










Tracing the spatial origins and spread of SARS-CoV-2 Omicron lineages in South Africa

Received: 25 September 2024

Accepted: 13 May 2025

Published online: 28 May 2025



Graeme Dor¹, Eduan Wilkinson ^{1,2}, Darren P. Martin³, Monika Moir ¹, Derek Tshiabula¹, Dikeledi Kekana⁴, Buhle Ntozini⁴, Rageema Joseph⁵, Arash Iranzadeh⁶, Martin M. Nyaga ⁷, Dominique Goedhals^{8,9}, Tongai Maponga ^{10,11}, Jean Maritz^{11,12}, Oluwakemi Laguda-Akingba ^{13,14}, Yajna Ramphal¹, Caitlin MacIntyre¹⁵, Lucious Chabuka¹, Sureshnee Pillay², Jennifer Giandhari², Cheryl Baxter ¹, Nei-yuan Hsiao ^{5,16}, Wolfgang Preiser ^{10,11}, Jinal N. Bhiman ^{4,17}, Mary-Anne Davies^{18,19}, Marietjie Venter ^{15,20}, Florette K. Treurnicht^{4,21}, Nicole Wolter ^{4,21}, Carolyn Williamson ^{16,22}, Anne von Gottberg ^{4,21,23}, Richard Lessells ², Houriiyah Tegally ^{1,24} ✉ & Tulio de Oliveira^{1,2,24} ✉

Since November 2021, five genetically distinct SARS-CoV-2 Omicron lineages (BA.1–BA.5) are believed to have emerged in southern Africa, with four (BA.1, BA.2, BA.4, and BA.5) spreading globally and collectively dominating SARS-CoV-2 diversity. In 2023, BA.2.86, a highly divergent BA.2 lineage that rose to prominence worldwide, was first detected in Israel and Denmark, but the subsequent diversity of South African sequences suggests it too emerged in the region. Using Bayesian phylogeographic inference, we reconstruct the origins and dispersal patterns of BA.1–BA.5 and BA.2.86. Our findings suggest that Gauteng province in South Africa likely played a key role in the emergence and/or amplification of multiple Omicron lineages, though regions with limited sampling may have also contributed. The challenge of precisely tracing these origins highlights the need for broader genomic surveillance across the region to strengthen early detection, track viral evolution, and improve preparedness for future threats.

As with most countries around the world, South Africa has experienced multiple severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) infection waves since the onset of the COVID-19 pandemic. South Africa's infection waves have had a global impact, with two of the five variants of concern (Beta and Omicron) first detected in the region^{1–3}. Despite its relative remoteness and the fact that it is home to only 0.7% of the world's population⁴, South Africa has had a disproportionately large impact on SARS-CoV-2 evolution and spread.

The rapid resurgence of infections that followed the discovery of the Omicron variant in November 2021 became South Africa's fourth wave. This wave was particularly unusual at the time in that it was

driven by three distinct lineages - designated BA.1, BA.2, and BA.3 - that were more divergent from one another than were the first four VOCs (Alpha, Beta, Gamma, and Delta) from one another. As inferred from genomic surveillance data, BA.1 accounted for the largest proportion of cases in the first part of the wave, followed by BA.2 in the latter part, with BA.3 accounting for a relatively small number of cases throughout this period². BA.1 and BA.2 went on to dominate global SARS-CoV-2 infections in the first half of 2022⁵.

Concomitant to the rise of these two lineages globally, another two distinct Omicron lineages, BA.4 and BA.5 were first detected in South Africa in approximately January and February 2022, respectively, and together drove a fifth wave of infections across the country².

A full list of affiliations appears at the end of the paper. ✉ e-mail: houriiyah@sun.ac.za; tulio@sun.ac.za

As with BA.1 and BA.2, BA.4 and BA.5 spread throughout the world to collectively dominate global SARS-CoV-2 infections by June 2022⁵.

Following the emergence in the USA and Singapore of XBB (a BA.2 recombinant), estimated to have emerged in July 2022⁶, and rising to dominance in South Africa and elsewhere in the world by February 2023, it seemed as though the stream of divergent new southern African Omicron lineages had ceased.

However, the discovery in July 2023 of BA.2.86, a new highly diverged BA.2 lineage first detected in Israel and Denmark⁷ indicated that this might not be the case. Reinvigorated genomic surveillance efforts in South Africa and other countries following the discovery of BA.2.86 soon revealed that the likely geographical origin of this lineage was the same region of southern Africa where BA.1, BA.2, BA.3, BA.4, and BA.5 likely emerged⁸. As with previous BA lineages, BA.2.86 also subsequently rose to dominance in South Africa and elsewhere around the world. Its descendant lineages remained dominant throughout 2024, underscoring the continued significance of BA lineages in the global epidemiological landscape.

Each of the first Omicron sequences carried over 50 mutations when compared to the ancestral SARS-CoV-2 genome, with as many as 30 mutations occurring in the spike glycoprotein which influenced antibody neutralisation and spike function^{3,9}. The extreme divergence of these first Omicron sequences relative to previous variants and the fact that multiple genetically distinct “BA” lineages emerged simultaneously, prompted intense speculation on how and where these lineages originated. The proposed explanations for the origins of the Omicron BA lineages include; prolonged evolution in one or more chronically infected individuals, possibly with uncontrolled HIV; undetected community transmission in an under-surveilled region of southern Africa; or spillover into an unidentified animal host followed by selection and spill-back into the general human population^{10–12}.

The shifting immunological landscape of the global human host population is likely to remain the key determinant of which newly arising SARS-CoV-2 lineages will have sufficient transmission potential to persist and spread throughout the world. Given the centrality of South Africa in the emergence of the Beta VOC and potentially all of the most divergent Omicron lineages that have achieved global dominance, it is sensible to more closely inspect the early variant dynamics in the country. From the perspective of understanding how and why southern African lineages have been so successful, it is especially important that we analyse the early transmission dynamics and geographic origins of these lineages. From a global pandemic preparedness and response perspective, this could illuminate potentially shared spatial origins and patterns of spread of these lineages that will enhance our understanding of how future southern African lineages might emerge and disseminate globally.

Here we aimed to reconstruct the early dispersal routes of Omicron lineages within South Africa. Before focusing on intra-country transmission dynamics, we first confirm that southern Africa is the inferred region of origin for these lineages by conducting a global ‘migration’ analysis. We employ spatial mapping and Bayesian phylogeographic analysis to explore the origins and dispersal patterns of BA.1 through BA.5, and BA.2.86, tracing each of the lineages back to the likely location within southern Africa where the most recent common ancestors of sampled viruses within each of these lineages occurred. The analysis suggests that the north-eastern region of South Africa, particularly Gauteng province, may function not only as a focal point for viral transmission but potentially as an area of initial emergence.

Results

Validation of global origins of SARS-CoV-2 Omicron BA lineages

To confirm the global origins and early dispersal patterns of the respective Omicron BA lineages, we mapped the discrete country

locations and inferred dispersal events from our ‘migration’ analysis. Ancestral state reconstruction of the early spread suggested South Africa as the origin for lineages BA.1, BA.2, BA.3, BA.4, and BA.2.86, while Eswatini, a neighboring country of South Africa, was the inferred origin of BA.5 (Fig. 1). Following their emergence in southern Africa, global dissemination was frequently inferred from high-connectivity regions, particularly parts of Europe and North America.

Genomic surveillance

We then focused on discovering what the locations of the earliest sampled sequences of the different lineages (those obtained during the first 8 weeks following their first detection) might reveal about both the geographical region(s) where the lineages first emerged, and the pathways of dispersal that occurred shortly after their emergence. We find that the earliest detected sequences across the respective lineages come predominantly from the north-eastern region of South Africa, yet there is no single common province of origin (Fig. 2b). Moreover, there is no common expansion or pattern of spread evident in the weeks following their first detection. It is important to note that testing rates in South Africa varied at a provincial level over this period, with Gauteng and the Western Cape provinces reporting the highest testing rates¹³. However, these differences were not substantial, particularly during low-incidence periods, and depicted a consistent trend following the first detection of the respective lineages. In addition, sequencing efforts were not evenly distributed across all provinces, with the Western Cape, Gauteng and KwaZulu-Natal provinces collectively accounting for over 66% of the total sequenced samples. That being said, Gauteng and KwaZulu-Natal provinces both had relatively low sequence to case proportions at 1.1%, compared to 2.1% for the Western Cape (Supplementary Fig. 1).

Phylogenetic and phylogeographic analyses

As has been reported elsewhere², the time-resolved maximum clade credibility (MCC) phylogeny confirmed that each of the Omicron BA lineages forms a phylogenetically distinct clade within the Omicron tree, each without clear basal progenitors (Fig. 3a). Specifically, BA.4 and BA.5 appear to be sister clades, with a more recent common ancestor than to the other four main Omicron lineages. BA.2.86 appears to branch from the rest of the tree as a sister clade of the BA.4/5 clade. All three of the latter appear to have a common ancestor sister to the main BA.2 clade, which itself is inferred to be a sister to BA.1 and BA.3. However, each of these lineages display genetically diverse characteristics and originated at varying time points. Specifically, the 95% highest posterior density (HPD) intervals suggest that BA.1 likely originated between mid-June and mid-September 2021, followed closely by BA.3 from early July to early October and BA.2 from late July to mid-October within the same year. Thereafter, BA.4 and BA.5 were estimated to have emerged between late November to mid-December 2021, and late December 2021 to late January 2022, respectively. These time to most recent common ancestor (tMRCA) estimates align well with those previously inferred^{2,3,8} although with slight differences potentially due to sampling strategies and sample size. BA.2.86 is estimated to have appeared between mid-January and late April 2023.

Despite the respective BA lineages forming phylogenetically distinct clades, phylogeographic inference revealed four of the six (BA.1, BA.2, BA.4, and BA.2.86) are inferred to have originated within Gauteng province before spreading to other regions within South Africa (Fig. 3b). To the east of Gauteng, Mpumalanga province and the country of Eswatini were inferred as the origins of BA.3 and BA.5 respectively. The temporal reconstruction reveals complex spatial patterns of lineage spread following emergence with no uniform trend across the BA lineages, however, early dispersal trends suggest initial expansion from Gauteng.

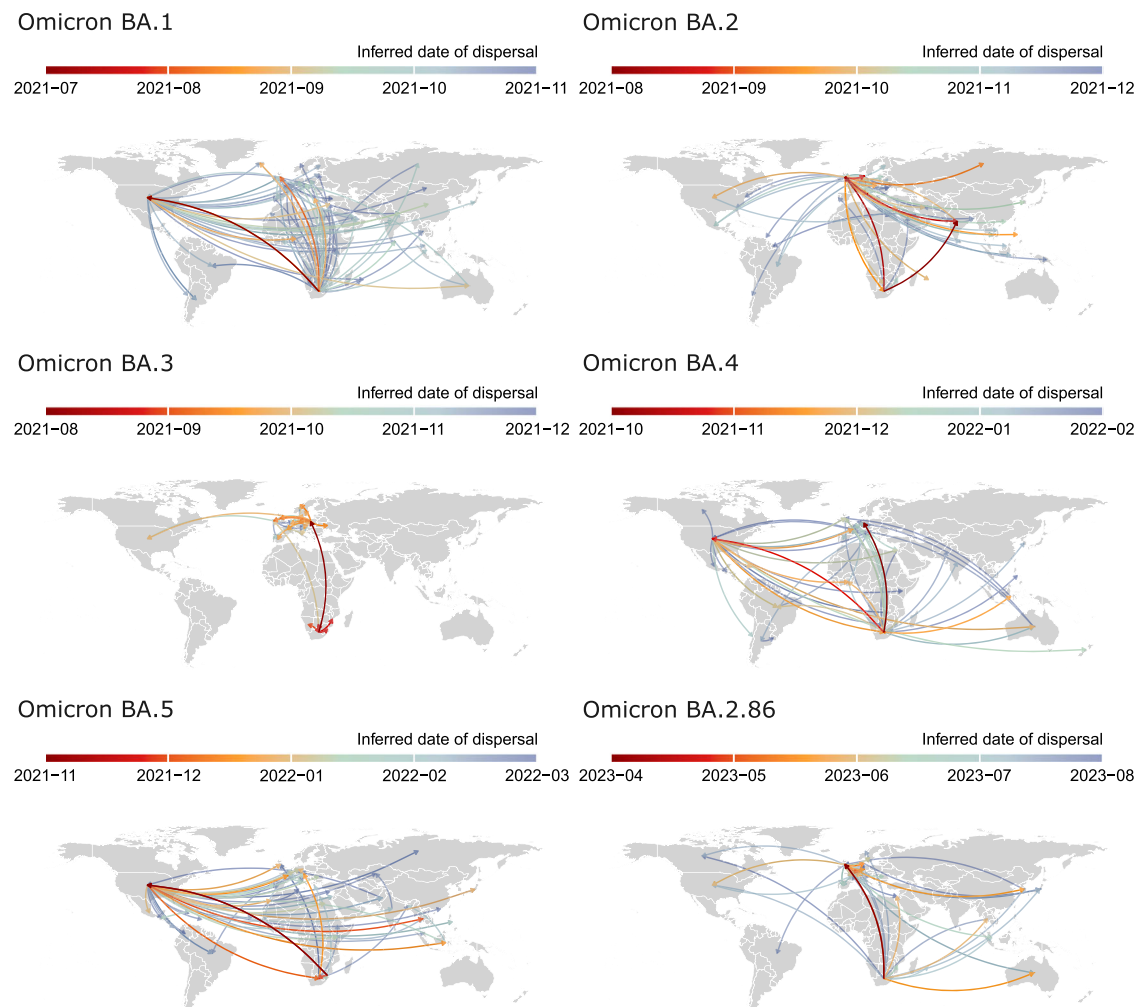


Fig. 1 | Inferred global origins and spatio-temporal dispersal patterns of Omicron BA lineages. The first 100 inferred movements are shown, with dark red lines depicting the earliest movement, following a counter-clockwise trajectory.

Overlaying the inferred most recent common ancestor (MRCA) locations and associated 95% HPD regions of all six BA lineages clearly highlights the influence of the northeastern region of South Africa, particularly Gauteng, and its role as both the region of origin and early dispersal point within South Africa (Fig. 4). The displacement of the inferred MRCA locations, as a measure of uncertainty, cluster over a relatively small area given the extent of sampling locations, while the 95% HPD intervals indicate the influence of Gauteng across all lineages.

Additionally, South Africa is shown to contain the earliest inferred MRCA locations across southern Africa for five of the six lineages, with the inferred spread propagating from South Africa to neighbouring countries for all six lineages (Supplementary Fig. 2). This analysis was performed to confirm whether there were potential introductions from neighbouring countries, considering Gauteng's role as a prominent transport hub.

Discussion

Here, we attempted to uncover the early spatio-temporal dispersal dynamics in South Africa of the six most globally significant SARS-CoV-2 Omicron lineages. Our findings position Gauteng province as the most likely geographical location of the MRCA of BA.1, BA.2, BA.4, and BA.2.86. Additionally, discrete phylogeographic analyses confirmed South Africa as the origin of these lineages within southern Africa, although this finding is potentially constrained by the limited genomic surveillance in the region.

While these analyses do not conclusively support the hypothesis that all of these lineages arose in the exact same area in South Africa, they are entirely consistent with this hypothesis. Specifically, our study indicates that, the northeastern region of South Africa, particularly Gauteng province and the region extending eastward is either (1) the location of a pool of divergent viral lineages that, while confined for the most part to evolving outside the context of general transmission within the human population (such as within chronically infected individuals), occasionally spills over into the general human population; or (2) is the primary amplification and dissemination hub for such lineages whenever these arise in neighbouring geographical regions of southern Africa.

The finding of Gauteng province as the most likely geographical location of the MRCA of four of these highly consequential BA lineages is, however, not strong evidence for a single discrete source for all the lineages; such as a single chronically SARS-CoV-2 infected individual or an isolated small community of chronically infected individuals. Strong support for such a source would have been if the MRCA of all six of the lineages had originated in a less populous and globally connected province (such as Mpumalanga) and had then all tended to move to Gauteng within a few weeks, which was only observed for two of the six lineages. Similarities in the dispersal dynamics of the six lineages might have only been detectable if more sequence data obtained closer to the locations and times when these lineages started circulating in the general population were available to us: locations that, unlike Gauteng, were very sparsely sampled in the months

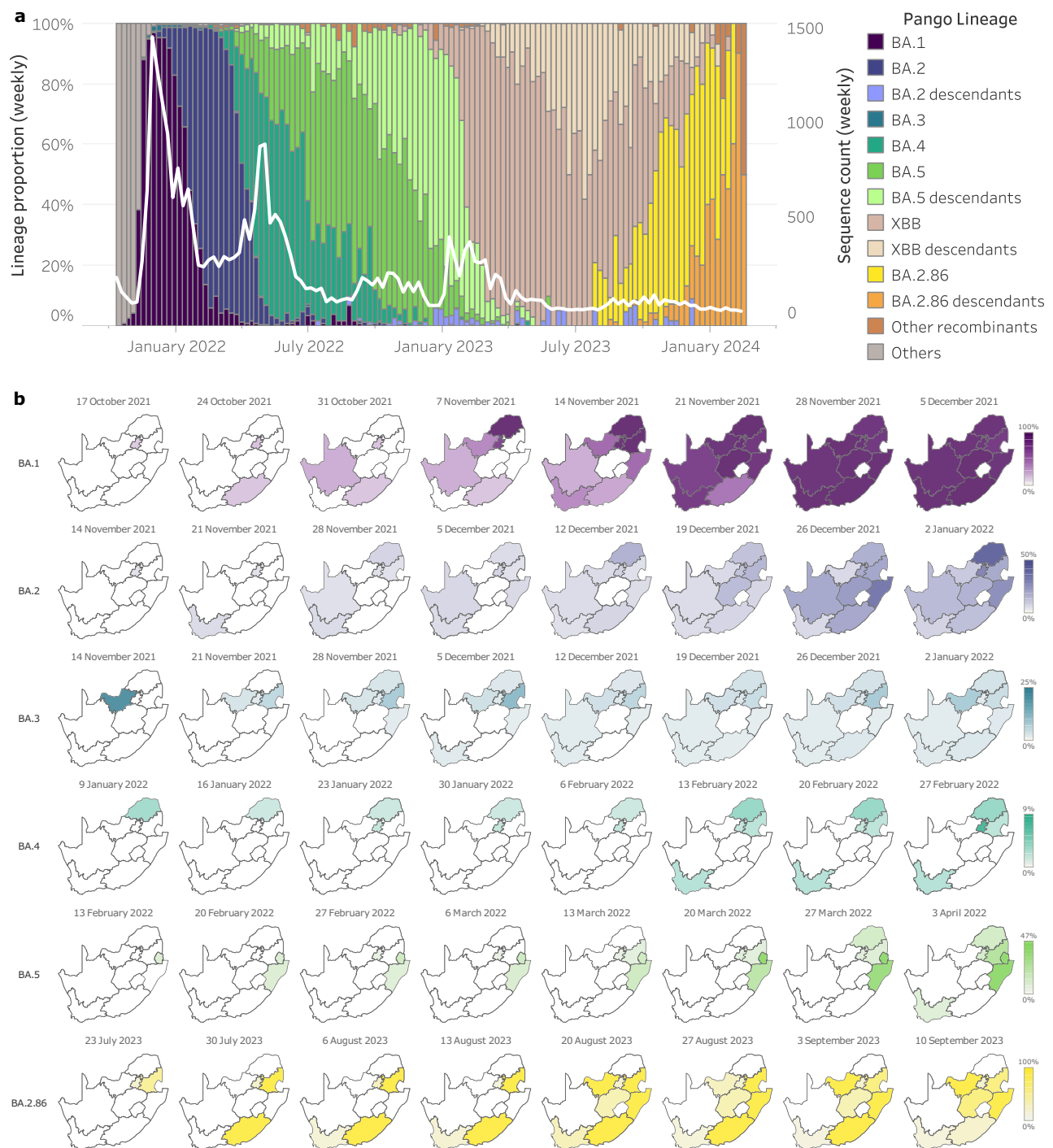


Fig. 2 | Distribution of Omicron BA lineages identified through genomic surveillance in South Africa. a The primary axis displays a stacked bar chart of the weekly proportion of lineages from all South African sequences, as assigned by Pangolin and extracted from GISAID, since the initial introduction of Omicron. The

secondary axis highlights the weekly count of sequences in the form of a line graph. **b** Time series maps displaying the prevalence of respective BA lineages at a province level, highlighting their first occurrence and spread thereafter. Colour gradients reflect the maximum prevalence for each lineage, respectively.

preceding the detection of these lineages. This is likely due to a shift to passive case finding and low testing rates during periods of low incidence. If we had inferred multiple non-overlapping geographical origins together with marked differences in the early dispersal dynamics of the lineages, this would have been consistent with the cryptic epidemiological confinement and evolution of the progenitors of the six lineages within a larger, more dispersed pool of individuals.

To increase the probability of detecting newly emerged divergent SARS-CoV-2 lineages closer to their origin locations, along

with their early transmission dynamics, it is imperative that South Africa sustains SARS-CoV-2 genomic surveillance in Gauteng and enhances surveillance in other north-eastern parts of the country. Additionally, fostering the establishment of stronger genomic surveillance in neighbouring southern African countries - specifically Eswatini, Lesotho, Zimbabwe, Mozambique, and Botswana - is critical. These efforts should focus not only on detecting new variants, but also on building a comprehensive regional surveillance network that could provide invaluable insights into the spread of

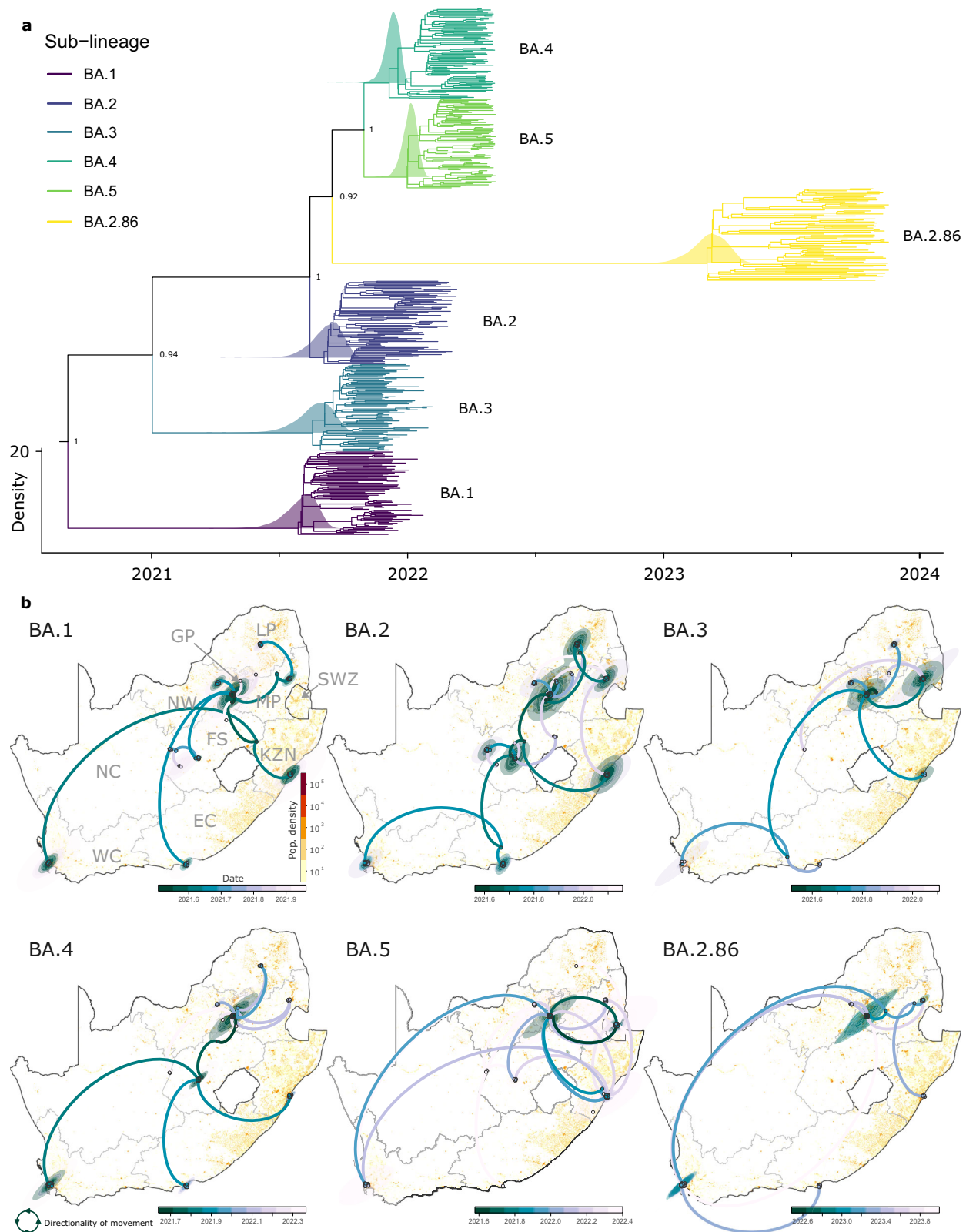


Fig. 3 | Molecular evolution and profile of Omicron BA lineages. a Time-Resolved MCC phylogenetic tree with marginal density plots unveiling the temporal dynamics and evolutionary relationships among BA lineages. **b** Phylogeographic reconstruction of BA lineages depicting the inferred origin and early

spatiotemporal dispersal patterns, displayed on a population density basemap (source: <https://hub.worldpop.org/>). The movement between locations follows a counter-clockwise direction along the arcs.

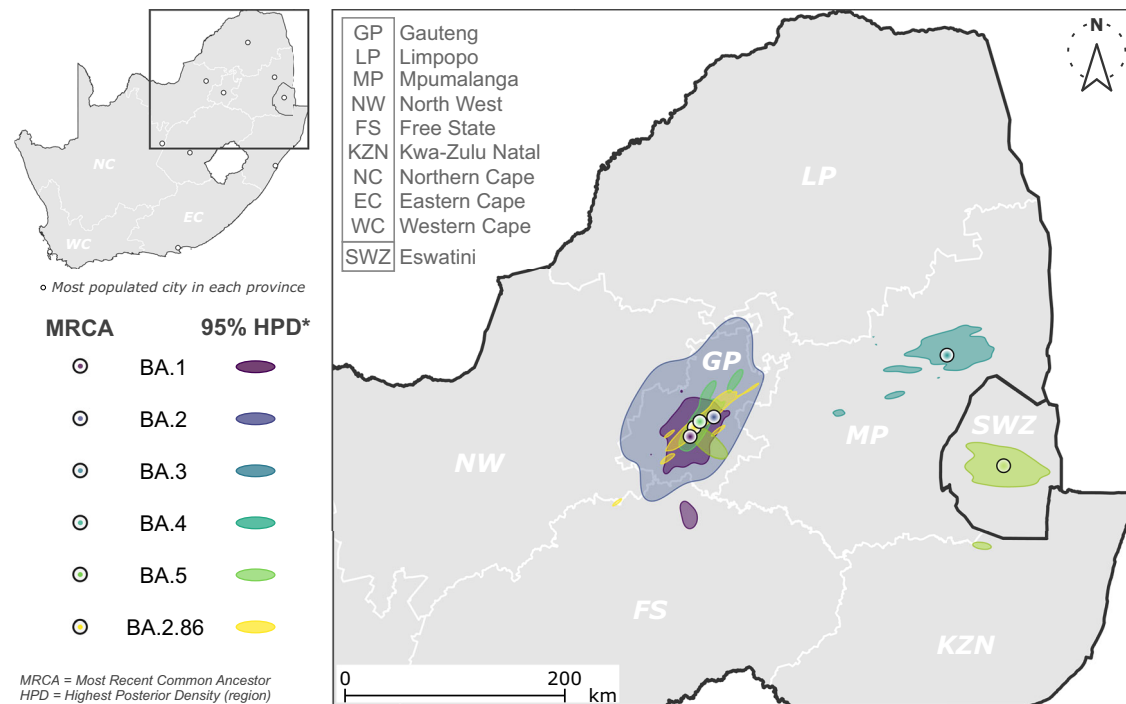


Fig. 4 | Inferred origins of Omicron BA lineages. Most Recent Common Ancestor (MRCA) inferred locations are represented by points and bounded by corresponding 95% Highest Posterior Density (HPD), indicating the spatial uncertainty associated with each MRCA location estimate.

future divergent lineages both locally and internationally. However, recognising the decreased public health impact of COVID-19, these recommendations should be balanced with the current capacity for implementation and future prioritisation for pandemic preparedness.

Although our study offers evidence supporting the hypothesis that the six major Omicron lineages all originated somewhere in the north-eastern parts of South Africa, potentially in discrete locations or reservoirs surrounding Gauteng province, the data on which our study is based have certain limitations. Notably, lower levels of testing and sequencing in neighbouring countries limit our ability to pinpoint the origins of these lineages with high accuracy, as Gauteng may just serve as a primary amplification and dissemination zone, potentially obscuring the true geographic sources. Likewise, in South Africa, the uneven distribution of testing and sequencing efforts across provinces results in sequence data that could disproportionately highlight certain regions. Moreover, sequencing prioritisation for samples that exhibited spike gene target failure (SGTF), as a result of a spike 69-70 deletion (a deletion found in BA.1, BA.4, BA.5 and BA.2.86), relied on the uneven distribution of these instruments. This likely introduced additional sampling bias, impacting the precision of our insights into the virus' origins. Quantifying and accounting for the effect that sequencing prioritisation had on the analysis could help refine the precision of our estimates and inferences. Further, systematic and standardised inclusion of finer-resolution geographic data (at finer resolution than the district-level) into sequence metadata could refine origin determination, especially if the lineages ultimately originate from a single source within the north-eastern part of South Africa.

The consistent and ongoing discovery of new divergent Omicron lineages in southern Africa - the latest being that of the BA.2.87.1 lineage¹⁴—implies this region is more likely than any in the world to be the origin of future highly diverged SARS-CoV-2 lineages of global relevance. It is likely an important factor that East and southern Africa is home to approximately half of the people in the world who are currently living with HIV. Uncontrolled HIV infections—

approximately 5 million of which occur in East and southern Africa¹⁵—are expected to cause a type of immune dysfunction that predisposes individuals to experience persistent SARS-CoV-2 infections. These persistent infections are believed to foster the evolution of highly diverged neutralising antibody resistant variants such as the Beta VOC and Omicron BA lineages^{16–18}. Given the potential global consequences of SARS-CoV-2 evolution in the context of unmanaged HIV infections, we strongly recommend renewing world-wide efforts to document and treat as many uncontrolled HIV infections as possible, as well as understand SARS-CoV-2 evolution in other types of immunocompromised conditions.

Methods

Genomic data acquisition and lineage classification

We queried the complete global dataset of 16.6 million whole genome sequences of SARS-CoV-2, along with related metadata, available on the GISAID database (<https://gisaid.org>) as of 28 February 2024. Of these, approximately 8 million sequences, classified using the 'Phylogenetic Assignment of Named Global Outbreak Lineages' (PANGOLIN) method¹⁹, belong to the BA lineages of the Omicron variant, which served as the focus of our study. To facilitate downstream analysis, we grouped lineages into their broader classification (BA.1 through BA.5, and BA.2.86) while distinguishing all descendant and recombinant Omicron lineages according to their respective pango-lineage designation, as outlined in the sequence designation list available on the PANGO GitHub repository (github.com/cov-lineages/pango-designation/milestones).

Sequence alignment and metadata extraction related to sequence quality were performed using Nextstrain's Nextclade CL^{20,21}. The resulting metadata was used to filter the dataset, retaining only sequences with complete collection dates and high coverage (>90% genome coverage) that passed Nextclade's quality check. Additionally, sequences with clade-level inconsistencies between GISAID's PANGOLIN assignment and Nextclade's PANGO lineage designation were excluded. For a detailed breakdown of sequence counts by BA lineage, refer to Supplementary Table 1.

Genomic sampling

Given the sensitivity of ancestral state reconstruction methods to sampling biases²², we employed a subsampling strategy to generate representative datasets for each lineage. For this purpose, we used *subsampler* (<https://github.com/andersonbrito/subsampler>), a pipeline designed to systematically subsample genomic data based on epidemiological time series data. This approach allocated sequences per location, by country and province for global and subnational-level investigations, respectively, in proportion to reported case counts accounting for geographic and epidemiologic diversity. To ensure temporal representativeness, we applied a weekly sampling criteria, aggregating reported cases and sequences on a weekly basis. A crude estimate of the number of weekly cases by lineage was derived by multiplying the number of reported cases by the proportion of sampled genomes designated to the respective BA lineages, using metadata from GISAID. This systematic subsampling approach ensured that the datasets accurately represented the temporal and proportional circulation of the lineages, aligning with the underlying epidemiological trends over time. Specific considerations for subsampling, according to the objectives of each analysis, are detailed below.

Global ‘mugration’ analysis

To establish and confirm the broader global origins of the respective BA lineages, we subsampled 5000 globally representative sequences per lineage using the *subsampler* tool based on available case count data (scaled by population) from the COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (<https://github.com/CSSEGISandData/COVID-19>). To infer the early movements and global origins of the respective lineages, we performed a ‘mugration’ analysis with TreeTime²³, modelling transitions between discrete geographic states at the country level. Maximum-likelihood tree topologies for each lineage were inferred using IQ-TREE version 2.3.6²⁴ under a GTR model of nucleotide substitution, with branch support estimated using ultrafast bootstrap approximation (1000 replicates) to ensure statistical robustness²⁵. Temporal molecular clock signals were evaluated using the clock functionality of TreeTime, and outlier sequences that violated the strict molecular clock assumption were identified and removed using the Ape package in R. The refined trees were then transformed into time-scaled phylogenies in TreeTime. Discrete country locations were mapped to tips, and ancestral country locations were inferred for internal nodes under a GTR model using the ‘mugration’ extension of TreeTime. Finally, a custom python script (https://github.com/CERIKRISP/SARS_CoV_2_VOC_dissemination) was used to derive the number of state changes over the span of the tree. The first 100 inferred transition states were visualised using gmap and ggplot2 in R on a global map with country-level boundaries from rnaturalearth, providing a comprehensive overview of global spatial dynamics during the early dispersal phase of each lineage.

Phylogenetic investigation

To investigate the evolutionary dynamics of the six lineages in the region associated with their emergence, we applied the previously described subsampling strategy, utilising provincial-level case data from COVID-19 South Africa (<https://www.covid19sa.org/provincial-breakdown>). For each lineage, we selected 200 sequences from the early phase of each wave, up to the peak in cases inferred from genomic data, while ensuring proportional representation across provinces and retaining the first 20 identified sequences per lineage. Preliminary maximum likelihood trees for each lineage, as well as for the combined set of lineages, were generated using IQ-TREE version 2.3.6²⁴. We then inspected these trees using TempEst version 1.5.3²⁶ to assess the strength of temporal signal and to examine the root-to-tip distance versus genetic regressions for outlier identification²⁷,

with further validation performed using TreeTime²³. Linear regression of root-to-tip genetic distances against sampling dates indicated that the sequences had been evolving in a strong clock-like manner (correlation coefficient = 0.98, $R^2 = 0.95$), warranting the use of a strict clock model for further Bayesian phylogenetic analyses. We then estimated time-calibrated phylogenies using BEAST version 10.5.0²⁸, applying a strict molecular clock model, the GTR + I + G nucleotide substitution model, and an exponential growth coalescent model²⁹. Model parameter selection was validated through marginal likelihood estimation (MLE) across a range of parameter combinations (Supplementary Table 2). We performed Markov Chain Monte Carlo (MCMC) sampling of the model parameter space with runs of 100 million states each, sampling every 10,000 steps. Proper mixing of the MCMC was verified using Tracer version 1.7.3³⁰. MCC trees were annotated with model parameters (and their 95% credibility intervals) and used to summarise the sampled MCMC chain states using TreeAnnotator after discarding 10% of samples as burn-in. The MCC tree was visualised using the R packages ggplot2 and ggtree^{31,32}.

Phylogeographic inference

To obtain information on the geographical origins of the six lineages within southern Africa we inferred the geographical locations of ancestors of the sampled sequences: information that, given the geographical coordinates where sequences were sampled, is expected to be encoded in the phylogenetic relationships of the sampled sequences. We constructed a time-resolved MCC phylogenetic tree to determine when the ancestral sequences represented by phylogenetic tree nodes existed: i.e., we determined the tMRCA of every pair of sampled sequences represented in the tree (Fig. 3a). We then performed a phylogeographic analysis, to identify both the most probable geographical location coordinates of the MRCAs of each of the BA.1, BA.2, BA.3, BA.4, BA.5 and BA.2.86 lineages and the 95% HPD regions that bounded these coordinates. Lastly, we used the inferred geographical coordinates of the sampled sequences in each lineage to infer the most likely geographical coordinates of the MRCAs of every pair of sequences and used these together with the temporal scaling of the phylogenetic trees to track the spatiotemporal dispersal dynamics of the lineages during the earliest stages of their emergence (Fig. 3b).

The same subset of sequences that was used to produce the time-scaled MCC tree was used to perform continuous phylogeographic reconstructions of the respective BA lineages. The spatial resolution was determined by the available sampling location information for each sequence, ranging from province (all) to district (~34%) level. For sequences with only province-level information, the coordinates of the province’s most densely populated city were used. We first constructed maximum likelihood trees for each lineage separately and visualised these in TempEst to assess whether each dataset contained a sufficiently strong temporal signal to warrant the use of a molecular clock model, and to remove potential outliers²⁷. We then used the ModelFinder³³ component of IQ-TREE to identify GTR + I + G as a suitable nucleotide substitution model for all of the datasets. To model the geographical movements of the respective BA lineages across South Africa, we used a flexible relaxed random walk diffusion model implemented in BEAST that accommodates branch-specific variation in rates of dispersal with a Cauchy distribution that is well-suited to assumptions of SARS-CoV-2 transmission dynamics, and which enables estimation of diffusion statistics^{34,35}. In addition to the spatial diffusion and nucleotide substitution models we applied a strict clock model and an exponential growth coalescent model, as identified through MLE model comparison (Supplementary Table 2).

MCMC chains were run in duplicate for 300 million steps with sampling every 30,000 steps, and convergence being assessed using Tracer. MCC trees were produced from the sampled MCMC chain states using TreeAnnotator after discarding 10% of samples as burn-in.

We used the R package *seraphim*³⁶ to extract and map spatiotemporal information embedded in the sampled trees.

To investigate whether the predicted origins of the respective BA lineages could be traced to one or more neighbouring countries of South Africa, we additionally conducted discrete phylogeographic reconstructions of each lineage at a regional level (South Africa and its neighbouring countries). We utilised the same sampling strategy as described for the continuous phylogeographic reconstruction, subsampling 200 sequences per lineage based on available case count data (scaled by population). We first reconstructed phylogenies using 100 million MCMC iterations without incorporating location data. We then parameterised the Continuous-Time Markov Chain (CTMC) model using *PrioriTree*³⁷, specifying a dispersal process with an asymmetric prior, where relative dispersal rates between countries were estimated while allowing for directional differences in movement probabilities. The CTMC models were run in duplicate for 50 million MCMC iterations each, with the MCC trees produced from the sampled MCMC chain states using *TreeAnnotator* after discarding 10% of samples as burn-in. The spatiotemporal information embedded in the sampled trees was extracted and visualised using *Spread3*³⁸ software to determine the dispersal dynamics of the respective BA lineages across the region.

Reporting summary

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

Data availability

All of the SARS-CoV-2 sequences analysed and presented here are publicly accessible through the GISAID platform (<https://www.gisaid.org/>), using the GISAID identifier: EPI_SET_250304tq (<https://doi.org/10.55876/gis8.250304tq>). For a summary of the sequences included in the respective analyses please refer to Supplementary Table 1. A detailed breakdown of sequences included in the respective analyses can be found on our Github repository: <https://github.com/CERI-KRISP/SARS-CoV-2-Omicron-origins-South-Africa>. Other publicly available data used in this study are as follows: sequence designation list available on the PANGO Github repository (github.com/cov-lineages/pango-designation/milestones); case count data from the COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (<https://github.com/CSSEGISandData/COVID-19>); case data from COVID-19 South Africa (<https://www.covid19sa.org/provincial-breakdown>); custom python script (https://github.com/CERI-KRISP/SARS-CoV-2_VOC_dissemination) used to derive the number of state changes over the span of the tree; geospatial boundary data from the Humanitarian Data Exchange programme (<https://data.humdata.org/>) and Natural Earth (<https://www.naturalearthdata.com/>); and gridded population count datasets from WorldPop (<https://hub.worldpop.org/>).

Code availability

All manuscript materials (e.g. BEAST XML input files, R code) can be found on our Github repository: <https://github.com/CERI-KRISP/SARS-CoV-2-Omicron-origins-South-Africa>.

References

- Tegally, H. et al. Detection of a SARS-CoV-2 variant of concern in South Africa. *Nature* **592**, 438–443 (2021).
- Tegally, H. et al. Emergence of SARS-CoV-2 Omicron lineages BA.4 and BA.5 in South Africa. *Nat. Med.* **28**, 1785–1790 (2022).
- Viana, R. et al. Rapid epidemic expansion of the SARS-CoV-2 Omicron variant in southern Africa. *Nature* **603**, 679–686 (2022).
- Stats, S. A. *Mid-Year Population Estimates*. https://www.statssa.gov.za/?page_id=1854 (2022).
- Chen, J. et al. Emerging dominant SARS-CoV-2 variants. *J. Chem. Inf. Model.* **63**, 335–342 (2023).
- Tamura, T. et al. Virological characteristics of the SARS-CoV-2 XBB variant derived from recombination of two Omicron subvariants. *Nat. Commun.* **14**, 2800 (2023).
- Rasmussen, M. et al. First cases of SARS-CoV-2 BA.2.86 in Denmark, 2023. *Eurosurveillance* **28**, 2300460 (2023).
- Khan, K. et al. Evolution and neutralization escape of the SARS-CoV-2 BA.2.86 subvariant. *Nat. Commun.* **14**, 8078 (2023).
- Cele, S. et al. Omicron extensively but incompletely escapes Pfizer BNT162b2 neutralization. *Nature* **602**, 654–656 (2022).
- Du, P., Gao, G. F. & Wang, Q. The mysterious origins of the Omicron variant of SARS-CoV-2. *Innovation* **3**, 100206 (2022).
- Mallapaty, S. Where did Omicron come from? Three key theories. *Nature* **602**, 26–28 (2022).
- Sun, Y., Lin, W., Dong, W. & Xu, J. Origin and evolutionary analysis of the SARS-CoV-2 Omicron variant. *J. Biosaf. Biosecurity* **4**, 33–37 (2022).
- NICD. COVID-19 Weekly Testing Summary. *NICD* <https://www.nicd.ac.za/diseases-a-z-index/disease-index-covid-19/surveillance-reports/weekly-testing-summary/> (2023).
- Lasrado, N., Rössler, A., Rowe, M., Collier, A. Y. & Barouch, D. H. Neutralization of SARS-CoV-2 Omicron subvariant BA.2.87.1. *Vaccine* **42**, 2117–2121 (2024).
- Global HIV & AIDS statistics — Fact sheet. <https://www.unaids.org/en/resources/fact-sheet> (2023).
- Cele, S. et al. SARS-CoV-2 prolonged infection during advanced HIV disease evolves extensive immune escape. *Cell Host Microbe* **30**, 154–162.e5 (2022).
- Lambarey, H. et al. SARS-CoV-2 infection is associated with uncontrolled HIV Viral load in non-hospitalized HIV-infected patients from Gugulethu, South Africa. *Viruses* **14**, 1222 (2022).
- Sigal, A., Neher, R. A. & Lessells, R. J. The consequences of SARS-CoV-2 within-host persistence. *Nat. Rev. Microbiol.* 1–15. <https://doi.org/10.1038/s41579-024-01125-y> (2024).
- Khare, S. et al. GISAID's Role in Pandemic Response. *China CDC Wkly* **3**, 1049–1051 (2021).
- Hadfield, J. et al. Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics* **34**, 4121–4123 (2018).
- Aksamentov, I., Roemer, C., Hodcroft, E. B. & Neher, R. A. Nextclade: clade assignment, mutation calling and quality control for viral genomes. *J. Open Source Softw.* **6**, 3773 (2021).
- Chen, Z., Lemey, P. & Yu, H. Approaches and challenges to inferring the geographical source of infectious disease outbreaks using genomic data. *Lancet Microbe* **5**, e81–e92 (2024).
- Sagulenko, P., Puller, V. & Neher, R. A. TreeTime: maximum-likelihood phylodynamic analysis. *Virus Evol.* **4**, vex042 (2018).
- Minh, B. Q. et al. IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).
- Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q. & Vinh, L. S. UFBoot2: improving the ultrafast bootstrap approximation. *Mol. Biol. Evol.* **35**, 518–522 (2018).
- Rambaut, A., Lam, T. T., Max Carvalho, L. & Pybus, O. G. Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus Evol.* **2**, vey007 (2016).
- Hill, V. & Baele, G. Bayesian estimation of past population dynamics in BEAST 1.10 using the skygrid coalescent model. *Mol. Biol. Evol.* **36**, 2620–2628 (2019).
- Suchard, M. A. et al. Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evol.* **4**, vey016 (2018).
- Griffiths, R. C., Tavaré, S., Bodmer, W. F. & Donnelly, P. J. Sampling theory for neutral alleles in a varying environment. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **344**, 403–410 (1997).
- Rambaut, A., Drummond, A. J., Xie, D., Baele, G. & Suchard, M. A. Posterior summarization in Bayesian phylogenetics using tracer 1.7. *Syst. Biol.* **67**, 901–904 (2018).

31. Wickham, H. ggplot2. *WIREs Comput. Stat.* **3**, 180–185 (2011).
32. Yu, G. Using ggtree to visualize data on tree-like structures. *Curr. Protoc. Bioinform.* **69**, e96 (2020).
33. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A. & Jermini, L. S. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* **14**, 587–589 (2017).
34. 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).
35. Dellicour, S. et al. Relax, keep walking — a practical guide to continuous phylogeographic inference with BEAST. *Mol. Biol. Evol.* **38**, 3486–3493 (2021).
36. Dellicour, S., Rose, R., Faria, N. R., Lemey, P. & Pybus, O. G. SERAPHIM: studying environmental rasters and phylogenetically informed movements. *Bioinformatics* **32**, 3204–3206 (2016).
37. Gao, J., May, M. R., Rannala, B. & Moore, B. R. PrioriTree: a utility for improving phylodynamic analyses in BEAST. *Bioinformatics* **39**, btac849 (2023).
38. Bielejec, F. et al. Spread3: Interactive visualization of spatio-temporal history and trait evolutionary processes. *Mol. Biol. Evol.* **33**, 2167–2169 (2016).

Acknowledgements

We gratefully acknowledge all data contributors, i.e., the authors and their originating laboratories responsible for obtaining the specimens, and their submitting laboratories for generating the genetic sequence and metadata and sharing via the GISAID Initiative, on which this research is based. We also thank Dr. Marcel Dunaïski and the Computer Science Division at Stellenbosch University for providing computational resources that assisted in running some of the analyses for this research. Sequencing and modelling activities at KRISP and CERi are supported in part by grants from the Rockefeller Foundation (HTH 017; T.d.O.), the Abbott Pandemic Defense Coalition (APDC; T.d.O.), the National Institute of Health USA (U01 AI151698; T.d.O.) for the United World Antivirus Research Network (UWARN; T.d.O.), the INFORM Africa project through IHVN (U54 TW012041; T.d.O.) and the eLwazi Open Data Science Platform and Coordinating Center (U2CEB032224; T.d.O.), the SAMRC South African mRNA Vaccine Consortium (SAM-VAC; T.d.O.), European Union's Horizon Europe Research and Innovation Programme (101046041; T.d.O.), the Health Emergency Preparedness and Response Umbrella Program (HEPR Program; T.d.O.), managed by the World Bank Group (TF0B8412; T.d.O.), the GIZ commissioned by the Government of the Federal Republic of Germany, the UK's Medical Research Foundation (MRF-RG-ICCH-2022-100069; T.d.O., H.T.), the Wellcome Trust for the Global.health project (228186/Z/23/Z; T.d.O., H.T.), and the Novo Nordisk Foundation (NNF24OC0094346; H.T.). Sequencing activities by the EViTOH group were funded by the Gates Foundation (grant number INV-045475; M.V.). Sequencing activities by the Global Immunology and Immune Sequencing for Epidemic Response (GIISER) program are supported by the Gates Foundation (INV-030570; J.N.B.). Sequencing activities for NICD are supported by a conditional grant from the South African National Department of Health as part of the emergency COVID-19 response; a cooperative agreement between the National Institute for Communicable Diseases of the National Health Laboratory Service and the United States Centers for Disease Control and Prevention (FAIN# U01P001048; NU51P000930; N.W., A.v.G.); the South African Medical Research Council (SAMRC, project number 96838; N.W., A.v.G.); the African Society of Laboratory Medicine (ASLM) and Africa Centers for Disease Control and Prevention through a sub-award from the Bill and Melinda Gates Foundation (INV-018978; A.v.G.); the UK Foreign, Commonwealth and Development Office and Wellcome (221003/Z/20/Z; A.v.G.); This work was partly funded by the SEQAFRICA project which is funded by the UK Department of Health and Social Care's

Fleming Fund using UK aid. NICD sequencing was also supported by The Coronavirus Aid, Relief, and Economic Security Act (CARES ACT) through the Centers for Disease Control and Prevention (CDC) and the COVID International Task Force (ITF) funds through the CDC under the terms of a subcontract with the African Field Epidemiology Network (AFENET) (AF-NICD-001/2021; A.v.G.). Hyrax Biosciences' Exatype platform, used for the assembly of SARS-CoV-2 genomes, was supported by the South African Medical Research Council (SAMRC) with funds received from the Department of Science and Innovation. Additional funds for NGS-SA were also routed through the University of KwaZulu-Natal from the SAMRC with funds received from the South African Department of Science and Innovation. The content, views and findings reported/illustrated are the sole deduction, view and responsibility of the researcher/s and do not reflect the official position and sentiments of the funders.

Author contributions

Genomic and data generation: T.d.O., H.T., E.W., D.P.M., D.T., D.K., M.M., B.N., R.J., A.I., M.M.N., D.G., T.M., J.M., O.L.-A., Y.R., C.M., L.C., S.P., J.G., C.B., N.-Y.H., W.P., J.N.B., M.-A.D., M.V., F.K.T., N.W., C.W., A.v.G., R.L. Sample data and metadata curation: G.D. Data analysis: G.D., H.T. Study design and data interpretation: T.d.O., H.T., G.D., D.P.M., E.W., R.L. Manuscript writing: G.D., D.P.M., H.T., T.d.O. All of the authors reviewed the manuscript.

Competing interests

N.W. has received grant funding from the US CDC, Gates Foundation and Sanofi. The remaining authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41467-025-60081-0>.

Correspondence and requests for materials should be addressed to Houriiyah Tegally or Tulio de Oliveira.

Peer review information *Nature Communications* thanks Marco Salemi and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

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

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

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025

¹Centre for Epidemic Response and Innovation (CERI), School for Data Science and Computational Thinking, Stellenbosch University, Stellenbosch, South Africa. ²KwaZulu-Natal Research Innovation and Sequencing Platform (KRISP), Nelson R. Mandela School of Medicine, University of KwaZulu-Natal, Durban, South Africa. ³Division of Computational Biology, Department of Integrative Biomedical Sciences, Institute of Infectious Diseases and Molecular Medicine, University of Cape Town, Cape Town, South Africa. ⁴National Institute for Communicable Diseases (NICD) of the National Health Laboratory Service (NHLS), Johannesburg, South Africa. ⁵Division of Medical Virology, Department of Pathology, University of Cape Town, Cape Town, South Africa. ⁶Computational Biology Division, University of Cape Town, Cape Town, South Africa. ⁷Next Generation Sequencing Unit and Division of Virology, Faculty of Health Sciences, University of the Free State, Bloemfontein, South Africa. ⁸Division of Virology, University of the Free State, Bloemfontein, South Africa. ⁹PathCare, Pretoria, South Africa. ¹⁰National Health Laboratory Service, Tygerberg, Cape Town, South Africa. ¹¹Division of Medical Virology, Faculty of Medicine & Health Sciences, Stellenbosch University, Stellenbosch, South Africa. ¹²PathCare Reference Laboratory, Cape Town, South Africa. ¹³National Health Laboratory Service, Port Elizabeth, South Africa. ¹⁴Department of Laboratory Medicine and Pathology, Faculty of Health Sciences, Walter Sisulu University, Eastern Cape, South Africa. ¹⁵Emerging Viral Threats, One Health surveillance and vaccines (EViTOH) Division, Infectious Disease and Oncology Research Institute (IDORI), School of Pathology, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa. ¹⁶National Health Laboratory Service, Cape Town, South Africa. ¹⁷South African Medical Research Council Antibody Immunity Research Unit, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa. ¹⁸Centre for Infectious Disease Epidemiology and Research, School of Public Health, University of Cape Town, Cape Town, South Africa. ¹⁹Health Intelligence, Western Cape Government Health and Wellness, Cape Town, South Africa. ²⁰Centre for Emerging Arbo and Respiratory Virus Research (CEARV), Department of Medical Virology, University of Pretoria, Pretoria, South Africa. ²¹School of Pathology, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa. ²²Institute of Infectious Disease and Molecular Medicine and Wellcome Centre for Disease Research in Africa, Faculty of Health Sciences, University of Cape Town, Cape Town, South Africa. ²³Division of Medical Microbiology, Faculty of Health Sciences, University of Cape Town, Cape Town, South Africa. ²⁴These authors jointly supervised this work: Houriiyah Tegally, Tulio de Oliveira.

✉ e-mail: houriiyah@sun.ac.za; tulio@sun.ac.za