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.