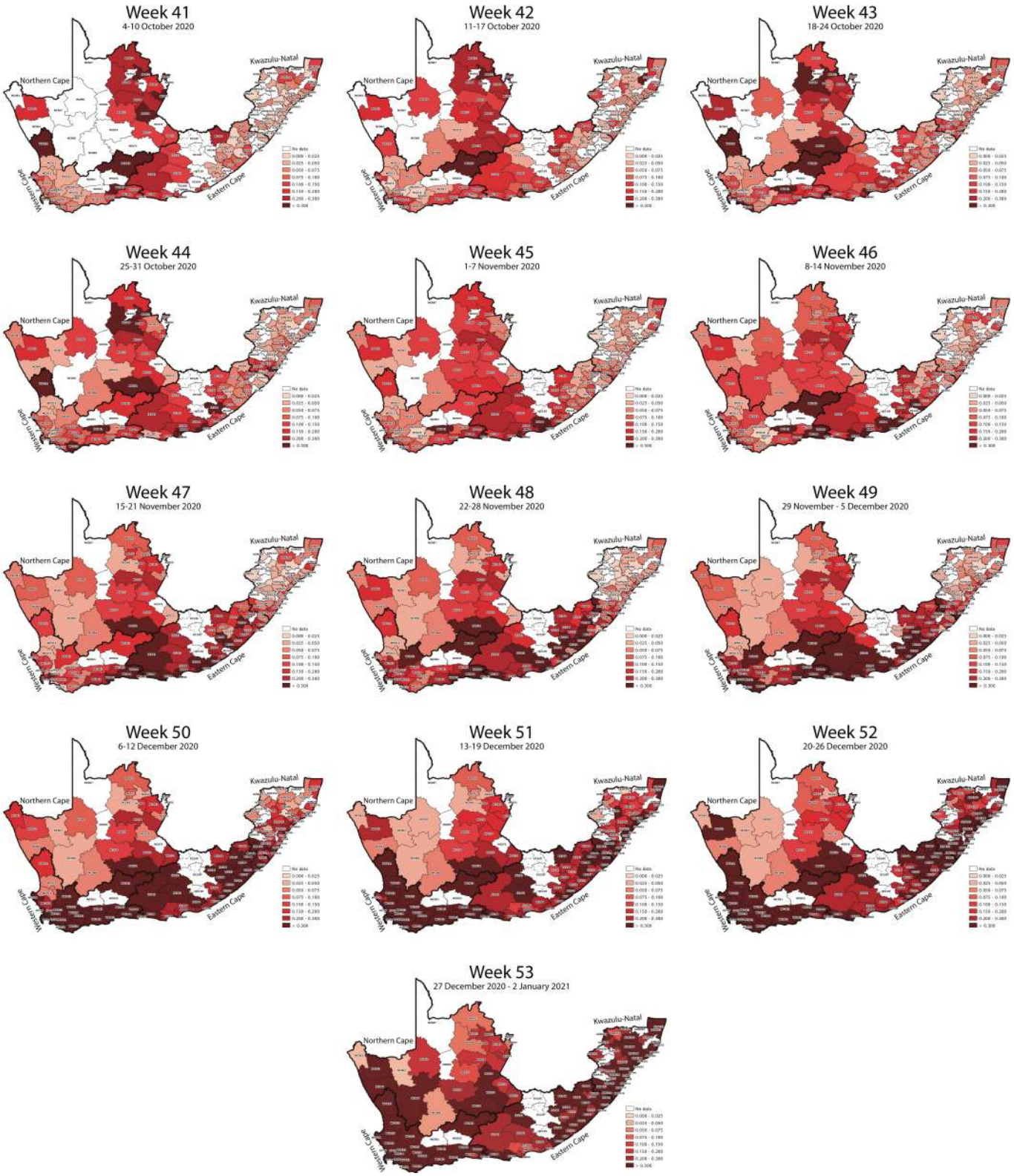
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-021-03402-9.

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

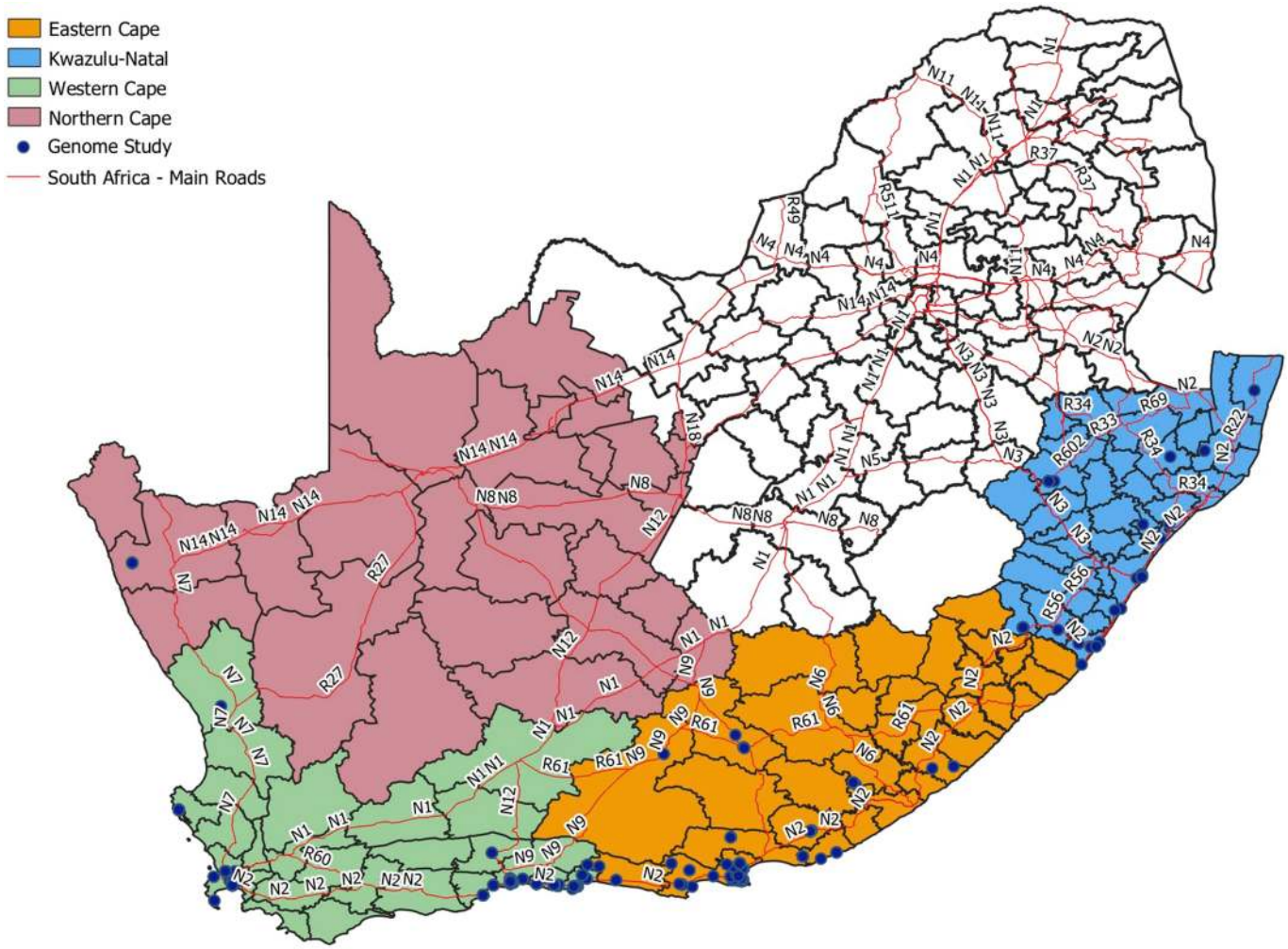
Peer review information *Nature* thanks Nathan Grubaugh, Tommy Tsan-Yuk Lam, Youchun Wang and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.



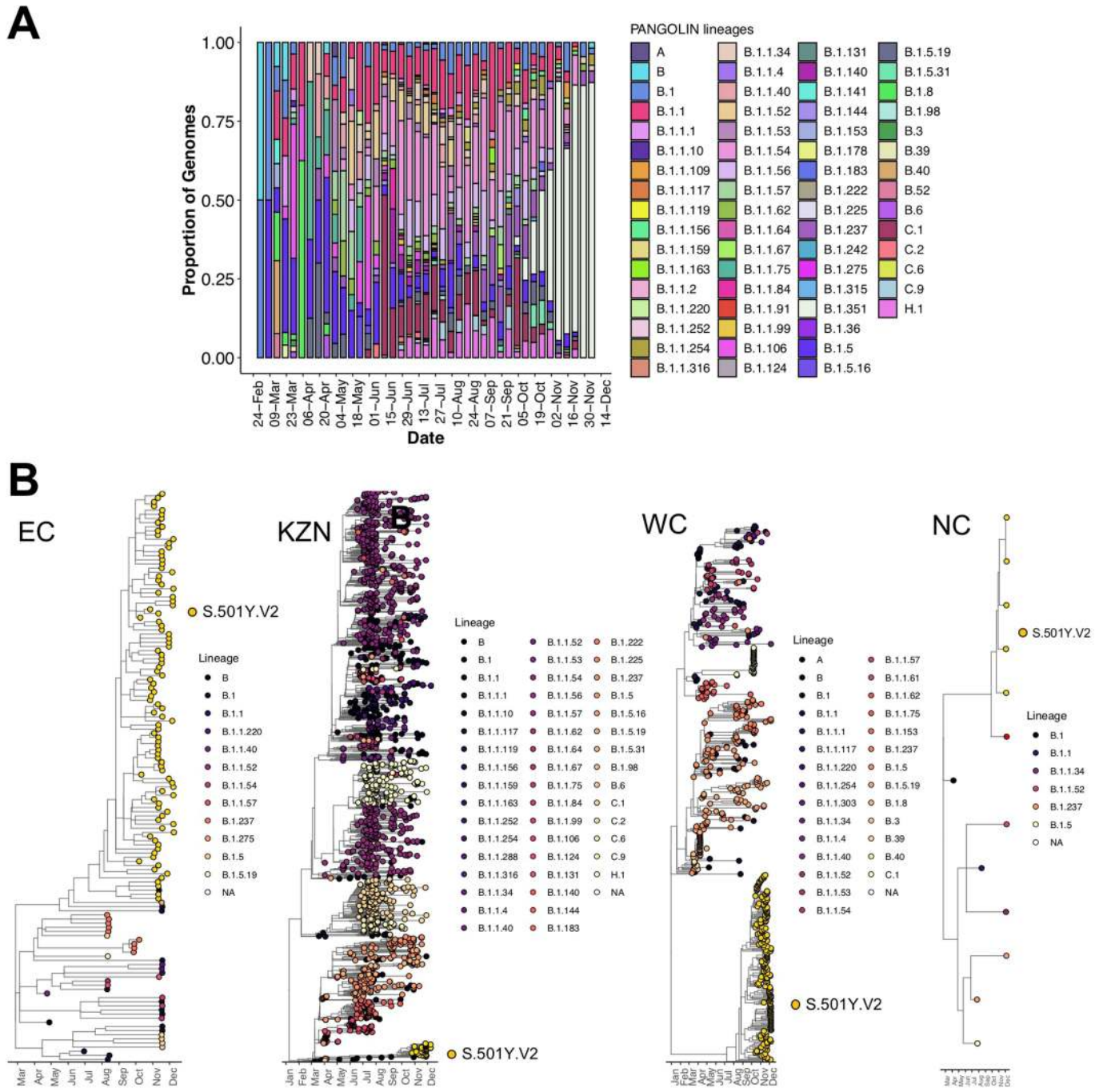
Extended Data Fig. 1 | Excess deaths per million individuals by province and metropolitan municipalities of South Africa. Data are shown for up until the week ending 8 September 2020 (immediately after the first peak of the epidemic peak). **a**, **b**, These graphs indicate the disproportionate effect 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.



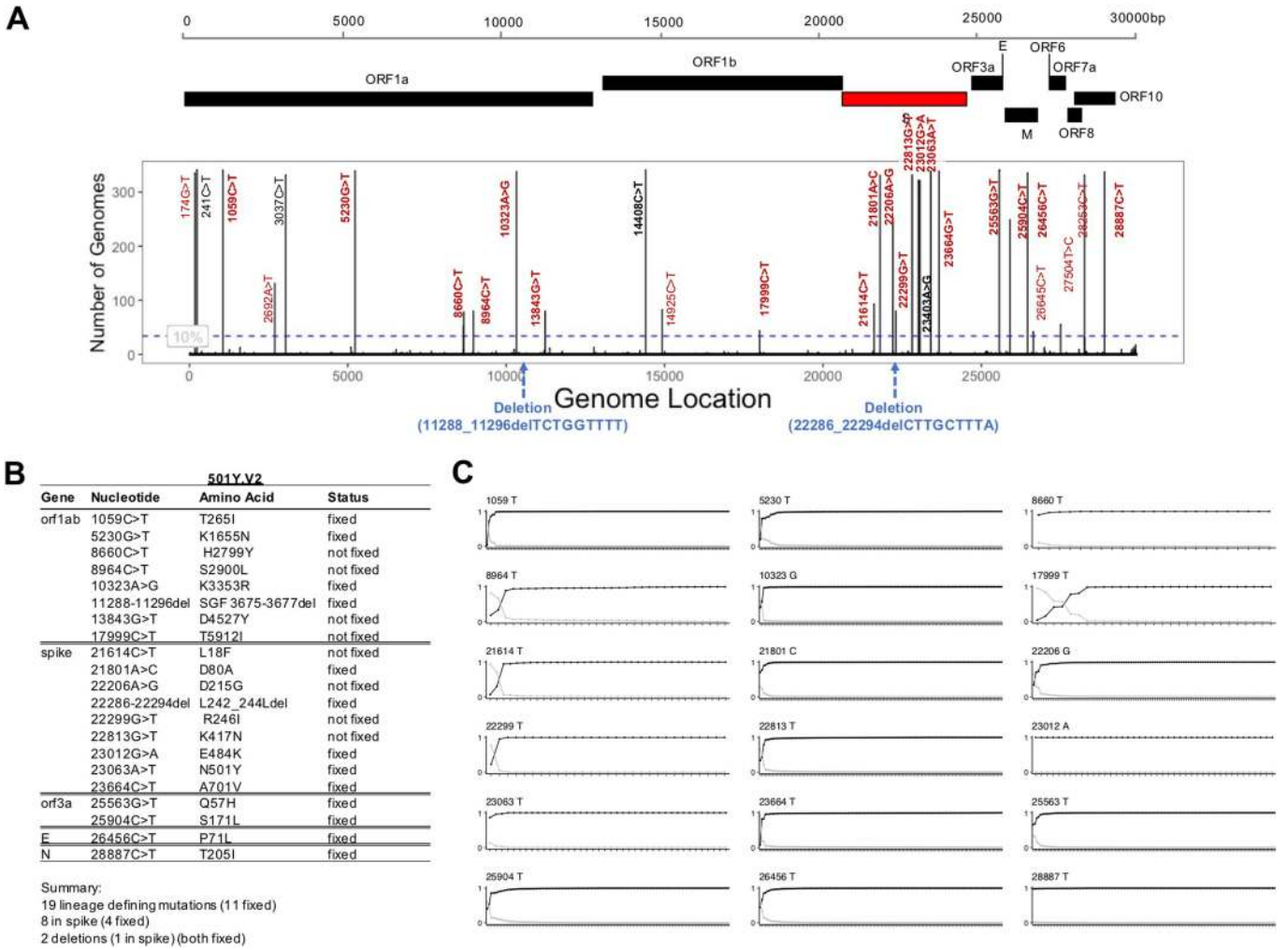
Extended Data Fig. 2 | Positivity test rates across four provinces of South Africa. Maps of the Northern Cape, Western Cape, the Eastern Cape and KwaZulu–Natal (the four provinces investigated in this Article) showing a weekly progression of SARS-CoV-2 prevalence per district, coloured by the rate of 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.



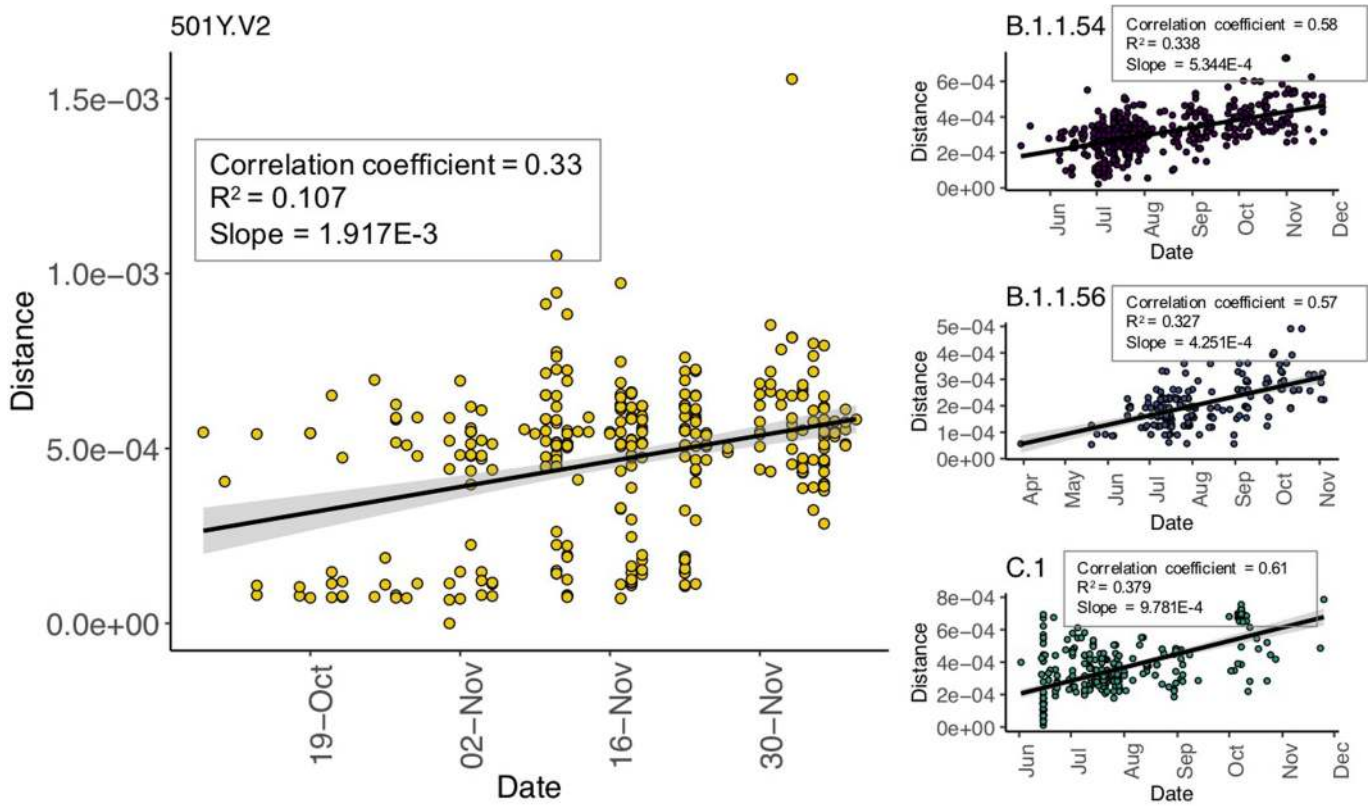
Extended Data Fig. 3 | Sampling location of 501Y.V2 genomes. A general map of South Africa, showing the sampling location of the 501Y.V2 genomes in this study (blue dots) in relation to the main road networks of the country, which hints at potential land transmission routes of this lineage along the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



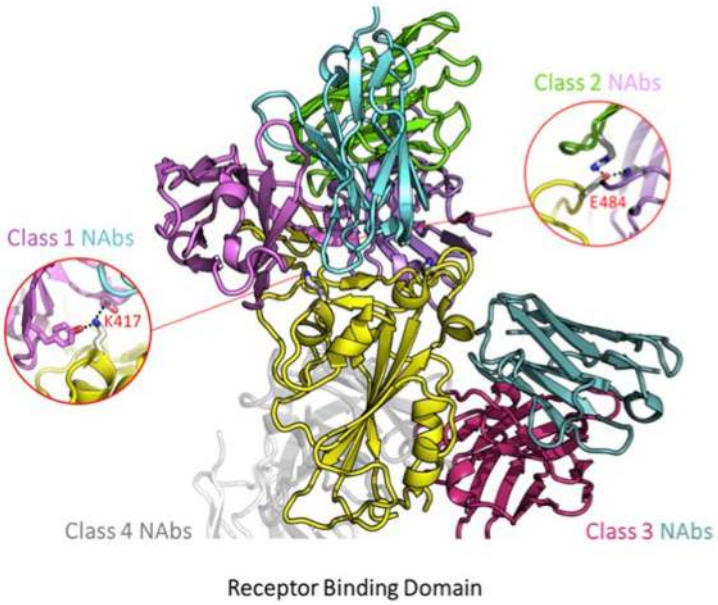
Extended Data Fig. 4 | Replacement of other lineages by the 501Y.V2 lineage. a, Progression of SARS-CoV-2 PANGOLIN lineages circulating in South Africa from March to December 2020, showing the overrepresentation of the 501Y.V2 lineage from October onwards (B.1.351, in off-white). **b**, Independent regional 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).



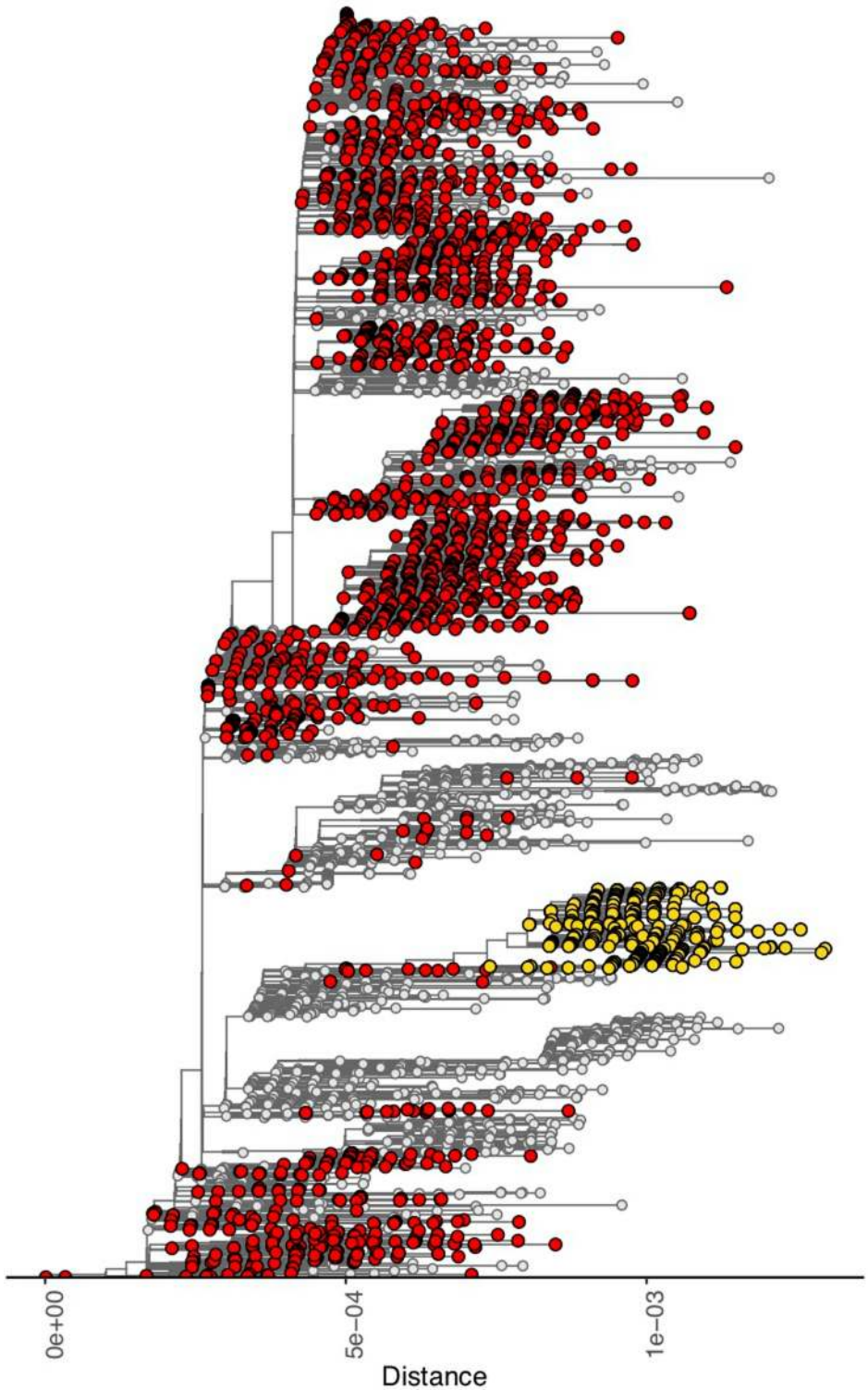
Extended Data Fig. 5 | Overview of mutations associated with the 501Y.V2 lineage. a, All nucleotide substitutions present in more than 10% of the genomes in the 501Y.V2 lineage, mapped on the SARS-CoV-2 genomic structure. Mutations present in the parent lineage (B.1) are marked in black, and mutations 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.



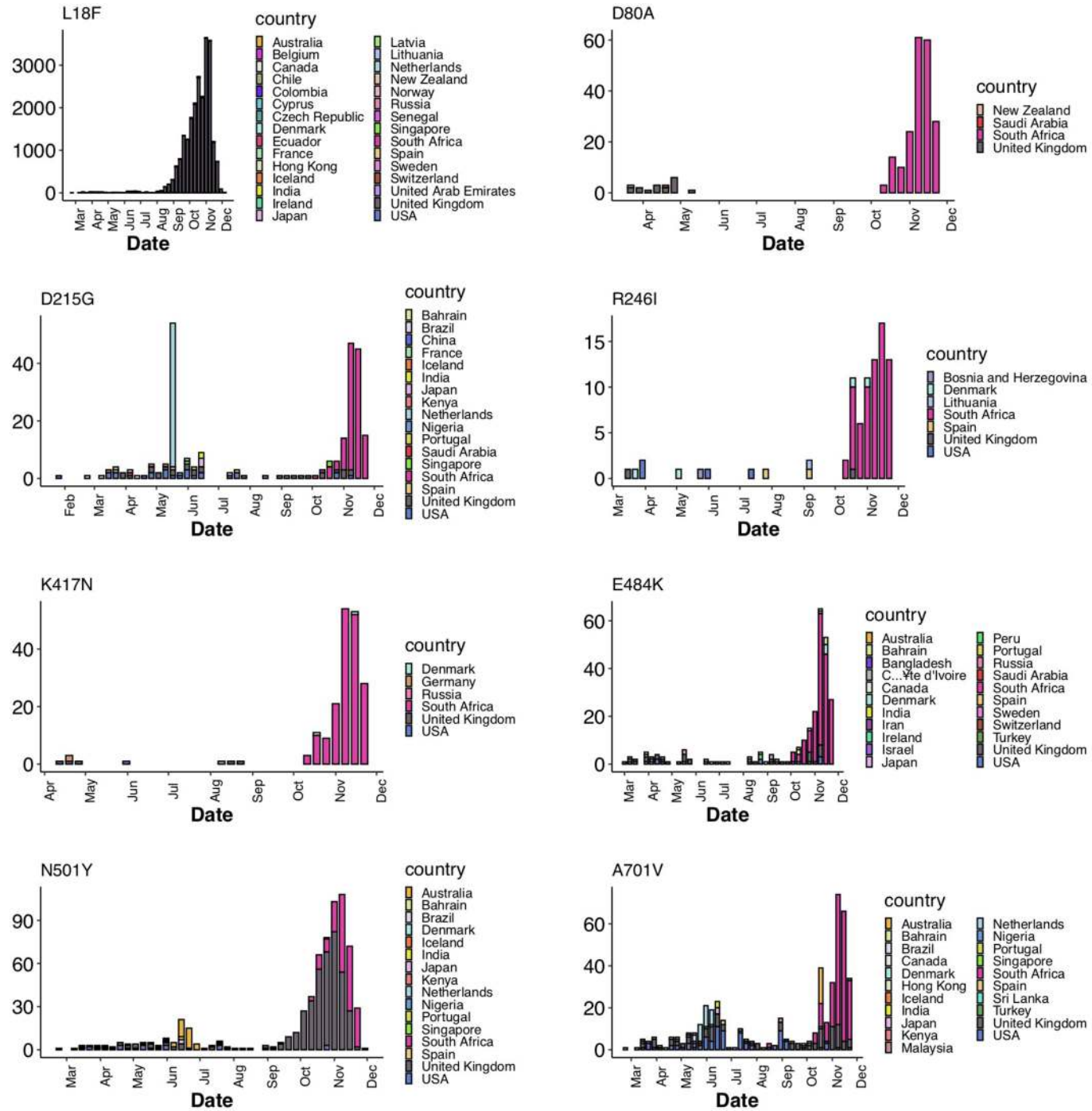
Extended Data Fig. 6 | Molecular clock signal of four main virus clusters that are spreading in South-Africa. Root-to-tip regression obtained from TempEst analysis for the 501Y.V2 lineage cluster ($n = 341$), showing a relatively strong clock-like behaviour (correlation coefficient = 0.33, $R^2 = 0.107$) 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.



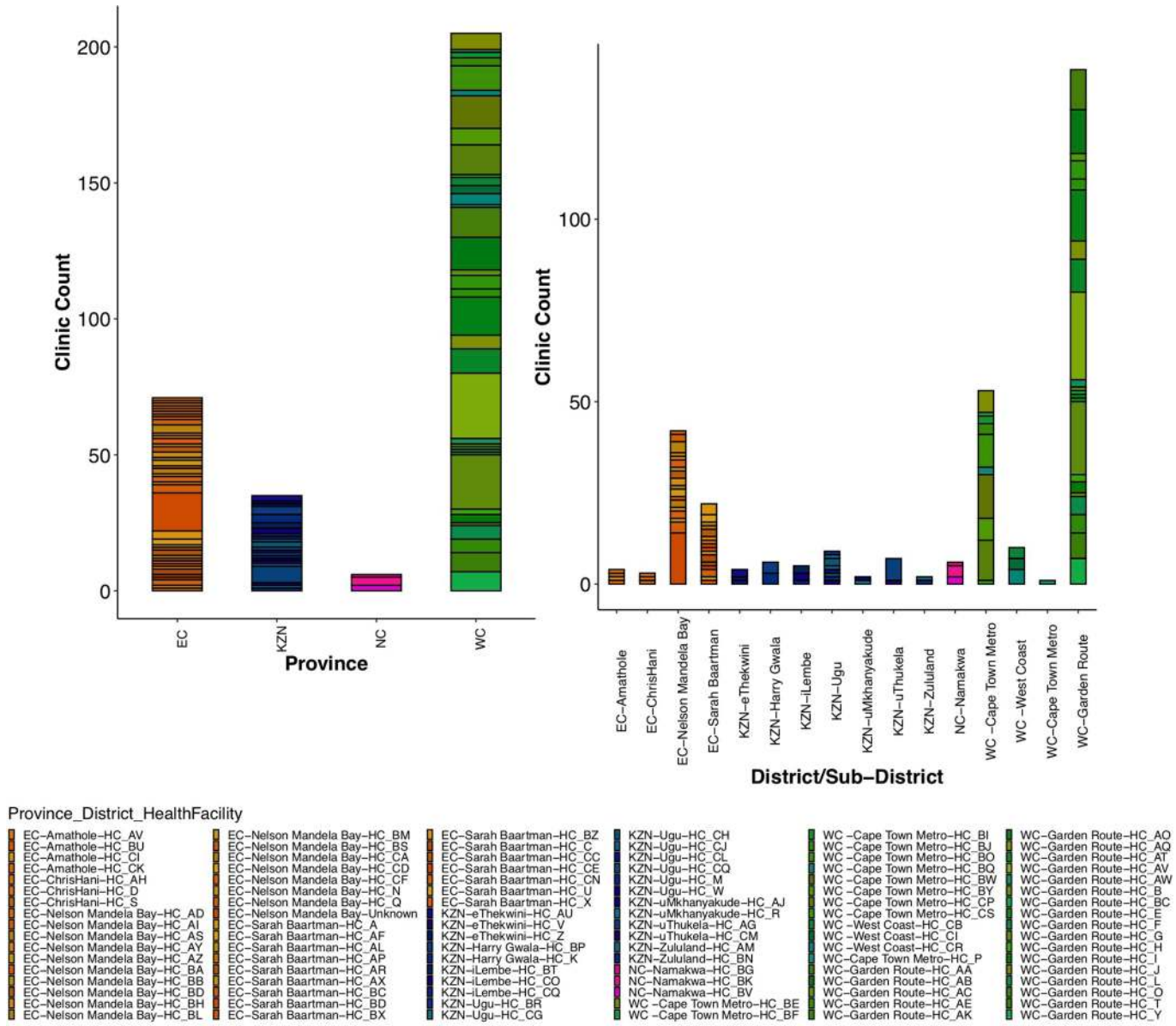
Extended Data Fig. 7 | SARS-CoV-2 RBD interactions with neutralizing antibodies. Model of the SARS-CoV-2 RBD in cartoon view (yellow), showing representative Fab domains for neutralizing antibodies (NAbs) from classes 1, 2, 3 and 4. Two zoomed-in insets show common, key interactions between the RBD residue 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.

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.