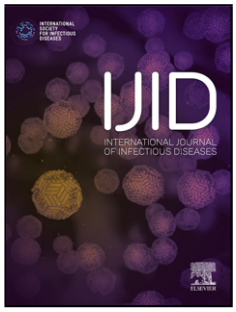
Journal Pre-proof



Early transmission of SARS-CoV-2 in South Africa: An epidemiological and phylogenetic report

Jennifer Giandhari, Sureshnee Pillay, Eduan Wilkinson, Houriiyah Tegally, Ilya Sinayskiy, Maria Schuld, Jose Lourenco, Benjamin Chimukangara, Richard Lessells, Yunus Moosa, Inbal Gazy, Maryam Fish, Lavanya Singh, Khulekani Sedwell Khanyile, Vagner Fonseca, Marta Giovanetti, Luiz Carlos Junior Alcantara, Francesco Petruccione, Tulio de Oliveira

PII: S1201-9712(20)32322-5

DOI: https://doi.org/10.1016/j.ijid.2020.11.128

Reference: IJID 4838

To appear in: *International Journal of Infectious Diseases*

Received Date: 18 June 2020

Revised Date: 2 November 2020

Accepted Date: 5 November 2020

Please cite this article as: { doi: https://doi.org/

Early transmission of SARS-CoV-2 in South Africa: An epidemiological and phylogenetic report

Authors: Jennifer Giandhari, PhD[1]*, Sureshnee Pillay, MSc[1]*, Eduan Wilkinson, PhD[1]*, Houriiyah Tegally, MSc[1]*, Ilya Sinayskiy, PhD[2,3], Maria Schuld, PhD [2], Jose Lourenco, PhD [4], Benjamin Chimukangara, PhD[1], Richard Lessells, PhD[1,4], Yunus Moosa, PhD [4], Inbal Gazy, PhD[1], Maryam Fish, PhD[1] Lavanya Singh, PhD[1] Khulekani Sedwell Khanyile, MSc[1], Vagner Fonseca, MSc[1,5,6] Marta Giovanetti, PhD[6], Luiz Carlos Junior Alcantara, PhD[5,6], Francesco Petruccione, PhD[2,3], Tulio de Oliveira, PhD[1,7,8].

Affiliations

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

[2]Quantum Research Group, School of Chemistry and Physics, University of KwaZulu-Natal, Durban, South Africa.

[3]National Institute for Theoretical Physics (NITheP), KwaZulu-Natal, 4001, South Africa [3]

[4]Department of Zoology, University of Oxford, Oxford OX1 3PS, UK.

[5]Laboratorio de Genetica Celular e Molecular, ICB, Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, Brazil

[6]Laboratório de Flavivírus, Instituto Oswaldo Cruz Fiocruz, Rio de Janeiro, Brazil

[7]Centre for Aids Programme of Research in South Africa (CAPRISA), Durban South Africa.

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

Corresponding Author:

Prof. Tulio de Oliveira

Emails: deoliveira@ukzn.ac.za and tuliodna@uw.edu

Address: 719 Umbilo Road,

Nelson R Mandela School of Medicine, UKZN

Durban, South Africa.

Highlights

Highlights for this research:

- Early epidemic in South Africa highly heterogeneous with marked differences in epidemiological progression between major provinces in the country
- Early transmission in KZN associated with multiple international introductions of SARS-CoV-2
- Dominating lineages during this time of the epidemic was B.1 and B, lineages originating mainly from Europe
- Evidence for transmission cluster during first month of the epidemic suggesting locally acquired infection
- Clustered outbreak linked to a major hospital outbreak in Durban, which inflated early mortality in KZN.

Evidence before this study

We searched PubMed, BioRxiv and MedRxiv for reports on epidemiology and phylogenetic analysis using whole genome sequencing (WGS) of SARS-CoV-2. We used the following keywords: SARS-CoV-2, COVID-19, 2019-nCoV or novel coronavirus and transmission genomics, epidemiology, phylogenetic or reproduction number. Our search identified an important lack of molecular epidemiology studies in the southern hemisphere, with only a few reports from Latin America and one in Africa. In other early transmission reports on SARS-CoV-2 infections in Africa, authors focused on transmission dynamics, but molecular and phylogenetic methods were missing.

Added value of this study

With a growing sampling bias in the study of transmission genomics of the SARS-CoV-2 pandemic, it is important for us to report high-quality whole genome sequencing (WGS) of local SARS-CoV-2 samples and in-depth phylogenetic analyses of the first month of infection in South-Africa. In our molecular epidemiological investigation, we identify the early transmission routes of the infection in the KZN and report thirteen distinct introductions from many locations and a cluster of localized transmission linked to a healthcare setting that caused most of the initial deaths in South Africa. Furthermore, we formed a national consortium in South Africa, funded by the Department of Science and Innovation and the South African Medical Research Council, to capacitate ten local laboratories to produce and analyse SARS-CoV-2 data in near real time.

Implications of all the available evidence

The COVID-19 pandemic is progressing around the world and in Africa. Early transmission genomics and dynamics of SARS-CoV-2 throw light on the early stages of the epidemic in a given region. This facilitates the investigation of localized outbreaks and serves to inform public health responses in South Africa.

Abstract

Objectives

To investigate introduction and understand the early transmission dynamics of the SARS-CoV-2 in South-Africa, we formed the Network for Genomic Surveillance in South Africa (NGS-SA)

Design

Here, we present the first results of this effort, which is a molecular epidemiological study of the first twenty-one SARS-CoV-2 whole genomes sampled in the first port of entry, KwaZulu-Natal (KZN), during the first month of the epidemic. By combining this with calculations of the effective reproduction number (R), we aim to shed light on the patterns of infections in South Africa.

Results

Two of the largest provinces, Gauteng and KwaZulu-Natal, had a slow growth rate on the number of detected cases, while in Western Cape and Eastern Cape the epidemic is spreading fast. Our estimates of transmission potential suggest a decrease towards R=1 since the first cases and deaths but a subsequent estimated R average of 1.39 between 6-18th of May 2020. We also demonstrate that early transmission in KZN was associated with multiple international introductions and dominated by lineages B1 and B and provide evidence for locally acquired infections in a hospital in Durban within the first month of the epidemic.

Conclusion

The COVID-19 pandemic in South Africa was very heterogeneous in its spatial dimension, with many distinct introductions of SARS-CoV2 in KZN and evidence of nosocomial transmission, which inflated early mortality in KZN. The pandemic at the local level is still

developing and the objective of NGS-SA is to clarify the dynamics of the epidemic in South Africa and devise the most effective measures as the outbreak evolves.

Keywords: South Africa; COVID-19; first introductions; national consortium; genomics; reproductive number; Phylogenetics; Molecular Epidemiology

Introduction

The novel coronavirus disease 2019 (COVID-19) was detected in China in late December 2019. On 30 January 2020, it was declared a Public Health Emergency of International Concern by the World Health Organization (WHO) (Sohrabi et al., 2020). By 15th of May 2020, there were 4,621,410 COVID-19 cases and 308,542 related deaths (Worldometer, 2020) worldwide involving almost every country in the world. Within five months, the virus had spread to Europe, America and eventually to Africa. The first case in Africa was reported in Nigeria on 28th of February 2020 (Adepoju, 2020), and at the time of writing, the pandemic has spread to almost all countries on the African continent. South Africa has had the highest number of COVID-19 cases to date with a total of 13,524 people infected and 247 deaths (as at 15th May)(COVID-19 WEEKLY EPIDEMIOLOGY BRIEF PROVINCES AT A GLANCE, n.d.).

The first confirmed case of COVID-19 in South Africa was reported on 5th of March 2020. Decisive early action was taken by the government: a national state of disaster was declared on 15th of March 2020, and a nationwide lockdown was enforced on 27th of March 2020 to avoid the first wave overwhelming the health system. While initially only people who had travelled to at-risk countries and their contacts received PCR tests for severe acute respiratory syndrome-related coronavirus 2 (SARS-CoV-2), the recommendation broadened to include all people with an acute respiratory illness. Furthermore, a program of community-based screening and testing was rolled out across the country (NICD, 2020). Testing increased rapidly and by the middle of May 2020, over 600,000 tests had been carried out in South Africa (approximately 10,000 per million population) (Roser M et al., 2020).

As the global pandemic has expanded, WGS and genomic epidemiology (Grubaugh et al., 2019) have been consistently used to investigate COVID-19 transmission and outbreaks (Deng et al., 2020; Eden et al., 2020; Gonzalez-Reiche et al., 2020; Grubaugh, 2020; Leung et al., 2020; Lu et al., 2020; Munnink et al., 2020). In response to the COVID-19 pandemic, the South African Network for Genomics Surveillance of COVID (NGS-SA) was formed (Msomi et al., 2020), which is a network of five large government laboratories and five public universities funded by the Department of Science and Innovation and the South African Medical Research Council. In this paper, our consortium focuses on a detailed analysis of the epidemic in South Africa and preliminary genomic analysis of some of the first introductions of SARS-CoV-2 in KwaZulu-Natal (KZN). We show that although the South African epidemic started in KZN, which have the first cases and deaths, other provinces in the country, namely the Western Cape (WC), Gauteng (GP) and the Eastern Cape (EC), have overtaken KZN in the number of confirmed cases. We also show evidence of many distinct introductions of SARS-CoV-2 in KZN and early evidence suggesting nosocomial transmission.

Methods

Data sources

We used publicly released data up to 11 May 2020 from the National Department of Health (NDoH) and the NICD in South Africa, which are collected in the repository of the Data Science for Social Impact Research Group at the University of Pretoria (Marivate et al., 2020), as well as global data on confirmed cases from the Johns Hopkins Coronavirus Resource Centre (Dong et al., 2020). The NDoH releases daily updates on the number of new confirmed cases, with a breakdown by province. In the early stages of the epidemic, individual-level information on sex, age and travel history was released, but detailed reporting was discontinued on 23rd of

March. In addition, the National Institute of Communicable Diseases (NICD) releases daily updates on the number of reverse-transcriptase polymerase chain reaction (RT-PCR) tests performed across all public and private sector laboratories, as well as the number of cases testing positive for severe acute respiratory syndrome-related coronavirus 2 (c). We also extracted information from government press releases and speech transcripts to chart a timeline of the government response to the epidemic. To understand the epidemic trajectory, we plotted the cumulative number of confirmed cases by province since the report of the hundredth case in the country by province.

Epidemiological analysis and reproductive number estimation

The effective reproduction number (R) was estimated by taking into account the observed epidemic growth rate r and two theoretical relationships (i, ii) of R with r previously described in the literature. (i) We used the relationship $R=(1+r/b)^a$ as described in Imperial College London's COVID-19 report 13 (Flaxman et al., 2020), where $a=m^2/s^2$ and $b=m/s^2$, m the serial interval (SID) mean and the SID standard deviation. The SID distribution used is the one estimated by Nishiura and colleagues (Nishiura et al., 2020), with m=4.7 and s=2.9. We term this approach the Flaxman et al. approach (Flaxman et al., 2020). (ii) We used the relationship R=(1+r/sigma)(1+r/delta), with 1/sigma the infectious period and 1/delta the incubation period, as described by Wallinga and Lipsitch (Wallinga and Lipsitch, 2007), which is based on an SEIR modelling framework and expects both periods to be exponentially distributed. We used exponential distributions with mean 5.1 days for incubation (Kucharski et al., 2020; Linton et al., 2020) and 4 days for the infection (Kucharski et al., 2020; Linton et al., 2020). We term this approach the Wallinga et al. approach. To obtain the epidemic growth rate r, we used maximum likelihood estimation in R (function optim), by fitting the exponential growth

model A_0e^{rt} to the reported time series of cases and deaths (independently), where t is time, A_0 is the number of reports at t=0, and r the growth rate. We used daily reported deaths and cases. The time periods for which we had data for deaths was 27th March to 11th May, and for cases was 5th March to 12th May. This approach is similar to that implemented by Xavier et al (Xavier et al., 2020).

Ethics statement

The project was approved by University of KwaZulu-Natal Biomedical Research Ethics Committee. Protocol reference number: BREC/00001195/2020. Project title: COVID-19 transmission and natural history in KwaZulu-Natal, South Africa: Epidemiological Investigation to Guide Prevention and Clinical Care.

SARS-CoV-2 sample collection and preparation

Remnant samples from nasopharyngeal and oropharyngeal swabs collected from symptomatic patients, were used for SARS-CoV-2 WGS. These samples comprised of either the primary swab sample or extracted RNA. The swab samples were heat inactivated in a water bath at 60°C for 30 minutes, in biosafety level 3 laboratory, prior to RNA extraction. RNA was extracted using the Viral NA/gDNA Kit on the Chemagic 360 system (Perkin Elmer, Hamburg, Germany) using the automated Chemagic 360 instrument (Perkin Elmer, Hamburg, Germany) or manually using the Qiagen Viral RNA Mini Kit (QIAGEN, California, USA).

Real Time RT-PCR

In order to detect the SARS-CoV-2 virus by PCR, the TaqPath COVID-19 CE-IVD RT-PCR Kit (Life Technologies, Carlsbad, CA) was used according to the manufacturer's instructions. The assays target genomic regions (ORF1ab, S protein and N protein) of the SARS-CoV-2 genome. RT-PCR was performed on a QuantStudio 7 Flex Real-Time PCR instrument (Life Technologies, Carlsbad, CA). Cycle thresholds (Ct) was analysed using auto-analysis settings with the threshold lines falling within the exponential phase of the fluorescence curves and above any background signal. To accept the results, we confirmed a Ct value for RNAse P (i.e. an endogenous internal amplification control) and or the target gene in each reaction, with undetermined Ct values in the no template control. The Ct values were reported for each target gene.

Tiling Polymerase Chain Reaction

cDNA synthesis was performed on the RNA using random primers followed by gene specific multiplex PCR using the ARTIC protocol (Quick, 2020). Briefly, extracted RNA was converted to cDNA using the Protoscript II First Strand cDNA synthesis Kit (New England Biolabs, Hitchin, UK) and random hexamer primers. SARS-CoV-2 whole genome amplification by multiplex PCR was carried out using primers designed on Primal Scheme (http://primal.zibraproject.org/) to generate 400bp amplicons with an overlap of 70bp that covers the 30Kb SARS-CoV-2 genome. PCR products were cleaned up using AmpureXP purification beads (Beckman Coulter, High Wycombe, UK) and quantified using the Qubit dsDNA High Sensitivity assay on the Qubit 4.0 instrument (Life Technologies Carlsbad, CA).

Illumina MiSeq Sequencing

PCR products for samples yielding sufficient material were included in this sequencing platform. Illumina® TruSeq® Nano DNA Library Prep kits were used according to the manufacturer's protocol to prepare uniquely indexed paired end libraries of genomic DNA. The libraries were quantified using the Qubit dsDNA High Sensitivity assay on the Qubit 4.0 instrument (Life Technologies) and the fragments were analysed using the LabChip GX Touch (Perkin Elmer, Hamburg, Germany). Sequencing libraries were normalized to 4nM, pooled and denatured with 0·2N sodium acetate. 12pM sample library was spiked with 1% PhiX (PhiX Control v3 adapter-ligated library used as a control). Libraries consisting of 12 samples each were loaded onto a 500-cycle MiSeq Nano Reagent Kit v2 nano v2 Miseq reagent kit and run on the Illumina MiSeq instrument (Illumina, San Diego, CA, USA).

Bioinformatics assembly of genomes

Raw reads coming from both Nanopore and Illumina sequencing were assembled using Genome Detective 1·126 (https://www.genomedetective.com/) and the Coronavirus Typing Tool (Cleemput et al., 2020; Vilsker et al., 2019). The initial assembly obtained from Genome Detective was polished by aligning mapped reads to the references and filtering out low-quality mutations using bcftools 1·7-2 mpileup method. All mutations were confirmed visually with bam files using Geneious software (Biomatters Ltd, New Zealand). All of the sequences were deposited in GISAID (https://www.gisaid.org/) (Shu and McCauley, 2017).

Reference dataset

We downloaded all sequences and associated metadata from the GISAID sequence database (https://www.gisaid.org/) (Shu and McCauley, 2017) as of 1st of May 2020 (n=15,793). Due to the low variability of SARS-CoV-2, we wished to only include high quality sequences in our downstream analyses. To this end, we filtered out sequences that were <25kbp in length as well as sequences with a high proportion of ambiguous sites (>5%). Additionally, we also removed sequences that lacked any geographic and or sampling date information. The resulting 10,959 sequences were analyzed along with 20 sequences that were generated by the laboratory at the KwaZulu-Natal Research Innovation and Sequencing Platform (KRISP). The dataset also contained one additional KZN sequences (*EPI_ISL_417186*) that were generated by the National Institute for Communicable Diseases (NICD) and represent a distant contact of the first diagnosed case in South Africa.

Lineage classification

Currently, no established nomenclature system exists for SARS-CoV-2. A dynamic lineage classification method proposed by Rambault et al., was used in this study (Rambaut et al., 2020) via the Phylogenetic Assignment of named Global Outbreak LINeages (PANGOLIN) software suite (https://github.com/hCoV-2019/pangolin). This is aimed at identifying the most epidemiologically important lineages of SARS-CoV-2 at the time of analysis, allowing researchers to monitor the epidemic in a particular geographical region more effectively. Two main SARS-CoV-2 lineages are currently recognized; lineage A, defined by Wuhan/WH04/2020 and lineage B, defined by Wuhan-Hu-1 strain. Although Wuhan-Hu-1 was the first published genome from SARS-CoV-2, it is classified as lineage B. Phylogenetic analyses of SARS-CoV-2 identified sequences from lineage A to be more closely related to a

bat corona virus (Rambaut et al., 2020), which suggest this to be the first lineage (hence A). Lineage A genomes are characterized by two unique mutations (8782C>T and 28144T>C), relative to lineage B. Lineage B, on the other hand, shares no common mutations since this lineage contains the global SARS-CoV-2 genome reference (Wuhan-Hu-1). From these lineages, sub-lineages (e.g. A·1, A·2, A·3 and so forth) are then designated, each defined by an additional set of unique mutations. For example, for sub-lineage A·1, these mutations would be; 11747C>T, 1785A>G and 18060C>T. Sub- lineages can further diversify into sub sub-lineages (e.g. A·1·1). Please refer to the schema provided in Supplementary Figure 5 for more information.

Phylogenetic analysis

10,959 GISAID reference genomes and 20 KRISP sequences were aligned in Mafft v7·313 (FF-NS-2) followed by manual inspection and editing in the Geneious Prime software suite (Biomatters Ltd, New Zealand). We constructed a maximum likelihood (ML) tree topology in IQ-TREE (GTR+G+I, no support) (Nguyen et al., 2015; Tavaré and Miura, 1986). Due to the large size of the alignment and the low variability, we opted to not infer support for splits in this tree topology. In any tree topology of SARS-CoV-2 the majority splits will be poorly supported with only the major splits separating the major lineages having good support. The resulting ML tree topology was transformed into a time scaled phylogeny using TreeTime (Sagulenko et al., 2018) with a clock rate of $8x10^{-4}$ and rooted along the branch of Wuhan-WH04 (GISAID: hCoV19/Wuhan/WH04/2020) and Wuhan-Hu1 (Genbank: MN908947). The resulting phylogeny was viewed and annotated in FigTree and ggtree.

Based on this large phylogeny of SARS-CoV-2, we randomly down sampled the GISAID reference sequences that passed initial sequence quality checks to ~10% of the original size. All African sequences in the GISAID subset, the 20 genotypes that were generated in this study, as well as a select few external references (e.g. Wuhan-Hu-1) were included. The resulting dataset of 1848 sequences was used in a custom build on the NextStrain analysis platform (James Hadfield, Colin Megill, Sidney M Bell, John Huddleston, Barney Potter, Charlton Callender, Pavel Sagulenko, Trevor Bedford, 2018). To infer support for the splits in this tree topology we inferred an additional 100 bootstrap trees in IQ-Tree under the same model parameters as NextStrain. These trees were then used to infer transfer support for splits in the phylogeny (Lemoine et al., 2018).

Bayesian Tree

Bayesian coalescent analyses were performed on major lineages of the NextStrain build in which KZN sequences fell. The purpose of these analyses were to: (*i*) confirm the estimated date of origin for SARS-CoV-2 as proposed in recent literature (Andersen et al., 2020; Li et al., 2020), (*ii*) infer the estimated date to the most recent common ancestor (MRCA) for major lineages and (*iii*) infer the estimated dates of viral introductions into South Africa.

Due to the dynamic lineage assignment system of pangolin, many sub sub-lineages (e.g. A·1·1 or A·1·1·1) have emerged since the start of the outbreak. In order to keep things neat and tidy we organized B lineages into B, B·1 and B·2. Due to the large number of B and B·1 lineages, we randomly down sampled these while retaining all South African genotypes. This resulted in three datasets for Bayesian coalescent inference: (B = 128, B·1 = 178 & B·2 =69). Since none of the KZN sequences were classified as lineage A we exclude A genotypes from our Bayesian analyses.

In short, sequences were aligned in mafft v7.313 and visualized and manually edited in Geneious software (Biomatters Ltd, New Zealand) as previously described. ML-tree topologies were inferred from each alignment in IQ-TREE v1·6·9 (GTR+G+I, with transfer support values) (Nguyen et al., 2015; Tavaré and Miura, 1986). Resulting tree topologies were analyzed in TempEst software suite for temporal clock signal (Supplementary Figure S4). Coalescent molecular clock analyses were performed in BEAST v1·8. In short, analyses were run under a strict molecular clock assumption at a constant evolutionary rate of $8x10^{-4}$ nucleotide substitutions per site per year and an exponential growth coalescent tree prior. The Markov Chains were run in duplicate for a total length of 100 million steps sampling every 10,000 iterations in the chains. Runs were assets in Tracer for good convergence (ESS>200) and TreeAnnotator after discarding 10% of runs as burn-in.

Data Availability

SARS-CoV-2 genome sequences generated in this study have been deposited in the GISAID database (https://www.gisaid.org/), under the following accession IDs: *EPI_ISL_421572, EPI_ISL_421573, EPI_ISL_421574, EPI_ISL_421575 EPI_ISL_421576 EPI_ISL_436684 EPI_ISL_436685 EPI_ISL_436686 EPI_ISL_436687*. In addition, raw short and long reads have been submitted to the Short Read Archive (SRA) and can be accessed under BioProject Accession: PRJNA636748 (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA636748).

Results

Epidemiology of COVID-19 in KZN and South Africa

The first confirmed case of COVID-19 in South Africa was reported on 5th of March 2020 in KZN, a South African citizen returning home from a skiing holiday in Italy. A steady increase in the number of confirmed cases in South Africa (all imported cases) followed over the next week, with the first suspected case of local transmission reported on 13th of March 2020 in Durban, KwaZulu-Natal. The early cases were predominantly located in the three provinces with the main urban populations and international travel hubs, namely GP (main cities Pretoria and Johannesburg), the WC (Cape Town) and KZN (Durban). In these three provinces, the doubling time for confirmed cases was approximately three days prior to the lockdown (Figure 1). However, since the lockdown on 27th of March 2020, the epidemic seems to be growing at different rates in South Africa.

The South African epidemic has been very heterogeneous. For example, the first cases and deaths happened in KZN and GP. This was more pronounced in KZN, as a large nosocomial outbreak in a private hospital in Durban caused KZN to lead the country in number of deaths until the WC overtook it on 21 April 2020. In addition, GP, home of the largest metropolitan area of Johannesburg, had an unusual epidemic, as the majority of initial cases were in middle age and rich individuals who traveled overseas for holidays. This translated in a very small number of deaths over time and infections were concentrated in the rich suburb of Sandton in Johannesburg. However, the epidemic expanded the fastest in the Western Cape (WC) province, especially in Cape Town, which is the capital and most populated city of the WC. At the time of writing this report, this province has over 60% of all of the cases and deaths in South Africa (Figure 2). There is mounting evidence that the Western Cape is seeding the growing epidemic in the Eastern Cape as the funerals from some of the deaths in the Western Cape are taking place in the Eastern Cape.

This dynamic and heterogeneous epidemic complicates the estimation of effective reproductive number (R) over time and space. For example, deaths, which is normally one of the gold standard data for estimation of R0 for South Africa in May 2020 were stable at 1·12 (1.0-1.2) (Supplementary figure 1). KZN, the first province affected by COVID-19, initially had the highest death rate but in the last period analyzed, had only 3 deaths. We have therefore attempted to estimate R from two data sources: aggregated reported cases and deaths at the country level (See Methods). Similarly, to that observed in other regions of the world, our estimates of R for South Africa suggest a decreasing transmission potential towards R=1 since the first cases and deaths have been reported, independently of the data source used. By the last period analyzed between 6-18th of May, using the Wallinga et al approach (Wallinga and Lipsitch, 2007), we find that R was still 1.39 (1.04 - 2.15, 95% CI), suggesting potential of sustained transmission for the near future.

SARS-CoV2 genomes from KZN

In order to determine the route of introduction of the SARS-CoV-2 in KZN, we assessed 27 of some of the first confirmed cases in the province. Samples obtained from nasopharyngeal swabs represented fourteen females and ten males between the ages of 23-74 years. We managed to produce 20 near-whole genome sequences (>90% coverage) from these samples, and six partial genomes (Supplementary Table S2, Table S3). To this dataset, we added an extra genome from the NICD, which was sampled in KZN (a close contact of the first reported case) on 7th of March 2020. The 21 KZN whole genomes (20 KRISP and one NICD) were

assigned to SARS-CoV-2 sub-lineages according to the nomenclature proposed and lineage classification obtained from >5000 genomes analyzed by Rambaut et al. (Rambaut et al., 2020). Given uncertainties pertaining to the low diversity of this virus (Moreno et al., 2020), we restricted lineage assignment to the four most prominent subgroups (A, B, B·1 and B·2). Of the 21 KZN isolates being investigated in the present study, one was assigned to lineage B (*KRISP-006*) and one to sub-lineage B·2 (*KRISP-002*) (Figure 3). The remaining 19 KZN sequences were all assigned to lineage B·1. The B·1 lineage consists primarily of cases originating in Europe (Figure 3c), suggesting that introductions from Europe accounted for many of the early cases in KZN.

Although our investigation contained only a small number of samples from the first month of the epidemic in South Africa, we identified at least 13 distinct introductions (Figure 3 and Figure 4) and one monophyletic cluster involving seven sequences. Three of the sequences (*KRISP-007*, *KRISP-010* and *KRISP-011*) were identical and contained five mutations (241C>T, 3037C>T, 14408C>T, 16376C>T and 23403A>G). After investigation, we found that these samples were from health care workers at a private hospital in Durban, KZN, with no history of travel outside the country. A detailed investigation is currently being conducted in this hospital, but preliminary findings suggest a point-source nosocomial outbreak (Lessells et al. manuscript in preparation). The other four sequences in the cluster contained two pairs (*KRISP-103; KRISP-104* and *KRISP-105; KRISP-106*). Samples 103 and 104 are identical to one another and are characterized by three additional mutations (5672C>A, 10592A>G, and 26063G>T) on top of the ones reported above (Supplementary Table S4). Samples 106 has acquired one additional mutation (24034C>T) on top of the five mutations common to the

hospital outbreak, while sample 105 acquired another two mutation (13766A>T and 18411T>C) on top of the mutations found in 106. These two pairs were derived from random sampling within the Durban metropolitan area suggesting early evidence of localized transmission in Durban (Figure 3b).

Time-resolved analysis of three main lineages circulating in KZN

To determine the evolutionary relationship of the KZN sequences to the world-wide SARS-CoV-2 pandemic, we conducted a Bayesian molecular clock analysis for each of the lineages found in KZN (Figure 4). Coalescent molecular clock analyses of lineage B places KRISp-006 at the base of a subclade along with a sequences from Canada with high posterior support (*P*=1.0). The remainder of the subclade contains sequences from a large number of Asia countries (Singapore, Phillippines & Malaysia), Australia, the United States and the United Kingdom. The B·1 Bayesian analysis, which contained 19 KZN sequences, suggest multiple introductions into KZN from European countries. Due to low diversity in this lineage the posterior support for splits in the tree were very low. Furthermore, due to the small number of nucleotide differences between isolates the monophyletic clade of seven KZN sequences observed previously now only contained five sequences. Samples 103 and 104, though clustering together with one another, were seperated from the rest of the clade by other European reference sequences. The time to the most recent common ancestor (tMRCA) for the monophyletic KZN clade of five sequences were inferred around 23rd of March 2020, with the 95% Highest Posterior Density between 10-31st of March 2020, which is consistent with the dates of the nosocomial outbreak in Durban (Lessells et al. manuscript in preparation).

Our coalescent analyses in BEAST placed the origin of B & B·1 around the first week of December 2019 [95% HPD], with the 95% highest posterior density (HPD) ranging between

mid October and the last week of December 2019. Coalescent analyses placed the time to the most recent common ancestor (MRCA) for sub-lineage B·2 around late December 2019 (95% HPD mid November 2019 – first week of January 2020). Our temporal estimates are consistent with the SARS-CoV-2 date of origin.

Discussion & Conclusion

The spread of SARS-CoV-2 across the globe has given rise to one of the largest evolving pandemics in modern times. South Africa currently has the highest number of infections in Africa. South Africa seems to be moving to the next stage of the COVID-19 pandemic, with increasing community transmission even during the stringent lockdown and the epidemic growing at different rates in different regions of the country. At the time of writing this report, Cape Town, the main city in the WC, has the fastest increase of new infections and deaths in South Africa. Recent data indicates that over 62% of the new infections and deaths are happening in this province, although only 17% of the South African population lives in this region. The fast spread of COVID-19 in the WC is not fully explained by the higher testing rates as this province has performed between 20-22% of the tests in South Africa, but the positivity rate has been around 9%, were as in the other provinces the positivity rate is around 1-2%. Our estimates of transmission potential for South Africa suggest a decreasing transmission potential towards R=1 since the first cases and deaths have been reported, similarly to that observed in other regions of the world. By the last period analyzed between 6-18th of May, when using the Wallinga et al estimation approach applied to time series of reported cases, we estimate that R was on average 1.39 (1.04 - 2.15, 95% CI). Overall, these results suggest of an epidemic still in expansion at that time, in spite of a very early lockdown.

Sequencing of viral isolates from early COVID-19 cases in KZN, which is the province of South Africa with the first infections and early deaths, provided useful insights into the origins and transmission of SARS-CoV-2. From the first twenty-one genomes analyzed, we found thirteen independent introductions in KZN. These introductions were related to lineages B, B·1 and B·2, which have spread widely in Europe and North America. We also found a cluster of cases in health care workers in Durban, highlighting the potential importance of nosocomial transmission in this pandemic and potentially two other transmission pairs. The production of genomes from the WC will be crucial to understand the drivers of transmission during the lockdown period, and particularly whether health care facilities, prisons, workplaces and other institutions are acting as amplifiers of transmission. This is one of the main activities that our consortium, NGS-SA, is currently working on.

Genomic analysis of SARS-CoV-2 in Africa has proved challenging on many fronts. First, sequencing of high-quality SARS-CoV-2 genomes is not a straightforward task. For example, a survey of thousands of sequences deposited in public databases has revealed a number of putative sequencing issues that appear to be the result of contamination, recurrent sequencing errors or hypermutability (Virological, 2020). These might arise from laboratory-specific techniques of sample preparation, sequencing technology or consensus calling. Furthermore, the low diversity of this virus and the small number of mutations that define lineages have prompted caution in the interpretation of early phylogenetic analysis worldwide (Lu et al., 2020). Often apparent local transmission clusters can in fact be the result of multiple introductions from under-sampled regions from non-uniform sequencing efforts (Grubaugh et al., 2019; Kraemer et al., 2019). To mitigate this we confirmed phylogenetic results by manual inspection of mutations relative to the reference of SARS-CoV-2 (Supplementary Table S4).

Second, the pandemic is still evolving and grouping of SARS-CoV-2 into lineages and sub-clades is likely to be dynamic at this stage and it is influenced by proportionally larger number of sequences produced in the northern hemisphere (Rambaut et al., 2020). Third, the travel histories of apparent community transmission need to be thoroughly investigated in order to elucidate the true dynamics of transmission in a particular area. In our case, a subsequent investigation into the samples comprising the monophyletic cluster revealed the association with a big hospital outbreak of SARS-CoV-2 infections in Durban, KZN (Lessells et al. manuscript in preparation 2020).

This paper has some important limitations. The first is related to estimation of R from a limited number of deaths in a high heterogeneous epidemic both in time and space - for which we were able to estimate R only at the aggregated country level. The second is a lack of well set up genomics laboratories that can sequence the virus in Africa. This is also amplified by the difficulty of acquiring reagents that are in high demand, coupled with the disruption of air freight. It is therefore a high priority for our consortium, NGS-SA, to evaluate and share protocols among national laboratories in South Africa that could generate sequences of high-quality and capacitate our laboratories with the protocols and bioinformatics pipelines to properly investigate virus introduction and to validate the call of variants with a detailed and reliable bioinformatics system. NGS-SA is also working with the Africa Center for Disease Control (CDC) and the World Health Organization (WHO) to strengthen genomics surveillance in the African continent.

In this paper, we provide an early analysis of COVID-19 pandemic in South Africa, showing very heterogeneuos epidemics in the different provinces. We also estimated SARS-CoV-2 genetic diversity in KZN using the first twenty one genomes from some of the first cases in the

country. We find that KZN had many distinct introductions of SARS-CoV2, but also had early evidence of nosocomial transmission. The pandemic at the local level is still developing and the objective of NGS-SA is to clarify the dynamics of the epidemic in South Africa and devise the most effective measures as the outbreak evolves.

Funding Statement

This work is based upon research supported by the UKZN Flagship Program entitled: Afrocentric Precision Approach to Control Health Epidemic, by a research Flagship grant from the South African Medical Research Council (MRC-RFA-UFSP-01- 2013/UKZN HIVEPI, by the the Technology Innovation Agency and the the Department of Science and Innovation and by National Human Genome Re- search Institute of the National Institutes of Health under Award Number U24HG006941. H3ABioNet is an initiative of the Human Health and Heredity in Africa Consortium (H3Africa). The content is solely the responsibility of theauthors and does not necessarily represent the official views of any of the funders.

Conflict of interest

The authors have no conflict of interest to declare.

Acknowledgments

We wish to extend our thanks to all laboratory personnel that have worked hard to genotype SARS-CoV-2 samples and who have generously made it public via the GISAID database. Without this free data-sharing environment, this research would not have been possible. A full list of acknowledgments to contributing laboratories can be found in Supplementary Table S8.

References

Adepoju P. Nigeria responds to COVID-19; first case detected in sub-Saharan Africa. Nat Med 2020;26:444–8. https://doi.org/10.1038/d41591-020-00004-2.

Andersen KG, Rambaut A, Lipkin WI, Holmes EC, Garry RF. The proximal origin of SARS-CoV-2. Nat Med 2020;26:450–2. https://doi.org/10.1038/s41591-020-0820-9.

Cleemput S, Dumon W, Fonseca V, Abdool Karim W, Giovanetti M, Alcantara LC, et al. Genome Detective Coronavirus Typing Tool for rapid identification and characterization of novel coronavirus genomes. Bioinformatics 2020;36:3552–5.

Coronavirus Death Toll and Trends - Worldometer. n.d. https://www.worldometers.info/coronavirus/coronavirus-death-toll/ (accessed May 7, 2020).

COVID-19 WEEKLY EPIDEMIOLOGY BRIEF PROVINCES AT A GLANCE. n.d.

Deng X, Gu W, Federman S, Du Plessis L, Pybus O, Faria N, et al. A Genomic Survey of SARS-CoV-2 Reveals Multiple Introductions into Northern California without a Predominant Lineage. MedRxiv 2020:2020.03.27.20044925. https://doi.org/10.1101/2020.03.27.20044925.

Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. Lancet Infect Dis 2020;20:533–4. https://doi.org/10.1016/S1473-3099(20)30120-1.

Eden J-S, Rockett R, Carter I, Rahman H, de Ligt J, Hadfield J, et al. An emergent clade of SARS-CoV-2 linked to returned travellers from Iran. Virus Evol 2020;6. https://doi.org/10.1093/VE/VEAA027.

Flaxman S, Mishra S, Gandy A, Unwin HJT, Coupland H, Mellan T, et al. Estimating the number of infections and the impact of non-pharmaceutical interventions on COVID-19 in 11 European countries. Imperial College COVID-19 response team. March 2020 2020.

Gonzalez-Reiche AS, Hernandez MM, Sullivan M, Ciferri B, Alshammary H, Obla A, et al. Introductions and early spread of SARS-CoV-2 in the New York City area. MedRxiv

2020:2020.04.08.20056929. https://doi.org/10.1101/2020.04.08.20056929.

Grubaugh JRFEPEBHKSHYEAGWCBFVAFBTAAMJRRDNRCALWCCKI. Coast-to-coast spread of SARS-CoV-2 during the early epidemic in the United States. Cell Press 2020. https://doi.org/10.1016/j.cell.2020.04.021.

Grubaugh ND, Ladner JT, Lemey P, Pybus OG, Rambaut A, Holmes EC, et al. Tracking virus outbreaks in the twenty-first century. Nat Microbiol 2019;4:10–9. https://doi.org/10.1038/s41564-018-0296-2.

Issues with SARS-CoV-2 sequencing data - Novel 2019 coronavirus / nCoV-2019 Genomic Epidemiology - Virological. n.d. http://virological.org/t/issues-with-sars-cov-2-sequencing-data/473 (accessed May 8, 2020).

James Hadfield, Colin Megill, Sidney M Bell, John Huddleston, Barney Potter, Charlton Callender, Pavel Sagulenko, Trevor Bedford RAN. Nextstrain: real-time tracking of pathogen evolution. Bioinformatics 2018;Volume 34:4121–4123.

Kraemer MUG, Cummings DAT, Funk S, Reiner RC, Faria NR, Pybus OG, et al. Reconstruction and prediction of viral disease epidemics. Epidemiol Infect 2019;147. https://doi.org/10.1017/S0950268818002881.

Kucharski AJ, Russell TW, Diamond C, Liu Y, Edmunds J, Funk S, et al. Early dynamics of transmission and control of COVID-19: a mathematical modelling study. Lancet Infect Dis 2020.

Lemoine F, Entfellner J-BD, Wilkinson E, Correia D, Felipe MD, De Oliveira T, et al. Renewing Felsenstein's phylogenetic bootstrap in the era of big data. Nature 2018;556:452–6.

Leung KS-S, Ng TT-L, Wu AK-L, Yau MC-Y, Lao H-Y, Choi M-P, et al. A territory-wide study of early COVID-19 outbreak in Hong Kong community: A clinical, epidemiological and phylogenomic investigation. MedRxiv 2020:2020.03.30.20045740.

https://doi.org/10.1101/2020.03.30.20045740.

Li X, Zai J, Zhao Q, Nie Q, Li Y, Foley BT, et al. Evolutionary history, potential intermediate animal host, and cross- species analyses of SARS- CoV- 2. J Med Virol 2020;92:602–11. https://doi.org/10.1002/jmv.25731.

Linton NM, Kobayashi T, Yang Y, Hayashi K, Akhmetzhanov AR, Jung S, et al. Incubation period and other epidemiological characteristics of 2019 novel coronavirus infections with right truncation: a statistical analysis of publicly available case data. J Clin Med 2020;9:538.

Lu J, du Plessis L, Liu Z, Hill V, Kang M, Lin H, et al. Genomic Epidemiology of SARS-CoV-2 in Guangdong Province, China. Cell 2020. https://doi.org/10.1016/j.cell.2020.04.023.

Marivate V, Arbi R, Combrink H, de Waal A, Dryza H, Egersdorfer D, et al. Coronavirus disease (COVID-19) case data - South Africa 2020. https://doi.org/10.5281/ZENODO.3819126.

Max Roser, Hannah Ritchie, Esteban Ortiz-Ospina JH. "Coronavirus Pandemic (COVID-19)." Publ Online OurWorldInDataOrg 2020.

Moreno GK, Braun KM, Halfmann PJ, Prall TM, Riemersma KK, Haj AK, et al. Limited SARS-CoV-2 diversity within hosts and following passage in cell culture. BioRxiv 2020:2020.04.20.051011. https://doi.org/10.1101/2020.04.20.051011.

Msomi N, Mlisana K, de Oliveira T, Msomi N, Mlisana K, Willianson C, et al. A genomics network established to respond rapidly to public health threats in South Africa. The Lancet Microbe 2020;1:e229–30. https://doi.org/10.1016/S2666-5247(20)30116-6.

Munnink BBO, Nieuwenhuijse DF, Stein M, O'Toole A, Haverkarte M, Mollers M, et al. Rapid SARS-CoV-2 whole genome sequencing for informed public health decision making in the Netherlands. BioRxiv 2020:2020.04.21.050633. https://doi.org/10.1101/2020.04.21.050633.

Nguyen L-T, Schmidt HA, Von Haeseler A, Minh BQ. IQ-TREE: a fast and effective

stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol 2015;32:268–74.

Nishiura H, Linton NM, Akhmetzhanov AR. Serial interval of novel coronavirus (COVID-19) infections. Int J Infect Dis 2020.

Quick J. Forked from Ebola virus sequencing protocol 2020. https://doi.org/10.17504/protocols.io.bbmuik6w.

Rambaut A, Holmes EC, Hill V, OToole A, McCrone J, Ruis C, et al. A dynamic nomenclature proposal for SARS-CoV-2 to assist genomic epidemiology. BioRxiv 2020:2020.04.17.046086. https://doi.org/10.1101/2020.04.17.046086.

Sagulenko P, Puller V, Neher RA. TreeTime: Maximum-likelihood phylodynamic analysis. Virus Evol 2018;4:vex042–vex042. https://doi.org/10.1093/ve/vex042.

Shu Y, McCauley J. GISAID: Global initiative on sharing all influenza data–from vision to reality. Eurosurveillance 2017;22:30494.

Sohrabi C, Alsafi Z, O'Neill N, Khan M, Kerwan A, Al-Jabir A, et al. World Health Organization declares global emergency: A review of the 2019 novel coronavirus (COVID-19). Int J Surg 2020.

Tavaré S, Miura RM. Some Mathematical Questions in Biology: DNA Sequence Analysis Lectures on Mathematics in the Life Sciences 1986.

Vilsker M, Moosa Y, Nooij S, Fonseca V, Ghysens Y, Dumon K, et al. Genome Detective: an automated system for virus identification from high-throughput sequencing data. Bioinformatics 2019;35:871–3.

Wallinga J, Lipsitch M. How generation intervals shape the relationship between growth rates and reproductive numbers. Proc R Soc B Biol Sci 2007;274:599–604.

Xavier J, Giovanetti M, Adelino T, Fonseca V, da Costa AVB, Ribeiro AA, et al. The ongoing COVID-19 epidemic in Minas Gerais, Brazil: insights from epidemiological data

and SARS-CoV-2 whole genome sequencing. MedRxiv 2020.

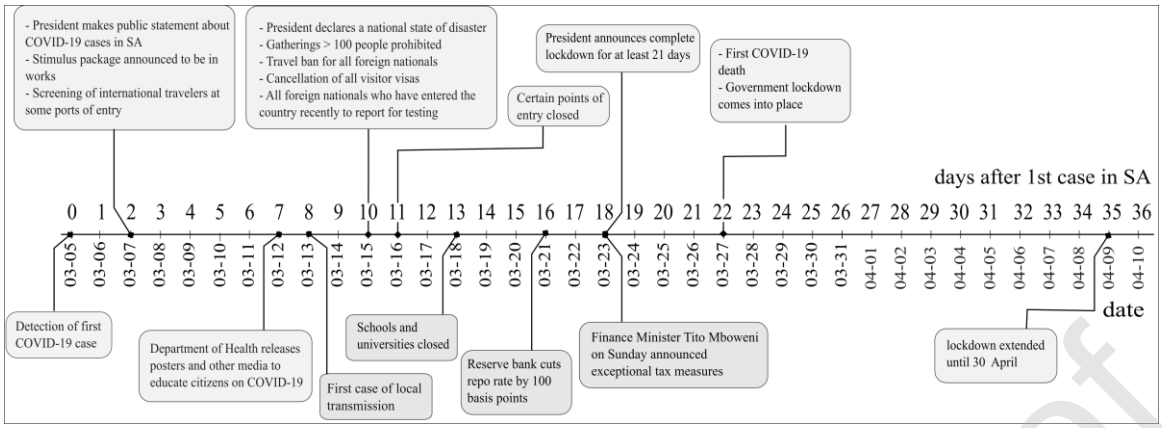


Figure 1: Timeline of measures implemented in South Africa from the first detected COVID-19 case on 5th of March to the expansion of the lockdown in April 2020.

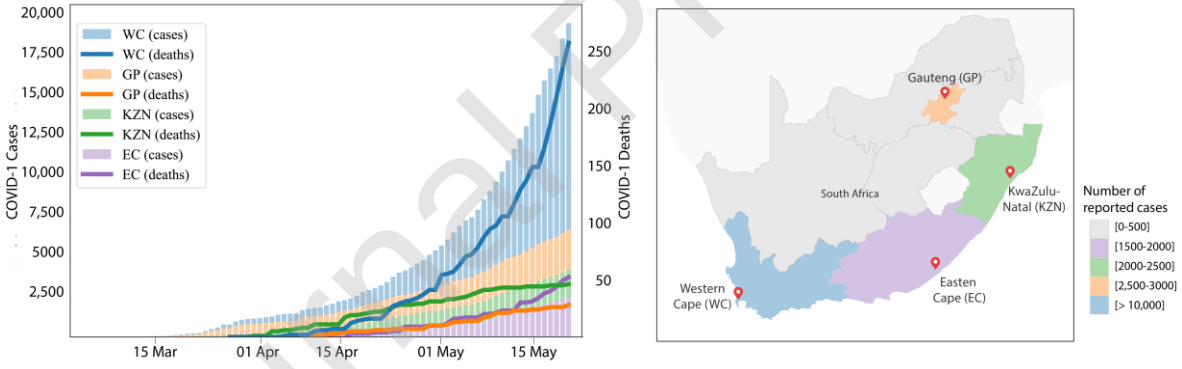


Figure 2: Summary of the COVID-19 epidemic in South Africa. A) Numbers of COVID-19 cases and deaths in the Western Cape (WC), Gauteng (GP), KwaZulu-Natal (KZN) and the Eastern Cape (EC). B) Geographic map showing the location of South African provinces.

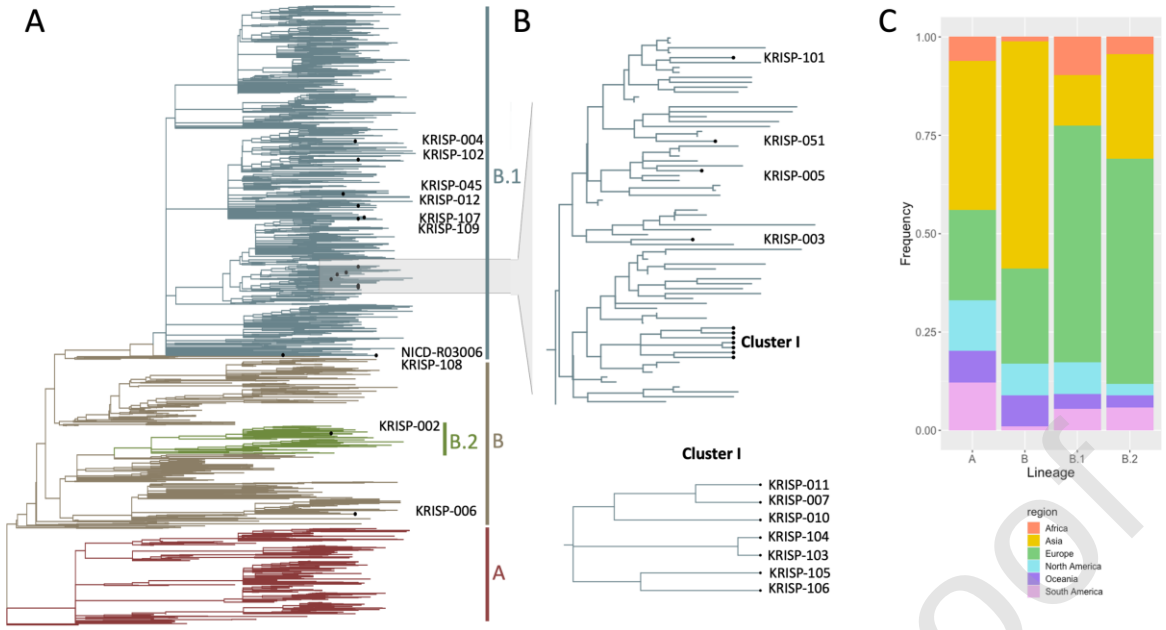


Figure 3: Phylogenetic analysis. (A) A time scaled Maximum likelihood tree of 1849 sequences including 21 genotypes from KwaZulu-Natal, South Africa. Major lineages of SARS-CoV-2 are labelled. B) Monophyletic cluster of KZN sequences. (C) Stacked barplot showing the lineage breakdown of the dataset by region.

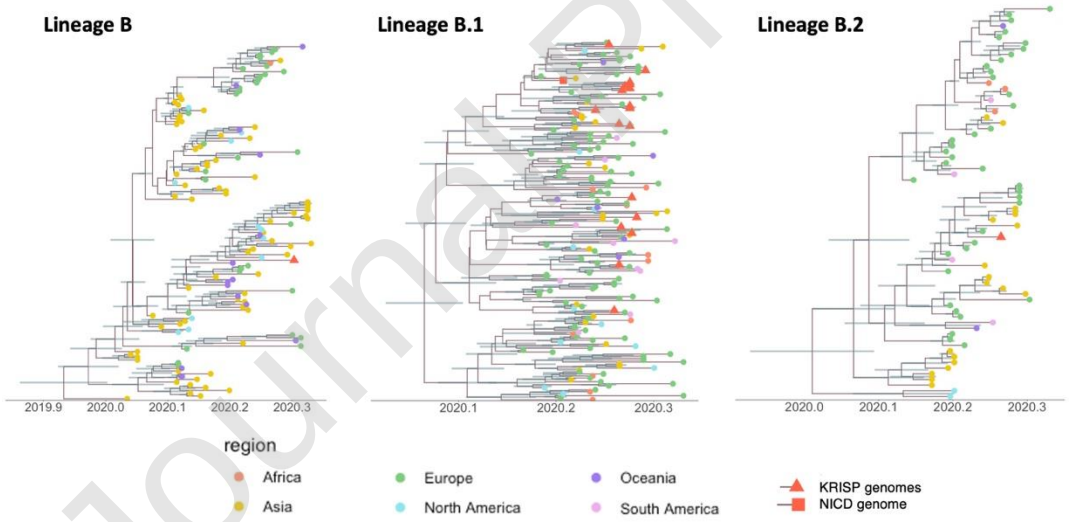


Figure 4: Time stamped phylogenetic trees of the three lineages of SARS-CoV-2 found in KwaZulu-Natal (KZN). The genomes produced in this study are marked with a red triangle, and the NICD genome by a red square. The geographic region of the other sequences is marked with coloured circles.