Comparison of SARS-CoV-2 sequencing using the ONT GridION and the Illumina MiSeq

By

Derek Tshiabuila (215024685)

A Dissertation by manuscripts submitted in fulfillment of the requirement of the degree of Master of Medical Science (MMedSc.)

Department of Virology, School of Laboratory Medicine and Medical Sciences, College of Health Sciences, University of KwaZulu-Natal, Durban, South Africa

Supervisor:

Prof Tulio de Oliveira

June 13, 2022

DECLARATION

I, Mr. Derek Tshiabuila, declare as follows:

1 That the work described in this dissertation has not been submitted to UKZN or other tertiary institutions for purposes of obtaining an academic qualification, whether by myself or any other party.

2 That my contribution to the project was as follows:

I conceptualized the project and performed the formal analysis and data curation as stipulated in the methodology section. Also, I performed the visualization of the data and wrote and reviewed this dissertation.

3 That the contributions of others to the project were as follows:

- Prof. Tulio de Oliveira: Supervisor
- Dr. Jennifer Giandhari: Methodology
- Dr. Sureshnee Pillay: Methodology
- Dr. Emmanuel James San: Visualization

4 This dissertation does not contain other persons' writing unless specifically acknowledged as being sourced from other researchers. Where other written sources have been quoted, then:

- Their words have been re-written but the general information attributed to them has been referenced.
- Where their exact words have been used, their writing has been placed inside quotation marks and referenced.

Signed  Date 29/03/2022

DEDICATION

This master's dissertation is dedicated to my late father who has been my inspiration and encouragement and has provided for my every need. I will forever love and cherish you. I dedicate the dissertation to my mother who continues to support me and loves and cares for me. You will forever remain an important part of my life. I thank the Lord almighty for providing for me and continuing to bless and protect me throughout the years. All thanks go to friends and family for the prayers and well wishes.

ACKNOWLEDGEMENTS

I would like to acknowledge professor Tulio de Oliveira for providing me with the opportunity to enroll for a master's degree through the Kwazulu-Natal Research Innovation and Sequencing Platform and for being an awesome supervisor and mentor. A special thank you goes out to the KRISP laboratory and bioinformatics teams for the constructive criticism and guidance provided throughout this academic journey. Many thanks go out to my family for the love and support that they have provided me with throughout my studies. Finally, a huge thank you to the Lord Almighty who has never left my side and has continued to bless me both academically and socially.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

+ssRNA	Positive Single-Stranded Ribonucleic Acid
β-CoV	Beta-coronavirus
ACE2	Angiotensin-Converting Enzyme 2
BREC	Biomedical Research Ethics Committee
COVID-19	Corona Virus Disease 2019
Ct	Cycle Threshold
DNA	Deoxyribonucleic Acid
GISAID	Global Initiative on Sharing All Influenza Data
KRISP	Kwazulu-Natal Research Innovation and Sequencing Platform
MERS	Middle Eastern Respiratory Syndrome
ML	Maximum Likelihood
NGS	Next Generation Sequencing
NGS-SA	Network for Genomic Surveillance in South Africa
NHLS	National Health Laboratory Service
NICD	National Institute for Communicable Diseases
ONT	Oxford Nanopore Technology
ORF1a/1b	Open Reading Frame 1a/1b
PCR	Polymerase Chain Reaction
RBD	Receptor Binding Domain
RNA	Ribonucleic Acid
RT-PCR	Real-Time Polymerase Chain Reaction
SARS-CoV-2	Severe Acute Respiratory Syndrome Corona Virus 2
SBS	Sequencing by Synthesis
VOC	Variant of Concern
VOI	Variant of Interest
WGS	Whole Genome Sequencing
WHO	World Health Organization

ABSTRACT

Corona Virus Disease 2019 (COVID-19) is an ongoing pandemic that has spread rapidly around the world and has seen over 431 000 000 identified cases and 5 930 000 deaths caused by this disease by the end of January 2022. Many viral lineages have arisen from Severe Acute Respiratory Syndrome Corona Virus 2 (SARS-CoV-2) as public health measures from numerous countries have failed to contain the spread of the virus. Sequencing of SARS-CoV-2 has enabled the identification and classification of the viral lineages, while real-time tracking of the emergence and spread of these lineages has been facilitated by the open sharing of genomic surveillance data and collaborative online platforms. Several studies have suggested that various mutations may have a functional effect on the virus, such as a substitution in the spike protein (D614G) may result in increased transmissibility whilst an N439K substitution in the receptor-binding domain (RBD) may assist in neutralizing monoclonal antibodies. It is therefore necessary that a fast and reliable sequencing technology be used to rapidly and correctly produce SARS-CoV-2 genomes that can be used to identify viral lineages. Many sequencing laboratories have begun using Nanopore sequencing as it promises high throughput, real-time sequencing, at an affordable cost and many of their sequencing platforms allow for portability. The sequencing technology has, however, not been verified to produce consensus SARS-CoV-2 genomes that are comparable to Illumina Sequencing which is currently the gold standard Next Generation Sequencing (NGS) technology for SARS-CoV-2 sequencing. In this study, we compared the Illumina and Nanopore sequencing platforms by comparing the SARS-CoV-2 genomes produced by the Illumina MiSeq and Oxford Nanopore Technology (ONT) GridION X5. The results show that the GridION is currently unsuitable for SARS-CoV-2 genomic surveillance as consensus genomes produced by the platform have a lower quality than those produced by the MiSeq which reduces the reliability of the data obtained from the genomes. These results can be used to better understand the Nanopore sequencing technology and how it differs from the Illumina technology which will help in updating the Nanopore technology to produce consensus genomes at a faster rate than the Illumina technology whilst still having a similar quality.

CHAPTER 1: INTRODUCTION

1.1 Background

COVID-19 is a viral pneumonia that started a pandemic in the Hubei province of China in December 2019 (Khan et al., 2020). The disease is caused by a type of coronavirus known as SARS-CoV-2 that belong to the clade *Riboviria*, kingdom *Orthornavirae*, phylum *Pisuviricota*, class *Pisoniviricetes*, order *Nidovirales*, suborder *Cornidovirineae*, family *Coronaviridae*, subfamily *Orthocoronavirinae*, genus *Betacoronavirus*, subgenus *Sarbecovirus* (Zhou et al., 2020).

SARS-CoV-2 is transmitted via droplet nuclei with an incubation period of 2 – 14 days (Lu et al., 2020). Common symptoms include fever, cough, shortness of breath, fatigue, muscle or body aches, headache, loss of taste or smell, nausea or vomiting, and/or diarrhea (Fiorillo et al., 2020). SARS-CoV-2 spread rapidly across the globe and on 11 March 2020, it was declared a global pandemic by the World Health Organization (WHO) (World Health, 2020b).

To help curb the spread of SARS-CoV-2 and end the COVID-19 pandemic, infected individuals needed to be rapidly screened and isolated to prevent transmission chains from occurring (Seemann et al., 2020). SARS-CoV-2 sequencing has allowed for the rapid identification of the virus and the development of new diagnostic tests and other tools as sequencing provides the genotypic information of the specific strain infecting the patient allowing for a rapid response to the COVID-19 pandemic (Seth-Smith et al., 2019, St Hilaire et al., 2020). NGS is a technique used to study pathogens by determining the order of nucleotides in the genomes. It has been crucial in the COVID-19 pandemic as it has allowed us to better understand the epidemiology of the virus and identify mutations critical in the development of vaccines and subsequent intervention methods. Obtaining sequences promptly and sequencer portability are major challenges associated with whole-genome sequencing (WGS) (Shaw and Sugden, 2018).

Sequencing technologies that have been used for SARS-CoV-2 include Sanger, Illumina, ION torrent, and Nanopore sequencing. The most widely used technology, however, is Illumina sequencing (GISAID, 2022). Relatively long sequencing times and the high costs associated with library preparation for high-throughput sequencing are limitations associated with Illumina sequencing (Gohl et al., 2020). This is overcome with Nanopore sequencing which sequences in real-time and is a long-read sequencing technology. The ONT flongle and minion are sequencing platforms that allow for sequencer portability but the technology is limited by the high number of false negatives, such as incorrect basecalling resulting in wrong variant calls, and low sensitivity to mutational changes (Wang et al., 2020b).

The purpose of this study was to compare the quality of consensus genomes produced by the ONT GridION and Illumina MiSeq for 2608 SARS-CoV-2 positive nasopharyngeal swabs received by the Kwazulu-Natal Research Innovation and Sequencing Platform (KRISP) for routine genomic surveillance. Amplicons for both the MiSeq and the GridION were generated using the ARTIC polymerase chain reaction (PCR) tiling method (Quick, 2020). Libraries for the MiSeq were prepared using the Nextera Flex deoxyribonucleic acid (DNA) library Preparation method with Nextera DNA CD Indexes, whilst libraries for the GridION were prepared using the Ligation sequencing kit 96 barcode plate (Pillay et al., 2020). Sequence alignment for the MiSeq was performed using Genome Detective, whilst the ARTIC protocol was used for the GridION (Cleemput et al., 2020, Nick Loman, 2020). Consensus genomes produced were uploaded to Nextclade Online Tool (https://clades.nextstrain.org/) for sequence analysis and the results were compared **(Figure 1)**. Although GridION sequencing was able to produce complete SARS-CoV-2 genomes, the sequence quality observed was not as good as that obtained with MiSeq sequencing

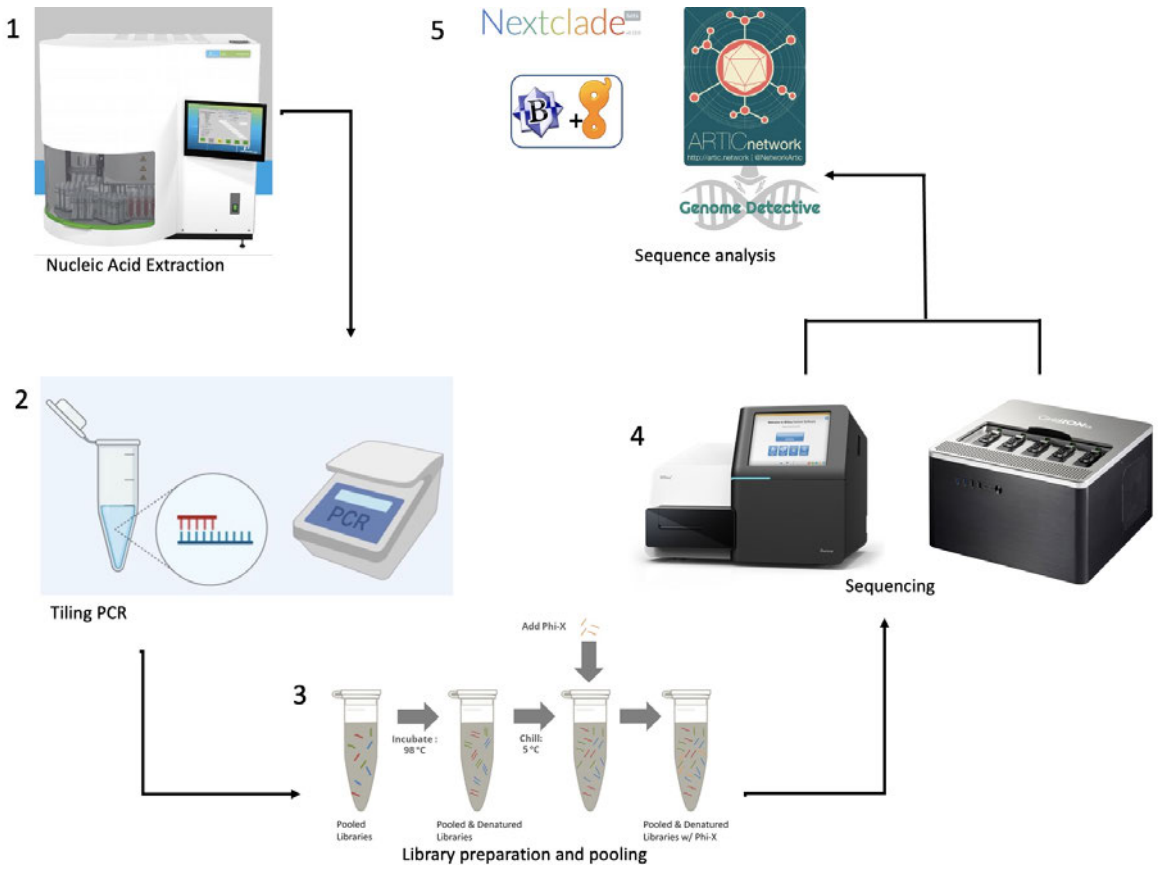


Figure 1. SARS-CoV-2 sequencing workflow for the GridION and the MiSeq: The figure above illustrates the workflow used for sequencing SARS-CoV-2 on the Illumina MiSeq and the ONT GridION. Nucleic acid extraction was performed on the Chemagic 360 nucleic acid extractor (Perkin Elmer) and the tiling PCR method was used for cDNA synthesis. Sequencing libraries for the MiSeq were prepared using the Nextera Flex DNA Library Preparation and Nextera DNA CD Indexes and

pooled prior to sequencing whilst sequencing libraries for the GridION were prepared using the Ligation sequencing kit 96 barcode plate. GridION libraries were also pooled prior to sequencing. Genome Detective was used to assemble sequences from the MiSeq, whilst the ARTIC network was used to assemble sequences from the GridION. Consensus genomes were then compared on the Nextclade online analysis tool.

1.2 Literature Review

SARS-CoV-2

SARS-CoV-2 is a ribonucleic acid (RNA) virus belonging to the order *Nidovirales*, suborder *Cornidovirineae*, family *Coronaviridae*, subfamily *Orthocoronavirinae*, genus *Betacoronavirus* (β-CoV), subgenus *Sarbecovirus*, and has structural similarities to other coronaviruses, such as SARS-CoV and Middle Eastern Respiratory Syndrome (MERS) (Zhu et al., 2020). The virus has a positive-sense, single-stranded RNA (+ssRNA) surrounded by a nucleocapsid (N) protein, while envelope (E), membrane (M), and spike (S) proteins form the viral envelope, see Figure 2 (Rahimi et al., 2021). These proteins are key to the virus's replication and infectivity and are thus the focus of studies on the viral structure and drug discovery (Scudellari, 2020). The viral genome is approximately 30 kb in length and comprises a structural gene unit that codes for the S, E, M, and N proteins and two large open reading frames (ORF1a and ORF1b) which encode sixteen non-structural proteins, including RNA-dependent RNA polymerase (Kim et al., 2020).

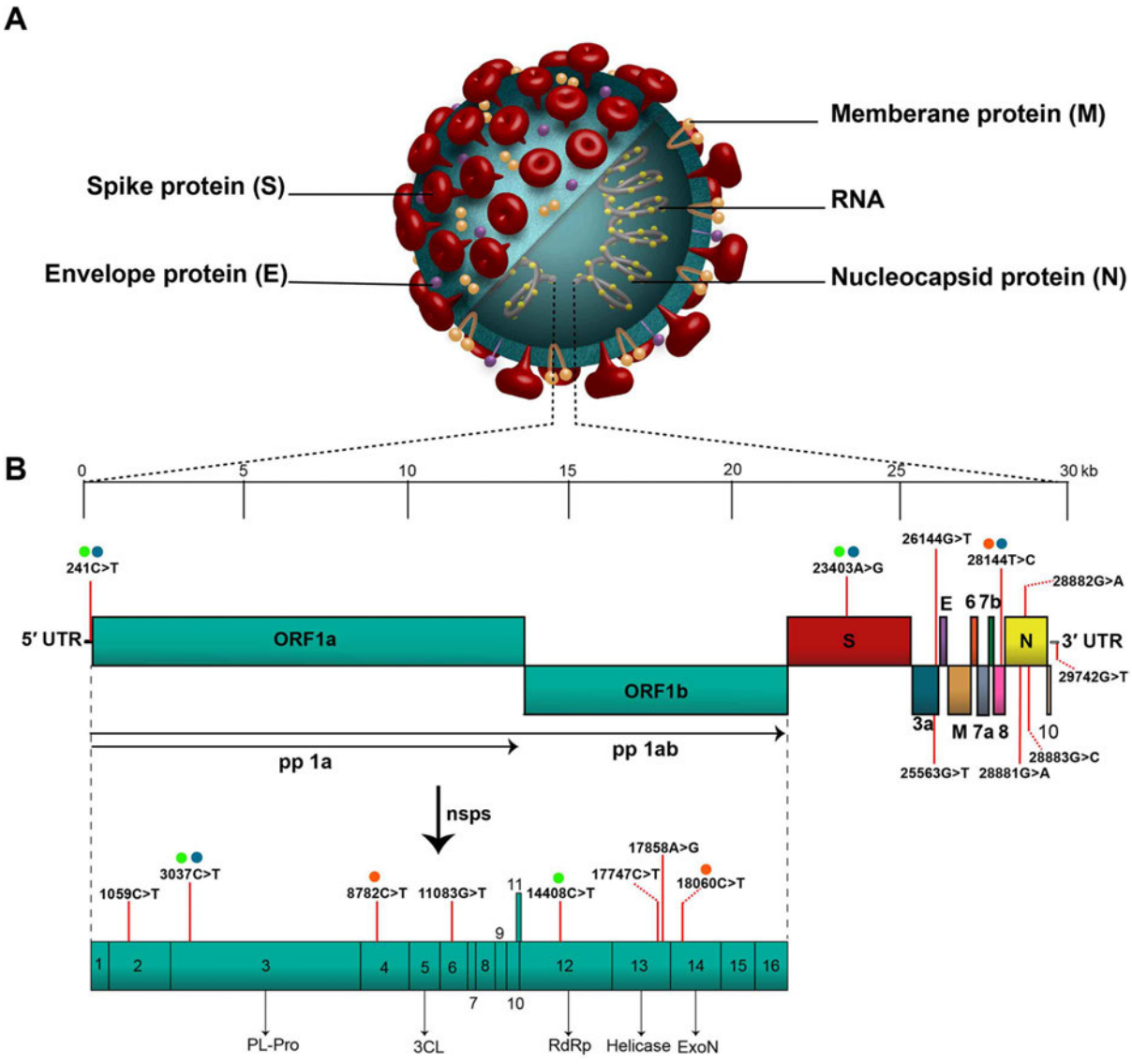


Figure 2. SARS-CoV-2 schematic figure (Rahimi et al., 2021). Structural proteins such as the Spike (S1 and S2), Nucleocapsid (N), Membrane (M), and Envelope (E) form the viral structure. The S, M, and E proteins are embedded in the viral envelope. *UTR, 5′ untranslated region; *OFR, open reading frame; *nsp, non-structural protein

Each of the viral spike proteins is made up of three identical spike glycoproteins that consist of two functional subunits (Walls et al., 2020). The subunits mediate the binding to the human Angiotensin-Converting Enzyme 2 (ACE2) receptor, found on lung and body cells, and the fusion of the cellular and viral membranes (Wan et al., 2020). Once the virus particle is fused to the cell, the viral genomic RNA is injected into the cell and the host cellular machinery translates the RNA into protein chains. Following viral entry, two large open reading frames, ORF1a and ORF1b, are immediately transcribed resulting in polyproteins pp1a and pp1ab. The polyproteins are processed into individual non-structural proteins that form the viral replication and transcription complex (V'kovski et al., 2021). Viral replication is then achieved by cleaving the protein chains, with a protease, into functional units to make new proteins required for replication (Zhang et al., 2020).

SARS-CoV-2 mutates at a relatively consistent and moderate rate, equating to approximately 33 changes per year across the viral genome (Candido et al., 2020, Laamarti et al., 2020). Viral genomes are used to identify regions of the genome that are subjected to greater selection pressure. This is important as it assists in the development of effective therapeutics, vaccines, and diagnostics that target unchanged, conserved parts of the viral genome (Wang et al., 2020a). Phylogenetic analysis of sequencing data classifies genomes into different sub-groups based on genetic similarity and observed mutations (Rambaut et al., 2020, Tegally et al., 2021b). There have been a number of variants of interest (VOIs) and five variants of concern (VOCs) that have been identified since December 2020. These have been summarized in Table 1.

A viral lineage is defined as a group of closely related viruses with a common ancestor. Lineages are further divided into clades by the presence of signature mutations (Tegally et al., 2021b). All clades share a common ancestor and all descendants of the clade and are named after the frequency of the clade has exceeded 20 %. A viral variant is a genome genetically distinct from the reference genome as it contains one or more mutations and a strain is a variant that has unique and stable phenotypic characteristics. The D614G variant carries a mutation in the spike glycoprotein and was first detected in early March 2020 at a significant level and over two months spread to global dominance (Korber et al., 2020). The mutation has been shown to enhance viral infectivity, replication fitness, and early transmission (Hou et al., 2020). A mutation in the RBD of the spike protein, Y453F, first identified in Denmark, showed an increased binding infinity for the ACE2 receptor (Lauring and Hodcroft, 2021).

The N501Y variant accumulated 17 lineage-defining mutations before its detection in early September 2020 and affects the RBD of the spike protein. Population genetic models suggest that it spreads 56 % more quickly than other lineages (Davies et al., 2021).

Currently, Delta (B.1.617.2 and AY lineages) and Omicron (B.1.1.529 and BA lineages) are the latest VOCs. It is important to study the functional effect that variants have on the virus and track their speed through different populations (Prevention, 2021).

Table 1. Summary of currently designated SARS-CoV-2 VOCs.

WHO Label	Pango lineage	GISAID clade	Nextstrain clade	Earliest documented samples	Date of designation
Alpha	B.1.1.7	GRY	20I (V1)	The United Kingdom, Sep-2020	18-Dec-2020
Beta	B.1.351	GH/501Y.V2	20H (V2)	South Africa, May-2020	18-Dec-2020
Gamma	P.1	GR/501Y.V3	20J (V3)	Brazil, Nov-2020	11-Jan-2021
Delta	B.1.617.2	GK	21A, 21I, 21J	India, Oct-2020	11-May-2021
Omicron	B.1.1.529	GRA	21K, 21L 21M	Multiple countries, Nov-2021	26-Nov-2021

Whole-Genome Sequencing

Since the beginning of the COVID-19 pandemic, sequencing technologies have been used to understand the virus's biology and epidemiology. Sequencing describes the process whereby the nature and order of nucleic acids for samples are converted into data for analysis (Behjati and Tarpey, 2013). NGS provides the highest resolution information regarding pathogen genomes as it allows full nucleotide sequences to be read and the discovery of new genomic variation at scale (Slatko et al., 2018). As of 05

March 2022, 9 021 143 SARS-CoV-2 consensus genomes had been shared via the Global Initiative on Sharing All Influenza Data (GISAID) (GISAID, 2022).

Various sequencing technologies are currently used for SARS-CoV-2 sequencing (GISAID, 2022). These can be broken down into groups referred to as generations: first (Sanger), second (high throughput), and third (long-read) generation sequencing (Heather and Chain, 2016). NGS is used to refer to second and third-generation sequencing technologies and includes Sequencing by Synthesis (SBS) (Illumina), Ion Torrent (Thermo Fisher Scientific), Single-molecule real-time (Pacific Bioscience), and Nanopore sequencing (ONT) (Liu et al., 2012).

NGS surveillance assists in identifying disease origins within human populations, management of outbreaks, identifying transmission chains, examining viral population structure, and tracking disease prevalence. It also plays a significant role in monitoring the trends in COVID-19 and deaths at a national and global level, monitors the spread and evolution of SARS-CoV-2 and the impacts it has on disease, and enables the rapid detection, isolation, testing and management of cases. Furthermore, surveillance helps to evaluate the impact of the pandemic on healthcare systems and plays a significant role in detecting and containing clusters and outbreaks, especially among vulnerable populations. NGS can also be used for diagnostic purposes to identify pathogens early in the outbreak, to identify regions of the genome for use in diagnostic testing, and in the detection of co-infections. As COVID-19 cases are still increasing, the WHO has stated that epidemiological analyses should include the identification and reporting of new cases, and the inclusion of consensus genomes within 24 hours of new infections (World Health, 2020a).

Illumina Sequencing

Illumina MiSeq sequencing is currently the most widely used SARS-CoV-2 sequencing technology (GISAID, 2022). The technology makes use of SBS which provides high throughput, short-read sequencing. With SBS, as new DNA forms, fluorescently labelled bases are incorporated into the DNA. The fluorescent tags are washed off as bases are incorporated and more modified bases are added (Liu et al., 2012). This is repeated until the maximum number of cycles, and thus read length, is achieved. During library preparation, which is performed before SBS, RNA is converted to cDNA via bridge amplification. Bridge amplification is a process by which complementary DNA strands, otherwise known as the reverse strand, are generated in a flow cell for further sequencing and analysis. The flow cell is coated with two types of oligos that are complementary to the two adapters of the fragment strand, respectively. As the fragment strand is added to the flow cell, it hybridizes into one of the oligos on the surface of the cell. A polymerase then moves along the strand and in the process creates the complementary DNA strand. The now double-stranded DNA is denatured and the forward strand, the original strand, is washed away. The reverse strand then folds over and hybridizes to the second oligo

using its adaptor region. A polymerase attaches to the reverse strand and generates a complementary strand that matches the forward strand. A double-stranded bridge is formed. The bridge is denatured resulting in two single-stranded copies of the DNA, forward and reverse strand, attached to the flow cell. The denaturation and extension processes are repeated to result in the amplification of millions of fragments forming localized clusters on the flow cell (Pettersson et al., 2009).

Illumina has produced a range of platforms to cover various sequencing applications. A few examples of Illumina sequencing platforms are summarized in Table 2 (Petersen et al., 2019).

Table 2. Overview of Illumina sequencing platforms.

	iSeq 100	MiniSeq	MiSeq	NextSeq 500	NextSeq 2000	NovaSeq 6000
Run time (hours)	9.5 – 19	4 – 24	4 – 55	12 – 30	24 – 48	13 – 44
Maximum data output	1.2 Gb	7.5 Gb	15 Gb	120 Gb	300 Gb	6 Tb
Maximum read length	1 x 150 bp	2 x 150 bp	2 x 300 bp	2 x 150 bp	2 x 150 bp	2 x 250 bp
Description	Illumina's smallest benchtop sequencer released in 2018. Low cost and low capacity.	Benchtop sequencer released in 2016. Low cost and low capacity.	Mid-range, benchtop sequencer providing longest reads. Released in 2011.	High-throughput, mid-ranged, benchtop sequencer. Released in 2015	High-cost, high-capacity benchtop sequencer. Released in 2020	Highest throughput and most expensive sequencer. Released in 2017
Estimated Cost	$19 900	$20 000	$125 000	$250 000	$335 000	$985 000

Advantages of Illumina sequencing include comparatively low-cost sequencing at high throughput, the availability of SARS-CoV-2 specific protocols and tools, high accuracy, a commonly used system in around 155 countries, and high levels of sample multiplexing. The limitations, however, are the longer sequencing run times, most platforms are large and expensive to purchase and require specialized

infrastructure to reduce preventable harm, and the relatively short reads may decrease the accuracy in certain genomic regions (Mantere et al., 2019).

Nanopore Sequencing

Nanopore sequencing makes use of long-read single-molecule sequencers that utilize nanopores to read longer contiguous strands of DNA than other NGS technologies. Reads produced may range between 10 000 and 100 000 bp in length and have the potential to produce molecules over 100 000 bp long (de Lannoy et al., 2017). A major advantage of long-read sequencing is that longer reads are more likely to be distinct from other reads. This allows for the sequencing of highly polymorphic or highly repetitive regions as the assemblies are less ambiguous. With some systems, amplification-free sequencing can be achieved. This facilitates the examination of epigenetic modifications and may remove some amplification bias (van Dijk et al., 2014).

ONT has produced a range of sequencers based on nanopores. These systems are designed to be relatively mobile, generate ultra-long reads, and require less experience and expertise to operate (Nicholls et al., 2019). ONT sequencing platforms are summarized in Table 3.

Table 3. Overview of ONT sequencing platforms.

	Flongle	MinION Mk1B	MinION Mk1C	GridION Mk1	PromethION 24	PromethION 48
Run time	1 min – 16 hrs.	1 min - 48 hrs.	1 min - 48 hrs.	1 min - 48 hrs.	1 min - 72 hrs.	1 min - 72 hrs.
Maximum data output	2 Gb	50 Gb	50 Gb	250 Gb	5.2 Tb	10.5 Tb
Maximum read length	Read lengths can exceed 2 Mb and are dependent on the length of the target molecule.					
Description	Lowest cost, reduced adapter for MinION.	Low cost, mobile, long-read sequencer.	Low cost, mobile, long-read sequencer with built-in analysis platform.	Medium capacity, desktop, long-read sequencer with built-in analysis platform.	High capacity, desktop, long-read sequencer.	Highest cost, high capacity, desktop long-read sequencer.
Estimated Cost	$90 per flow cell	$1 000	$4 900	$49 955	$20 000/year Subscription	$20 000/year Subscription

Prior to the COVID-19 pandemic, the MinION had been extensively used for microbial processing. Its portability allowed it to be used for outbreaks such as Ebola, for surveillance, and diagnostics. Hospital settings have also used the sequencer to monitor the spread of nosocomial infections (Xu et al., 2020). The ARTIC network has published various protocols for the preparation and sequencing of SARS-CoV-2 using Nanopore sequencing (Network, 2020).

ONT sequencing is advantageous, as it is rapid and flexible, sequences in real-time, is relatively low-cost at low throughput, allows for mobile sequencing, has a simple user interface and analysis platform, and has well-established sequencing protocols for SARS-CoV-2. The limitations are that high-throughput sequencing is expensive as barcodes are limited, raw output files are relatively large making file storage difficult, and there is a high error rate in homopolymeric regions (Nicholls et al., 2019).

1.3 Research Question

Will the ONT GridION produce SARS-CoV-2 consensus genomes that are comparable to consensus genomes produced by the Illumina MiSeq?

1.4 Hypothesis

The ONT GridION will produce SARS-CoV-2 consensus genomes with a similar quality, when comparing genome coverage, concordance to reference, and the number of mutations and gaps, to consensus genomes produced by the Illumina MiSeq.

1.5 Aims and Objectives

Aim: To compare the sequence quality of consensus genomes produced by the MiSeq and the GridION for positive COVID-19 nasopharyngeal and oropharyngeal swabs and determine the effect of cycle threshold (Ct) score on sequencing coverage.

Objectives:

1. To compare the coverage, concordance, mutations, and gaps for consensus genomes obtained from the GridION and the MiSeq for COVID-19 positive swabs.
2. To determine the effect of sample Ct score on the genome coverage and number of reads of SARS-CoV-2 sequences produced by the GridION and the MiSeq.

1.6 Methodology

The study population was comprised of 2608 COVID-19-positive male and female patients whose nasopharyngeal and oropharyngeal swabs, collected between December 2020 and March 2021, underwent RNA extraction on the automated Chemagic 360 system (Perkin Elmer) using the NA/gDNA kit (Figure 3).

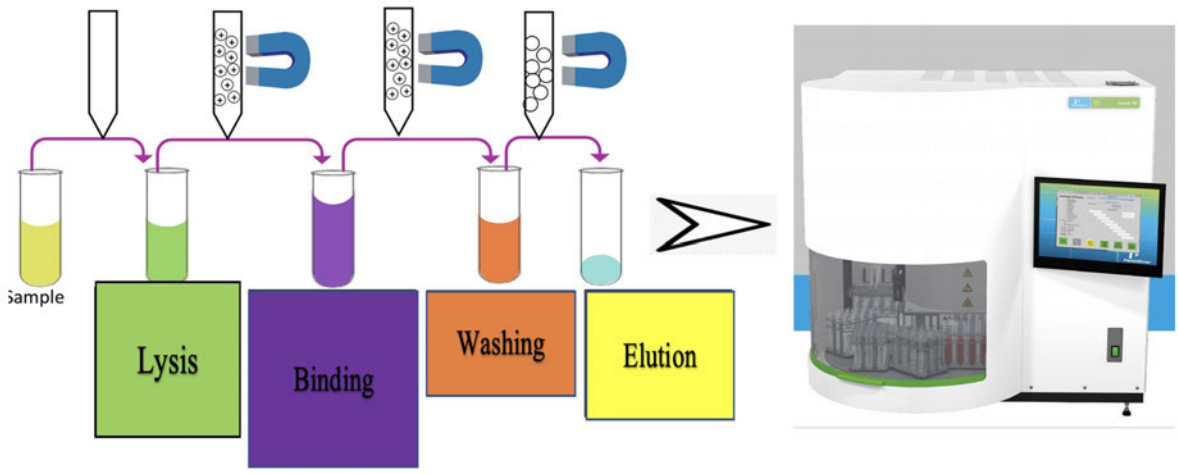


Figure 3. Diagrammatic representation of RNA extraction using Chemagic 360.

Figure 3 illustrates the RNA extraction protocol for naso- and oropharyngeal swabs using the Chemagic 360. Swabs underwent lysis using lysis buffer and proteinase K, binding was via the silica magnetic beads, and washing and elution were performed using the wash and elution buffer respectively.

Tiling PCR was then performed for complementary cDNA synthesis using SuperScript IV reverse transcriptase (Life Technologies) and the ARTIC protocol. Illumina library preparation was achieved using the Nextera DNA Flex Library Prep Kits as per the manufacturer's instructions, and sequencing was performed on the Illumina MiSeq. The ARTIC protocol was used as per the manufacturer's instructions to prepare libraries on FLO-MIN106 flowcells, and sequencing was performed using the ONT GridION X5 (Figure 4).

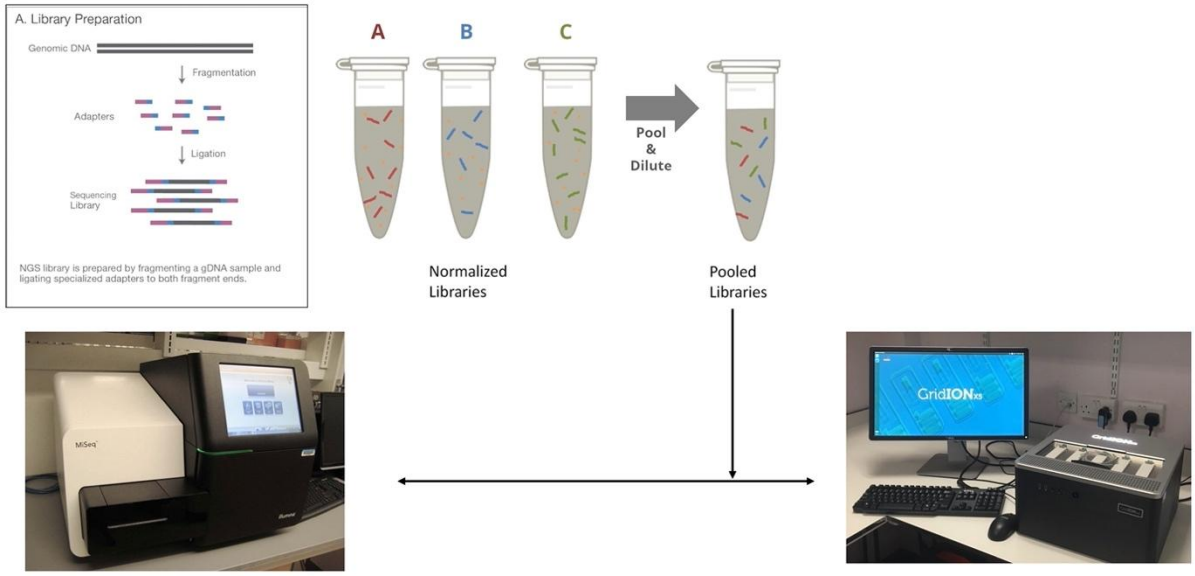


Figure 4. Library preparation and sequencing on the Illumina MiSeq and ONT GridION X5.

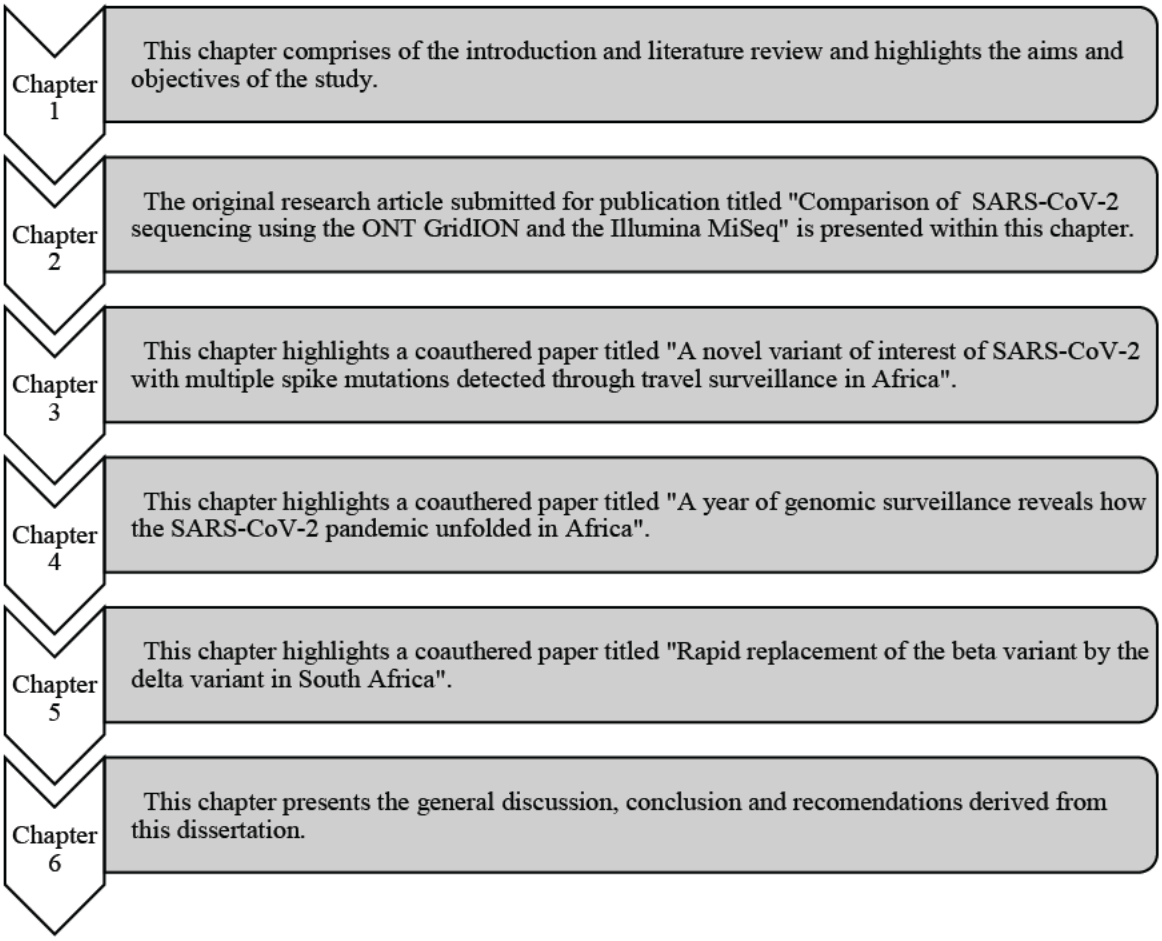
Figure 4 illustrates the library preparation and sequencing workflow for the MiSeq and the GridION X5. cDNA synthesis was achieved using SuperScript IV reverse transcriptase. The NEXTERA DNA FLEX and ARTIC protocol were used for the MiSeq and GridION, respectively. Libraries were normalized before being pooled and sequenced on the Illumina MiSeq and ONT GridION X5.

The Nextclade online tool was used to analyze consensus genomes produced by both the Illumina MiSeq and the ONT GridION. Wilcoxon tests were used to compare the sequence quality, coverage, number of mutations, and type of mutation detected by both platforms. Sample Ct score was then correlated to sequence coverage and the number of reads for both platforms.

This study was approved by the University of KwaZulu–Natal Biomedical Research Ethics Committee (BREC) (BREC/00001195/2020) and (BREC/00002764/2021) found in Annexures 1 and 2 respectively.

1.7 Structure of Dissertation

This dissertation is comprised of 6 chapters with an overall aim of comparing two SARS-CoV-2 sequencing platforms to identify a sequencing platform that can rapidly and accurately produce SARS-CoV-2 consensus genomes. The reference list found under the heading "REFERENCES" only applies to Chapter 1 and Chapter 6 of this dissertation.



Bridging Chapters 1 and 2

Chapter 1 provides a general overview of the SARS-CoV-2 viral structure and a summary of the virus's replication cycle and host immune evasion. It also assesses the different VOIs and VOCs and how the WGS of the virus was achieved. The advantages and disadvantages of SARS-CoV-2 sequencing using Illumina and Nanopore sequencing platforms are highlighted and the different sequencing platforms are compared. COVID-19 is an ongoing pandemic that has seen the SARS-CoV-2 virus mutate multiple times resulting in different VOIs and VOCs. Rapid identification of these variants may increase the understanding of transmission chains. Chapter 2 compares the Illumina MiSeq and ONT GridION to determine whether sequences produced by the GridION can be accurately used to identify variants swiftly. The article was accepted by *BMC Genomics* on 08 April 2022 and published on 22 April 2022 (https://doi.org/10.1186/s12864-022-08541-5). In this study, I assisted in RNA extraction and sequencing and performed sequence analysis and curation. I also wrote the original draft used for submission.

CHAPTER 2

COMPARISON OF SARS-COV-2 SEQUENCING USING THE ONT GRIDION AND THE ILLUMINA MISEQ

Comparison of SARS-CoV-2 sequencing using the ONT GridION and the Illumina MiSeq

Derek Tshiabuila[1,4*], Jennifer Giandhari[1,4], Sureshnee Pillay[1,4], Upasana Ramphal[1,2,4], Yajna Ramphal[1,4], Arisha Maharaj[1,4], Ugochukwu Jacob Anyaneji[1,4], Yeshnee Naidoo[1,4], Houriiyah Tegally[1,4], Emmanuel James San[1,4], Eduan Wilkinson[1,4], Richard J. Lessells[1,4], Tulio de Oliveira[1,2,3,4]

1. KwaZulu-Natal Research Innovation and Sequencing Platform (KRISP), School of Laboratory Medicine & Medical Sciences, University of KwaZulu-Natal, Durban 4001, South Africa.
2. Centre for the AIDS Programme of Research in South Africa (CAPRISA), Durban, South Africa.
3. Department of Global Health, University of Washington, Seattle, WA, USA.
4. Centre for Epidemic Response and Innovation (CERI), Stellenbosch University, Stellenbosch, South Africa.

- Correspondence: derektshiabuila@gmail.com; Tel.: +27 685 232 796

This study received no external funding and was approved by the Biomedical Research Ethics Committee (BREC)-UKZN (BREC/00002764/2021, 04 October 2021)

Abstract

Background: Over 4 million SARS-CoV-2 genomes have been sequenced globally in the past 2 years. This has been crucial in elucidating transmission chains within communities, the development of new diagnostic methods, vaccines, and antivirals. Although several sequencing technologies have been employed, Illumina and Oxford Nanopore remain the two most commonly used platforms. The sequence quality between these two platforms warrants a comparison of the genomes produced by the two technologies. Here, we compared the SARS-CoV-2 consensus genomes obtained from the Oxford Nanopore Technology GridION and the Illumina MiSeq for 28 sequencing runs.

Results: Our results show that the MiSeq had a significantly higher number of consensus genomes classified by Nextclade as good and mediocre compared to the GridION. The MiSeq also had a significantly higher genome coverage and mutation counts than the GridION.

Conclusion: Due to the low genome coverage, high number of indels, and sensitivity to SARS-CoV-2 viral load noted with the GridION when compared to MiSeq, we can conclude that the MiSeq is more favourable for SARS-CoV-2 genomic surveillance, as successful genomic surveillance is dependent on high quality, near-whole consensus genomes.

Keywords: SARS-CoV-2; Illumina MiSeq; Oxford Nanopore Technology GridION; Nanopore sequencing, Next Generation Sequencing, Bioinformatics

Background

December 2019 saw a novel viral pneumonia emerge from a seafood market in Wuhan China later found to be a new type of Coronavirus, now known as Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) (1, 2). On 11 March 2020, after approximately 118 000 cases had been reported globally, the World Health Organization (WHO) declared SARS-CoV-2 a global pandemic (3, 4). SARS-CoV-2 is an ongoing pandemic that requires continuous surveillance with approximately 270 031 622 cases confirmed globally as of 14 December 2021 (3, 5).

Sequencing of SARS-CoV-2 allowed for the rapid identification of the virus and the development of diagnostic tests and other tools for a rapid response to the pandemic (6). Sequencing provides genotypic information about a patient's infection, which can be used to gain knowledge on the specific infecting strain, assist in identifying transmission within communities, and advance the development of new diagnostic methods, vaccines, and antivirals (7). Multiple next generation sequencing (NGS) technologies have been used for SARS-CoV-2 sequencing, including Sanger, Illumina, ION torrent, and Oxford Nanopore

Technology (8). However, Illumina sequencing remains the most commonly used technology (9). As of 05 November 2021, 4 892 742 SARS-CoV-2 consensus genomes had been deposited into the Global Initiative on Sharing all Influenza Data (GISAID) with over 65 % from Illumina and approximately 25 % from Oxford Nanopore Technology (ONT) (10).

A major challenge with whole-genome sequencing (WGS) is obtaining whole viral genomes from clinical samples promptly (11). Illumina SARS-CoV-2 sequencing is generally limited by long sequencing times and the high cost and labour associated with library preparation for high-throughput sequencing (12). Another limitation is their relatively short reads (2 x 300 bp), as genomes generally contain multiple repeated sequences, known as tandem repeats, that may be longer than the NGS reads and may result in gaps and misassemblies (13). Due to the large footprint of most sequencers, portability can be a challenge which is unfortunate as there is generally a large distance between sample collection sites and sequencing laboratories (14). Nanopore sequencing overcomes these challenges as they sequence in real-time and are long-read sequencing technologies that allow for portability and have a relatively low initial investment on sequencing equipment with the MinION costing $1000 (15). ONT sequencing is, however, limited by the high number of false negatives and low sensitivity (16).

Short-read sequencing technologies are useful for population-level genetic analysis and clinical variant discovery as they provide low-cost, high-accuracy data when done in large batches. Long-read sequencing approaches, however, are well suited for de novo genome assembly, sequencing of genomes with long repetitive regions, copy number alterations, and complex structural variations (17). Several studies have compared the sequencing of SARS-CoV-2 between Illumina and ONT platforms and have shown that despite the high error rates observed with ONT sequencing, highly-accurate SARS-CoV-2 consensus genomes can be achieved (18). ONT sequencing, however, failed to detect short indels identified by Illumina sequencing

(18). There has also been a lower raw-read accuracy with nanopore sequencing when compared to Illumina sequencing (18, 19).

A comparison of SARS-CoV-2 WGS genomic coverage and variant detection between Illumina and Nanopore sequencing is necessary as it allows us to determine whether SARS-CoV-2 genomes produced by Nanopore sequencing can be reliably used for genomic surveillance and the development of diagnostic measures. As SARS-CoV-2 lineages differ by geographic location, this study aimed to determine whether Nanopore sequencing is a viable alternative to Illumina sequencing for rapidly identifying SARS-CoV-2 variants found within African countries. We hypothesize that Nanopore sequencing will produce consensus genomes that are comparable to consensus genomes produced by Illumina sequencing at a faster rate. SARS-CoV-2 sequencing results, for multiple runs, from the Illumina MiSeq and the ONT GridION were compared and although Nanopore sequencing was able to produce complete SARS-CoV-2 genomes, the quality observed was not as good as those obtained with Illumina sequencing. The ONT GridION can sequence up to 5 flowcells with 96 samples in a single run and is cheaper than sequencing with the Illumina MiSeq. These advantages can allow for more clinical facilities to sequence SARS-CoV-2 allowing for a greater response to the COVID-19 pandemic.

Materials and Methods

Study Population

The study population consisted of positive COVID-19 male and female patients whose nasopharyngeal swabs were sent from routine PCR diagnostic services for genomic surveillance to the Kwazulu-Natal Research Innovation and Sequencing Platform (KRISP). A total of 2608 COVID-19 positive nasopharyngeal swabs were used for sequencing from 28 different runs split evenly between the GridION and MiSeq. Samples were randomized and were from South Africa, Angola, Malawi, Mozambique, and Zimbabwe.

Real-Time PCR Assays

Sample Ct scores were present in the metadata files accompanying samples brought in for sequencing. There were three RT-PCR assays used for these samples. Namely; Seegene-AllplexTM 2019-nCoV Assay, Roche-Cobas® SARS-CoV-2 Qualitative assay, and Thermofisher-TaqPath™ COVID 19 CE IVD RT PCR Kit.

Total Nucleic Acid Extraction

RNA was extracted using the NA/gDNA kit on the automated Chemagic 360 system (Perkin Elmer) as per the manufacturer's instructions. Briefly, samples were lysed using lysis buffer and proteinase K, followed by binding to silica magnetic beads. The beads were then washed to remove unbound samples, and the RNA was eluted. Extracted RNA was stored at -80 °C before use.

Tiling PCR

Complementary DNA synthesis was performed using SuperScript IV reverse transcriptase (Life Technologies) in combination with random hexamer primers. This was then followed by gene-specific multiplex PCR using the ARTIC protocol (20). Primers were designed on a primal scheme (http://primal.zebraproject.org/) to cover the SARS-CoV-2 whole genome. Primers generated were 400 base pair (bp) amplicons, with an overlap of 70 bp to cover the 30 kilobases (kb) SARS-CoV-2 genome. Purification of PCR products was performed using AmpureXP purification beads in a 1:1 ratio (Beckman Coulter, High Wycombe, UK) and quantification was performed using the Qubit double-strand DNA (dsDNA) High Sensitivity Assay Kit on a Qubit 4.0 instrument (Life Technologies).

Illumina MiSeq Library Preparation and Sequencing

Sequencing libraries were generated using the amplicons generated by tiling PCR as described above. Indexed paired-end libraries were prepared using the Nextera DNA Flex Library Prep

Kits (Illumina) as per the manufacturer's instructions. Briefly, amplicons were tagmented to allow for unfragmented DNA to be cleaved and tagged. Each sample was barcoded with a unique barcode using the Nextera CD Indexes (Illumina) to enable downstream pooling of all libraries. Libraries were purified and normalized to 4 nM prior to pooling. The pooled library was denatured using 0.2 N sodium acetate and then diluted to a final concentration of 8 pM. The library was spiked with 1 % PhiX Control v3 (adapter-ligated library used as a control), and the libraries were sequenced using a 500-cycle v2 MiSeq Reagent Kit on the Illumina MiSeq instrument (Illumina, San Diego, CA, USA). The full details of the amplification and sequencing have been previously published (21). Fastq files produced from Illumina MiSeq were assembled using Genome Detective (https://www.genomedetective.com/) and the coronavirus typing tool (22). Genome detective is a web-based application that is user-friendly and is used for the assembly of known viral genomes from NGS datasets (22). Fastq files are uploaded to the application and read quality is visualized using FastQC. Low-quality reads are then filtered and the adapters trimmed with Trimmomatic (23). DIAMOND, a protein-based alignment method, is used to identify candidate viral reads (24). The Swissprot UniRef90 protein database viral subset is used to improve speed and sensitivity (22). Short reads are sorted and placed into groups and metagenomic de novo assembly is performed on each group using SPAdes for single-ended reads or metaSPAdes for paired-end reads (25). Each group is then identified using the taxonomy ID of the lowest common ancestor of the hits identified by DIAMOND (24). Blastx and Blastn are used to search for candidate reference sequences against the NCBI RefSeq virus database. The results for all detected contigs are combined by the Advanced Genome Aligner and a score is calculated by Genome Detective at the amino acid and nucleotide level. The five best scoring references for each config are then used for the alignment (22).

ONT GridION Library Preparation and Sequencing

Amplicons generated using the tiling PCR were prepared for nanopore sequencing using the ONT Native Barcoding Expansion Kits as per the manufacturer's guidelines. Libraries were multiplexed on FLO-MIN106 flowcells and run on the GridION X5. Furthermore, a no-template control from the PCR amplification step was added to each plate before running. Sequencing performance was monitored, in real-time, using the MinKNOW software app. Sequencing was terminated after 21hrs and the resulting reads were base-called using Guppy (4.0.14) and aligned to the Wuhan-Hu-1 reference genome (MN908947.3) using minimap2 (2.17-r941). Primer sequences were trimmed from the termini of read alignments and sequencing depth was capped at a maximum of 400-fold coverage using the ARTIC tool align_trim. Variant candidates were identified using Nanopolish (26).

Sequence Analysis

Consensus genomes produced by both platforms were uploaded to Nextclade Online Tool v1.4.2 (2021-10-26) (https://clades.nextstrain.org/) for genome clade assignments, mutation calling, quality checks, and to determine the genome position on the SARS-CoV-2 phylogenetic tree. Nextclade is built on Nextalign and consists of three tools; Nextclade Web, Nextclade CLI, and Nextalign CLI, which all share the common C++ library of algorithms. Nextclade starts by performing a pairwise alignment of the query sequence to a reference sequence using Nextalign that uses a banded local alignment algorithm with affine gap-cost that are determined through seed matching. Alignment is only performed on sequences longer than 100 nucleotides by default, but this can be changed, as alignment of shorter sequences may be unreliable. Mutation calling is achieved by comparing the aligned nucleotide sequences, one at a time, with the reference nucleotide sequence. Depending on their nature, they are reported differently. The number of missing, and ambiguous bases are also reported. Nextclade places each query sequence on the reference phylogenetic tree by comparing the mutations on the query sequence with the mutations of every node and tip in the reference tree,

and finding the node which has the most similar set of mutations. Clade assignment is achieved by placing sequences on a phylogenetic tree annotated with clade definitions (27). A Maximum-likelihood (ML) tree was constructed using IQ-TREE and was visualized using FigTree v1.4.4 (https://github.com/rambaut/figtree/releases) (28). Data visualization and statistical analysis were performed using ggplot2 v3.3.1 package and R v.4.1.1.

Statistical Considerations

The non-parametric nature of the data influenced the use of a Wilcoxon test to compare the number of consensus genomes produced by the GridION and the MiSeq classified within each category of the online Nextclade sequence analysis tool. The Wilcoxon test was also used to compare the difference in genomic coverage, number, and type of mutations detected between the GridION and the MiSeq. Statistical correlations were performed between Ct score and genome coverage and Ct score and the number of reads for both platforms.

Ethics

The University of KwaZulu-Natal Biomedical Research Ethics Committee waived the requirement for informed consent and approved the study (protocol reference no. BREC/00001195/2020; project title: COVID-19 transmission and natural history in KwaZulu-Natal, South Africa: epidemiological investigation to guide prevention and clinical care). All methods were performed in accordance with the relevant guidelines and regulations. We also used de-identified remnant nasopharyngeal and oropharyngeal swab samples from patients testing positive for SARS-CoV-2 by RT–qPCR from public health laboratories in South Africa. Informed consent for study participation was not applicable for this study because de-identified (anonymous) remnant samples, which would have been otherwise discarded, were used.

Results

Comparison of sequencing performance

To compare sequencing performance and runtime between the MiSeq and the GridION, Run116 was sequenced on both platforms **(Table 1)**. A total of 93 samples were sequenced and 93 consensus genomes were produced after assembly using Genome Detective. The sequencing runtime for the MiSeq was 36 hrs, whilst the GridION had a runtime of 21 hrs. The MiSeq had an overall higher average coverage than the GridION, having coverages of 94.34 % and 72.96 %, respectively. There was also a higher number of consensus genomes that passed the QC used for GISAID submissions (>80 % genome coverage) from the MiSeq, 83 (89.2 %), than the GridION, 29 (27.9 %). The average coverage across the genome for the GridION **(Figure 1-A)** was less uniform than that of the MiSeq **(Figure 1-B).**

Table 1. Comparison of sequencing Run116 on both the MiSeq and the GridION

	Run116	
	MiSeq	**GridION**
Runtime (hrs.)	36	21
No. of samples sequenced	93	93
Data obtained	7246.3 MB	1999,4 MB
Consensus genomes	93	93
Average coverage (X)	94.34%	72.96%
Passing GISAID QC (>+80%)	83 (89.2%)	29 (27.9%)
Clusters passing QC	70%	-
Q30 score	73.1%	-
Pores on flowcell	-	1012 pores

The table above summarizes the sequencing of Run116 on both the MiSeq and the GridION. The sequencing runtime for the MiSeq was 36 hrs, whilst that of the GridION was 21hrs. The MiSeq had a Q30 score of 73.1% with 70% of the clusters passeing QC The flowcell used for the GridION had 1012 pores available for sequencing. Of the 93 samples sequenced by both platforms, 93 consensus genomes were produced by each. Consensus genomes from the MiSeq had an average coverage of 94.34 % with 89.2 % having a coverage of 80 % and over. Consensus genomes from the GridION had an average coverage of 72.96 % with 27.9 % having a coverage of 80 % and over.

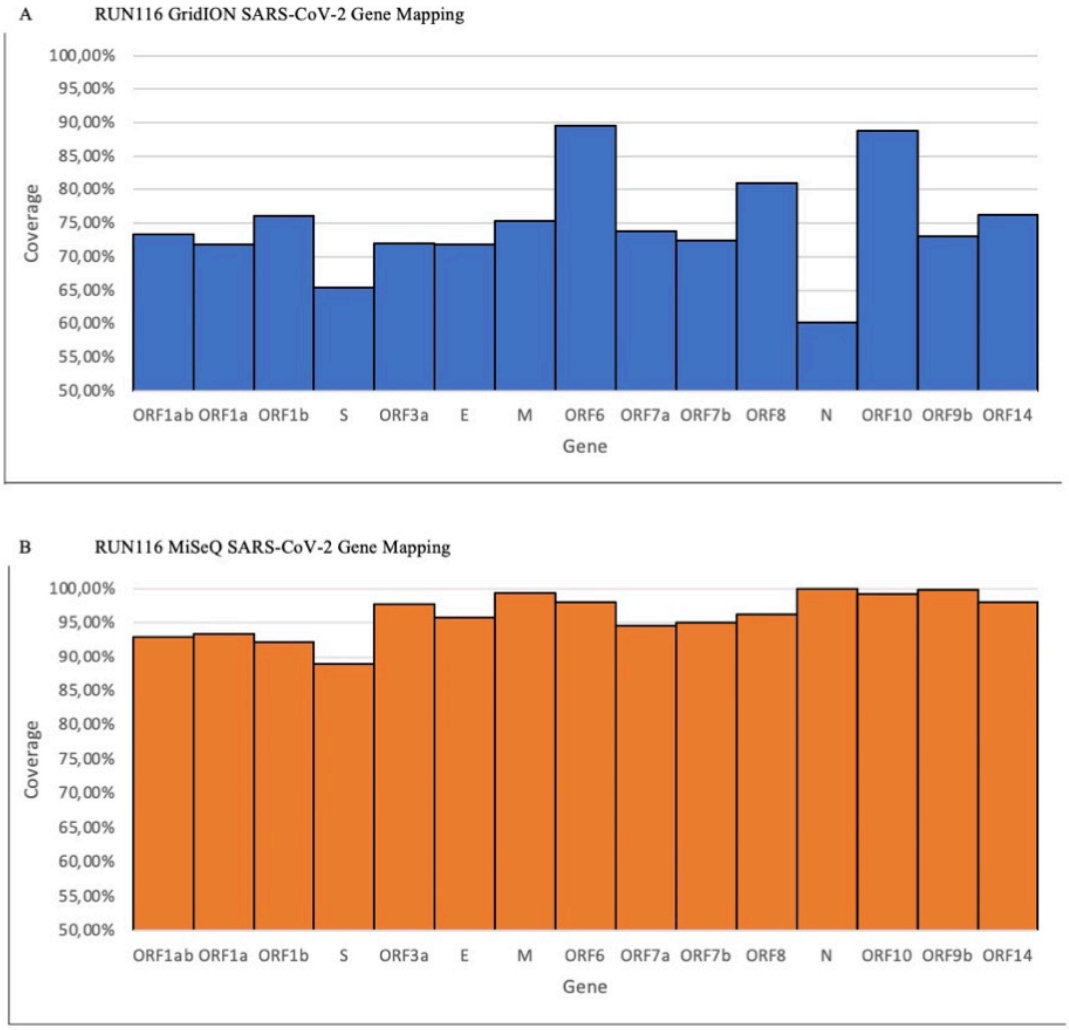


Figure 1. Comparison of GridION and MiSeq gene mapping for RUN116: Sequencing files from both the Illumina MiSeq and the ONT GridION were assembled using Genome Detective and average coverage across the 15 known genes was calculated to determine the sequencing coverage across the genome.

Comparison of consensus genome quality of Nanopore and Illumina sequencing

Consensus genomes produced by the GridION and the MiSeq were uploaded to Nextclade to determine the genome quality. Nextclade classifies genomes as either good, mediocre, or bad, based on the amount of missing data, and the number of mixed sites, private mutations, clustered mutations, frameshifts, and misplaced stop codons. Both the GridION and the MiSeq had a total of 14 runs with 1255 and 1183 consensus genomes, respectively. The total number of consensus genomes produced by the GridION and the MiSeq was significantly different (p

= 0.0053).. The number of genomes the two platforms classified as good (p = 0.00280), mediocre (p = 0.00250), and bad (p = 0.00037) also differed significantly **(Figure 2)**.

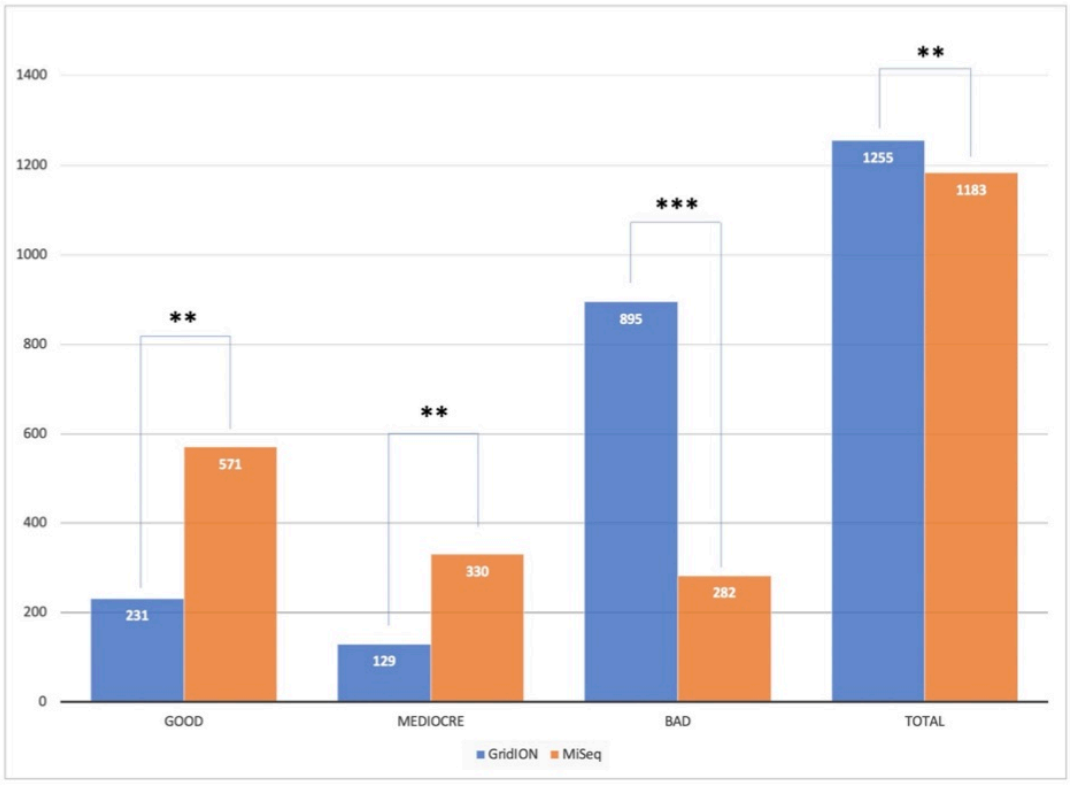


Figure 2. Comparison of consensus genome quality obtained from the GridION and the MiSeq and analyzed on Nextclade: To compare the quality of consensus genomes obtained from the GridION and the MiSeq, consensus genomes from both platforms were uploaded to Nextclade and the results plotted on a double bar graph. Genome quality was broken down into three groups; good, mediocre, and bad, with the GridION represented in blue and the MiSeq represented in orange. Statistical significance (Wilcoxon rank sum tests) is represented by "*" (**: $p < 0.01$, ***: $p < 0.001$). Sequencing scores ranging between 0 – 29 are classified as good, 30 – 99 are classified as mediocre, whilst 100 and above are classified as bad.

Comparison of genome coverage generated by the GridION and MiSeq

Identical samples (RUN116) were sequenced on both the GridION and the MiSeq and the genomic coverage was compared to determine the effect of sample quality on sequencing **(Figure 3-A)**. All the runs for both platforms were then compared **(Figure 3-B)**. A total of 86

consensus genomes were used from RUN116 after removing genomes with more than 100 mutations. Samples run on the MiSeq had a significantly greater genome coverage than the GridION ($p = 8.1e-16$). GridION genomes ranged from 35 – 100 %, whilst MiSeq genomes ranged from 80 – 100 %. The consensus genome coverage for all runs, 2351 genomes, was then compared. There was a significantly higher overall genome coverage observed with the MiSeq than with the GridION ($p < 2.2e-16$).

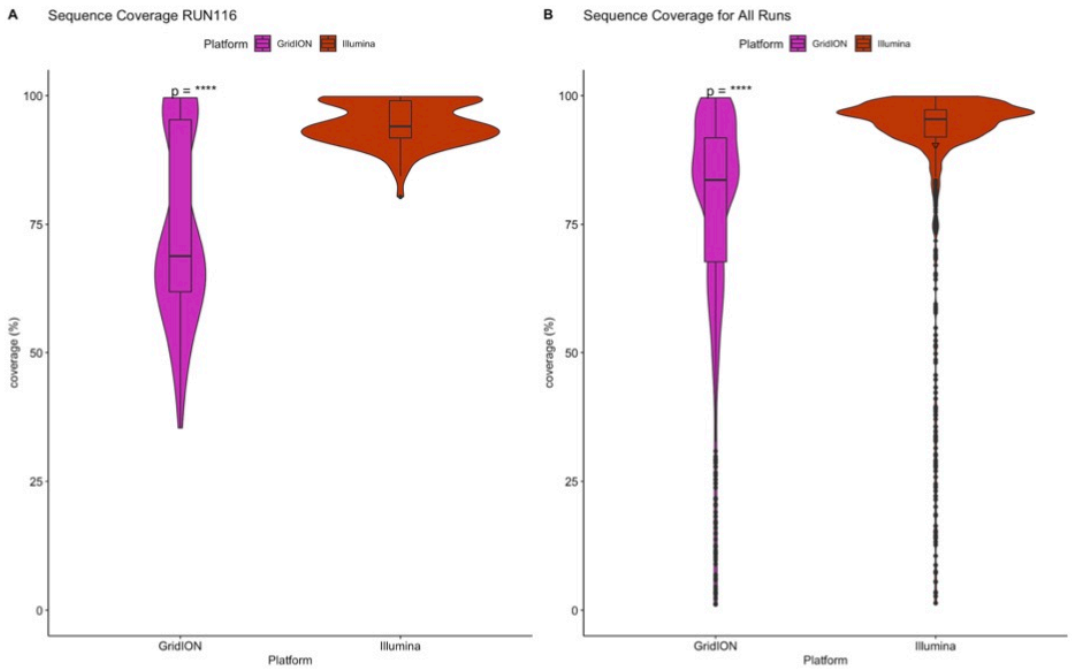


Figure 3. Comparison of GridION and MiSeq genome coverage: Fastq files for RUN116 from both the MiSeq and the GridION were assembled using Genome Detective and the consensus genome coverage was compared (A). The same was done for all genomes for both platforms (B). GridION samples are presented in purple, whilst Illumina MiSeq samples are presented in red. Statistical significance (Wilcoxon rank sum tests) is represented by "*" (****: $p < 0.0001$).

Comparison of Orf1ab- and S-gene coverage for GridION and MiSeq sequencing

To compare the depth of coverage of the ORF1ab- and S-gene for the GridION and the MiSeq, fastq files produced from both platforms were assembled on Genome Detective to produce consensus genomes. The results for each consensus genome were obtained and the coverages

for the ORF1ab-gene **(Figure 4-A)** and S-gene **(Figure 4-B)** were compared. All 14 runs for each platform were compared and Wilcoxon rank sum tests were performed. The ORF1ab-gene coverage ranged from 35 – 100 % for the GridION and 80 – 100 % for the MiSeq. The S-gene coverage ranged from 25 – 100 % for the GridION and 80 – 100 % for the MiSeq. There was a statistically significant difference in coverage for both genes on the GridION and the MiSeq with p = 1.2e-15 (RUN116) and p = 1.7e-15 (all genomes).

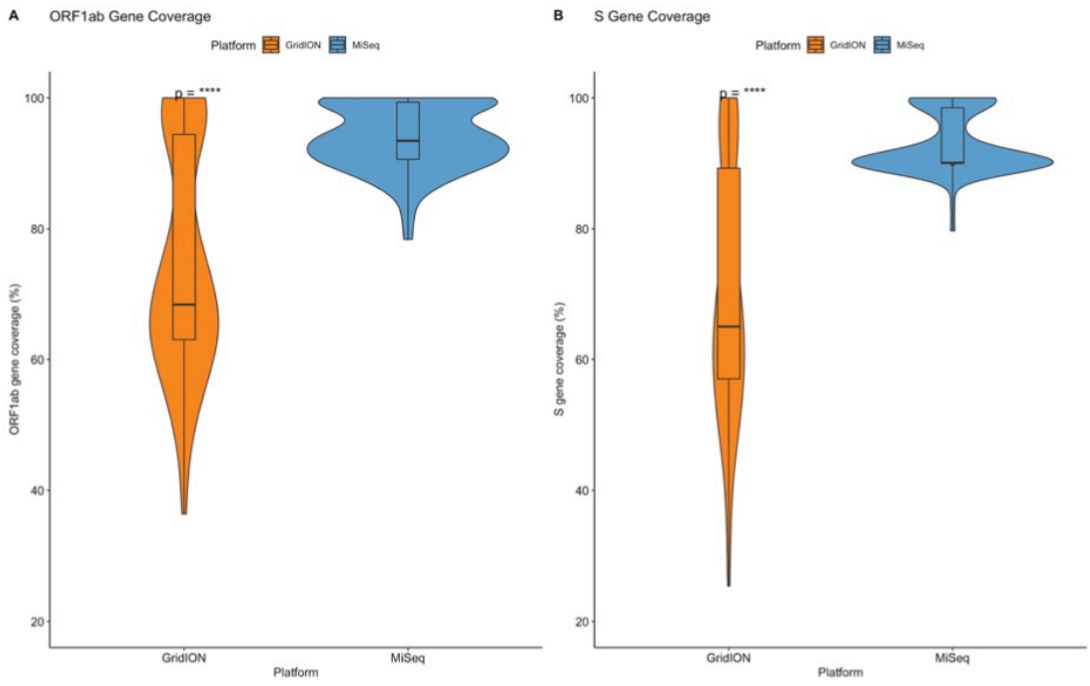


Figure 4. Comparison of ORF1ab- and S-gene coverage on the GridION and the MiSeq: Fastq files produced by both platforms were assembled on Genome Detective and the coverage for the ORF1ab- (A) and S-gene (B) was compared. Consensus genomes from the GridION are represented in orange and genomes from the MiSeq are represented in blue. Statistical significance (Wilcoxon rank sum tests) is represented by "*" (****: $p < 0.0001$).

Effect of Ct score on sequencing using the GridION and MiSeq

A correlation was performed to determine the effect of Ct score on genome coverage **(Figure 5)** and the number of reads produced by the GridION and the MiSeq during sequencing **(Figure 6)**. Due to the availability of Ct scores, three runs were used for each platform. Run101 (35 samples), Run111 (91 samples), and Run123 (64 samples), represented by graphs A, B, and C,

respectively, were used for the GridION. Run100 (68 samples), Run109 (54 samples), and Run122 (88 samples), represented by graphs D, E, and F, respectively, were used for the MiSeq. A negative correlation was observed between Ct Score and genome coverage for all six runs. The GridION's Runs 101, 111, and 123 had correlation coefficients of R = -0.88 (p = 4.5e-12), R = -0.45 (p = 7.2e-06), and R = -0.31(p = 0.012), respectively. The MiSeq's Runs 100, 109, and 122 had correlation coefficients of R = -0.35 (p = 0.0039), R = -0.19 (p = 0.18), and R = -0.33 (p = 0.0017), respectively. We note a significantly strong negative correlation between Ct score and number of reads for all GridION runs, whereas a significantly negative correlation was only noted for Run122 sequenced on the MiSeq. Run100 and Run109 showed non-significant correlations.

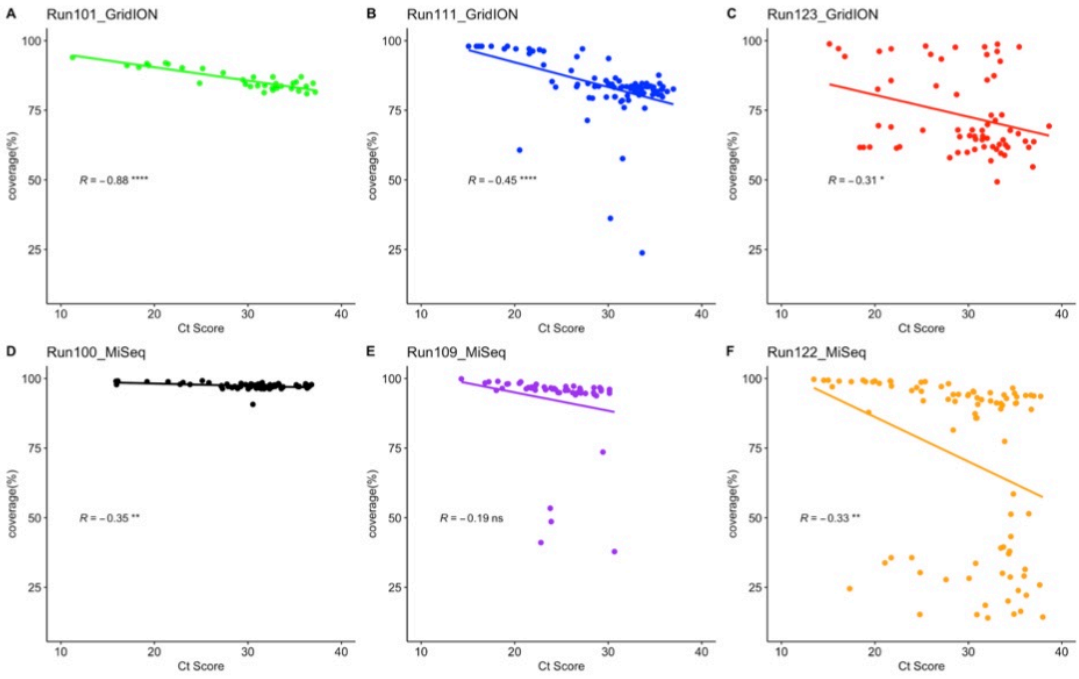


Figure 5. Correlation between genome coverage and Ct score for samples sequenced on the GridION and MiSeq: A correlation was performed to determine the effect of Ct score on the consensus genome coverage obtained from the GridION and the MiSeq. Genome coverage was plotted on the y-axis, whilst the sample's average Ct score was plotted on the X-axis. GridION runs are represented by graphs A (Run101), B (Run111), and C (Run123), which are represented as green, blue, and red, respectively. MiSeq runs are represented by graphs

D (Run100), E (Run109), and F (Run122) and are represented as black, purple, and gold, respectively. Statistical significance (Spearman's rank correlation test) is represented by "*" (ns: non-significant, *: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$, ****: $p < 0.0001$). For both platforms, as the Ct score increased, there was a decrease in genomic coverage.

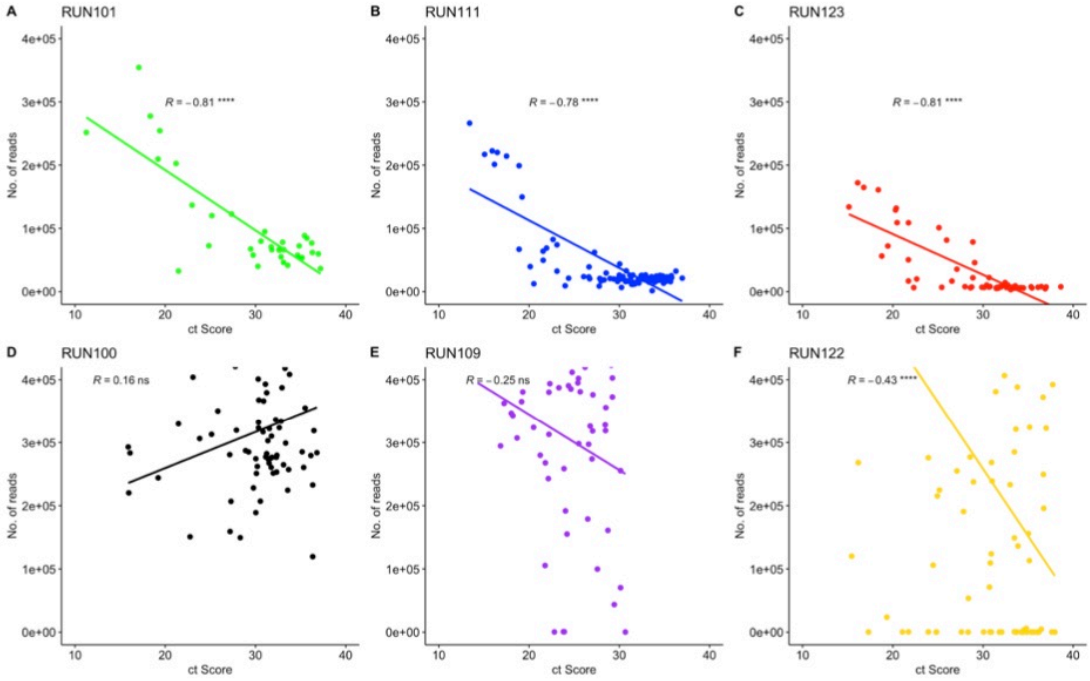


Figure 6. Correlation between the number of reads produced during sequencing and sample Ct Score: A correlation was performed for the number of reads produced by the GridION and the MiSeq and Ct score for SARS-CoV-2 samples. The number of reads was plotted on the Y-axis, whilst each sample's average Ct score was plotted on the X-axis. GridION runs are represented by graphs A (Run101), B (Run111), and C (Run123)and are shown as green, blue, and red, respectively. MiSeq runs are represented by graphs D (Run100), E (Run109), and F (Run122)and are shown as black, purple, and gold, respectively. Statistical significance (Spearman's rank correlation test) is represented by "*" (ns: non-significant, ****: $p < 0.0001$). An increase in Ct score resulted in a decrease in the number of reads produced for all GridION runs and 1 Illumina MiSeq run (Run122).

Mutation analysis

To determine whether the number of mutations detected by GridION and MiSeq differed significantly, the number of mutations detected for each sample was compared for Run116 **(Figure 7-A)** and all the runs **(Figure 7-B)**. The total number of insertions, deletions, and

substitutions detected by both platforms were also compared for Run116 **(Figure 7-C)** and all the runs **(Figure 7-D)**. A total of 181 consensus genomes obtained from the GridION and the MiSeq for Run116 were analyzed and a significant difference was noted in the number of mutations detected by each platform (Wilcoxon, p = 3.7e-08) with a greater number of mutations detected by the MiSeq (8 – 96 mutations) than the GridION (6 – 56 mutations). We also noted a significant difference (Wilcoxon, p = 1.5e-09) between the number of mutations detected from the genomes obtained from the MiSeq (1183 genomes) and the GridION (1255 genomes). There was a significant difference in the number of insertions (Wilcoxon, p = 8.2e-04) and substitutions (Wilcoxon, p = 5.3e-06) detected by both platforms for RUN116. However, when all runs were analyzed; only the number of insertions were significantly different between the two platforms (Wilcoxon, p = 7.5e-15).

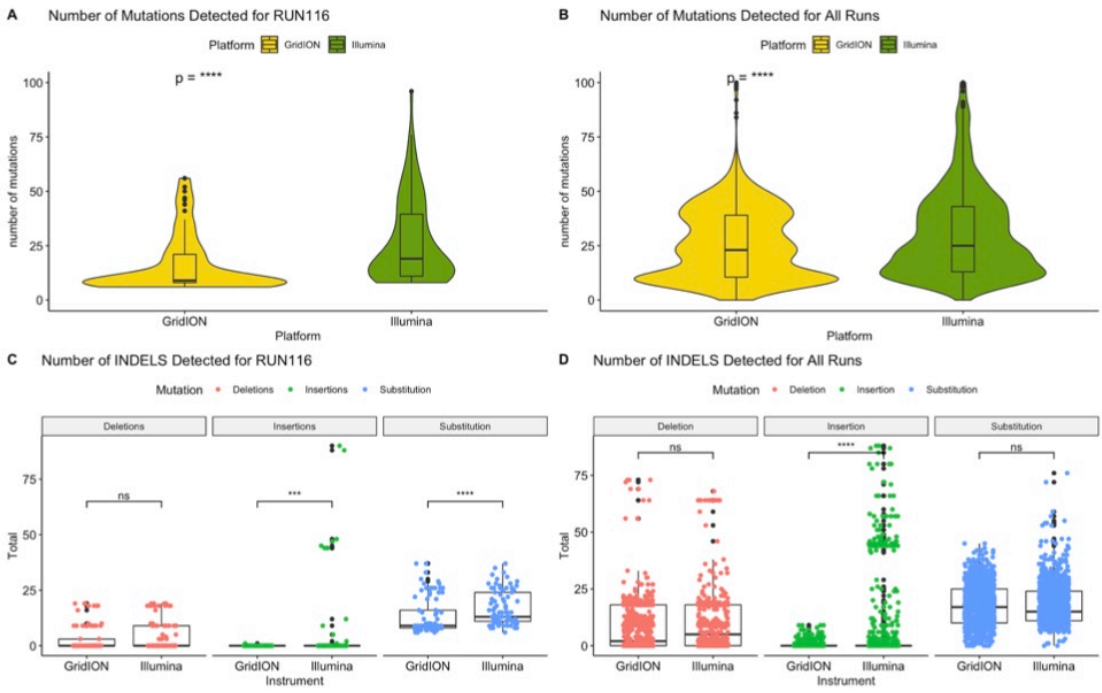


Figure 7. Analysis of mutations in samples sequenced on the GridION and the MiSeq: Consensus genomes produced by Genome Detective were uploaded to Nextclade and the results were analyzed. RUN116 was run on both platforms and the number and type of mutations detected by each platform was compared using a Wilcoxon

rank sum test **(Figure 7-A and -C)**. A consensus file for all runs, for each platform, was produced and uploaded to Nextclade and a Wilcoxon rank sum test was performed to compare the number and type of mutations detected by both platforms **(Figure 7-B and -D)**. GridION samples are represented in yellow, whilst MiSeq samples are presented in green. Deletions, insertions, and substitutions are represented in pink, green, and blue, respectively. Statistical significance (Wilcoxon p tests) is represented by "*" (ns: non-significant, ***: $p < 0.001$, ****: $p < 0.0001$).

Phylogenetic analysis

To determine whether there was a difference in the phylogenetic inference between consensus genomes generated by the GridION and the MiSeq, Run116 samples were sequenced on both platforms. A total of 93 consensus genomes from both the GridION and the MiSeq were uploaded to Nextclade and the results were compared. Of the 93 samples, 27 samples were classified within different clades **(Table 2).** A phylogenetic tree of the 27 samples was then created using IQTREE and visualized using FigTree **(Figure 8)**. Of the 27 samples, only one sample, highlighted in blue, was grouped on the same branch.

Table 2. Comparison of the genome coverage and assigned clade for run116 samples on Nextclade

Sample	Coverage (%)		Clade	
	GridION	MiSeq	GridION	MiSeq
K013400	72	92	20C	20A
K013408	57	86	20H (Beta, V2)	20C
K013410	70	90	20C	20A
K013411	76	93	20H (Beta, V2)	20A
K013415	63	91	20C	20A
K013417	63	94	20C	20A
K013418	62	94	20C	20A
K013423	63	91	20C	20A
K013425	60	93	20C	20A
K013426	57	89	20C	20A
K013429	64	95	20H (Beta, V2)	20C
K013432	65	93	20C	20A
K013433	50	94	20C	20A
K013434	68	94	20C	20A
K013437	35	92	20H (Beta, V2)	20C
K013445	65	91	20C	20A
K013447	92	98	20A	20D
K013449	49	94	20C	20A
K013450	68	97	20C	20A
K013452	68	94	20C	20A
K013454	56	90	20C	20A
K013462	51	92	20C	20A
K013465	60	94	20C	20A
K013467	50	92	20C	20A
K013470	72	89	20C	20H (Beta, V2)
K013476	69	91	20C	20H (Beta, V2)
Total	20A		1	20
	20C		22	3
	20D		0	1
	20H (Beta, V2)		4	3

The table above highlights the 27 samples which were sequenced on both the MiSeq and the GridION but were classified in different clades by Nexclade. Clades identified by the GridION include 20A (n = 1), 20C (n = 22), and 20H (Beta, V2) (n = 4). Clades identified by the MiSeq include 20A (n = 20), 20C (n = 3), 20D (n = 1), and

20H (Beta, V2) (n = 3). There was also an overall higher genomic coverage for sequences from the MiSeq when compared to the GridION.

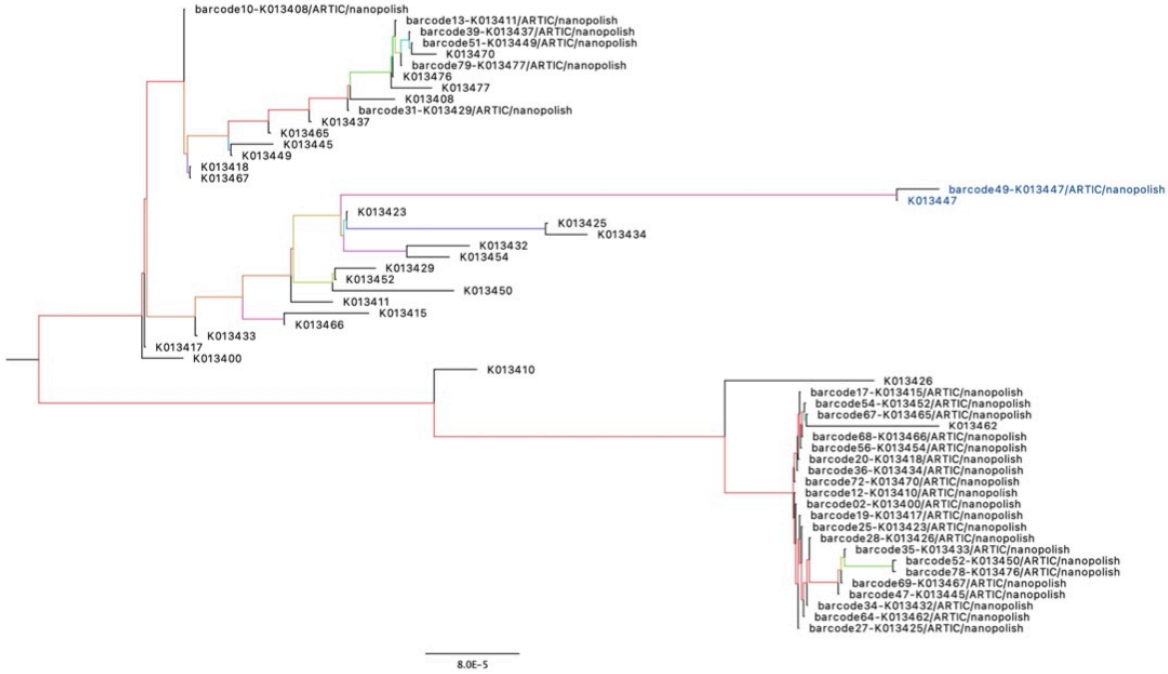


Figure 8. Phylogenetic comparison between identical samples sequenced using both the GridION and MiSeq: A phylogenetic tree was created using IQTREE and visualized using FigTree for samples from Run116 sequenced on both the GridION and the MiSeq but classified in different clades by Nextclade. Only one of the 27 samples, represented in blue, clustered on the same branch. GridION genomes are annotated as 'barcode*', whilst MiSeq genomes are annotated as 'K0*'.

Discussion

SARS-CoV-2 has caused a global health crisis as it is highly infectious and risks mutations that could result in more lethal variants (1, 29). A major factor in helping curb the spread of the virus and decreasing the infection rate is rapidly sequencing the virus to detect new strains and identify transmission chains (7). The sequencing runtime on the MiSeq for Run116 was 36 hours, whilst on the GridION it was 21 hours. This 10-hour decrease in sequencing time allows for 480 samples to be sequenced each day on the GridION in comparison to the 96 that can be

sequenced on the MiSeq every 36 hours. This is in agreement with reports that nanopore sequencing takes approximately 20 hours as a rapid library prep kit supplied by ONT can be used (30, 31). The lack of an image analysis step during nanopore sequencing facilitates real-time base-calling, which allows for the rapid detection of DNA for pathogen screening from clinical samples (32).

Studies have shown that Illumina sequencing may still be the most accurate way to sequence viruses (33). The majority of errors noted between Nanopore and Illumina consensus genomes have been attributed to Nanopore sequencing errors (34). Run116 samples were sequenced on both platforms to determine whether there was a significant difference in the sequencing coverage regardless of the sample. Consensus genome coverage was significantly greater with the MiSeq when compared to the GridION and this result was also observed when comparing all sequence runs. Genomic coverage can be affected by sequencing time and thus GridION coverage may have increased if left to sequence for longer. We also note a statistically significant higher sequencing coverage for the S-gene and ORF1ab-gene with the MiSeq than with the GridION. Nanopore technology has been shown to provide lower per-read sequencing coverage when compared to short-read sequencing (35). Coverage biases seen with ONT's sequencing protocol can be a result of truncated reads caused by pore blocking or fragmentation during library prep as transcripts are sequenced from the 3' to 5' end (36). ONT has made error correction tools such as Nanopolish available to try and reduce the error rate observed with Nanopore sequencing (37). In this study, variant calling was achieved using Nanopolish but we still note a significantly lower genome quality obtained from the GridION than the MiSeq. These low-quality genomes cannot be used to confidently acquire information on the infecting viral strain and are generally removed through a series of quality control checks (38). Although more consensus genomes can be produced using the GridION than the MiSeq, the low-quality genomes which are removed would eliminate the advantage of having a large number of

consensus genomes produced. It should be noted that the quality and coverage of consensus genomes for the ONT GridION can be increased by pooling lower samples as the number of reads and data produced will be shared across a smaller group.

Although Bull et al. 2020 shows that Nanopore sequencing was able to produce consensus genomes that were high quality, the SARS-CoV-2 viral variants that were available for analysis may not have been as diverse as the variants analysed in this study. This may have been due to the number of samples that were used for the study and the diversity of the samples as was as 157 samples were used in the study all of which came from Wales and Metropolitan Sydney. Furthermore, Samples were collected between March and April 2020 which may suggest that the viral variants in circulation were not as diverse as analysing samples from different African regions within a one year time frame as seen in this study.

Higher genomic coverage for the Illumina MiSeq has been associated with lower Ct scores (21). Ct score is a value that refers to the number of cycles required to amplify viral RNA to a detectable level. There is therefore an inverse relationship between Ct score and viral load (39). In this investigation, we also noted an inverse relationship between Ct score and genome coverage for both GridION and MiSeq sequencing. There is, however, a significantly stronger negative correlation seen with the GridION than the MiSeq, which may imply that the MiSeq's sequencing capabilities are less affected by sample Ct score and as a result, can be used for sequencing of samples within the early stages of infection when viral load is still low. This was, however, limited by not having the same runs to compare between the GridION and the MiSeq. Further analysis is required as the number of samples analyzed for each run was low and inconsistent due to the availability of Ct scores received with sample metadata. Additional analyses should be conducted to understand characteristics such as coverage bias, sequence biases, and reproducibility for the GridION sequencing platform (35). Sample quality may also

have an effect on sequencing and thus it is very important to maintain a cold chain during storage of swabs and RNA.

Identifying mutations involves aligning a consensus genome to a reference genome and identifying changes within the consensus genome. This is important, as it allows us to identify gene variants that may play a major role in the diagnosis of diseases (40). It has been shown that long-read sequencing platforms have a high error rate, which is mostly indels that are assumed to be randomly distributed within each read (41, 42). Prediction and interpretation of protein sequences may, therefore, be critically affected due to frameshifts and premature stop codons that may be introduced by the indels (43).

There was a significantly greater number of mutations detected by the MiSeq than the GridION for identical samples sequenced on both platforms. Although Nanopore platforms have been shown to make a large number of indel errors, in this study the MiSeq had a significantly higher number of insertions than the GridION. Paired-end sequencing, utilized by Illumina MiSeq, produces twice the number of reads, for the same sample and library preparation efforts, as single-end sequencing. This allows for a more accurate read alignment and detection of indel variants (44). Short read lengths have been shown to hinder the assignment of reads to parts of the genome that are complex, phasing of variants, resolving regions that are repeated, and the introduction of gaps and ambiguous regions in de novo assemblies. Longer reads can be used for sequencing of extended repetitive regions, allowing for the identification of mutations that are generally associated with disease (45). The higher number of indels noted with GridION sequencing highlights that genomic surveillance using Nanopore sequencing should be conducted cautiously as incorrect information on a viral strain can be obtained.

The rapid increase in COVID-19 cases has been linked to different SARS-CoV-2 viral lineages (46). Viral lineages are separated based on the number and type of mutations they contain that

differ from the parent strain (47). From the 93 consensus genomes analyzed from both platforms, 27 genomes were classified within different clades. These genomes had unique mutations and the clade differences noted between the two platforms were 20A – 20C and 20C – 20H(Beta, V2). As the number of indels and substitutions produced by the MiSeq and the GridION were significantly different, we can expect there to be differences in clade classifications as viral clades are subject to viral-defining mutations (29). Table 2 shows that genomes from the GridION have lower coverages than genomes from the MiSeq. This may be one of the factors causing a difference in the clade assignment as errors arising from the amplification and sequencing process may result in incomplete genome coverage, which affects phylogenetic inference (48). Rambaut et al., 2020 suggests that new lineages should only be proposed if the genome coverage exceeds 70 % of the coding region. Degradation of RNA can result in the introduction of mutations, which may cause a variant change (49). The GridION library for RUN116 was prepared simultaneously with that of the MiSeq and the amount of RNA used is also lower. Therefore, we can eliminate the possibility of RNA degradation and RNA input amount as factors that may have caused a difference in the variants called by each instrument. Lineages identified by the GridION need to be further analyzed to determine whether the mutations are valid or are a result of sequencing errors. Accurate identification of lineages can assist in identifying transmission chains and allow for the development of diagnostic methods and treatments (46).

Conclusions

The results of this study show that the ONT GridION is less ideal for SARS-CoV-2 genomic surveillance than the Illumina MiSeq but can be used to produce consensus genomes from samples of high quality and low CT scores. Healthcare facilities can, however, use ONT sequencing platforms to rapidly diagnose patients as the GridION can sequence up to 480

samples every 21 hours. This may allow for the identification and isolation of isolate infected individuals, thus aiding in stopping the spread of the disease.

List of Abbreviations

1. SARS-CoV-2
2. WHO
3. NGS
4. GISAID
5. ONT
6. WGS
7. COVID-19
8. KRISP
9. RNA
10. DNA
11. PCR
12. dsDNA
13. ML
14. Ct
15. QC
16. ns

Declarations

Ethics approval and consent to participate

This study was approved by the Biomedical Research Ethics Committee (BREC)-UKZN (BREC/00002764/2021, 04 October 2021). Consent to participate was waivered as the study made use of de-identified remnant nasopharyngeal and oropharyngeal swab samples from patients testing positive for SARS-CoV-2.

References:

1. World Health O. Clinical management of severe acute respiratory infection when novel coronavirus (2019-nCoV) infection is suspected: interim guidance, 28 January 2020. Geneva: World Health Organization; 2020 2020. Contract No.: WHO/nCoV/Clinical/2020.3.

2. Zhu N, Zhang D, Wang W, Li X, Yang B, Song J, et al. A Novel Coronavirus from Patients with Pneumonia in China, 2019. N Engl J Med. 2020;382(8):727-33.

3. Esbin MN, Whitney ON, Chong S, Maurer A, Darzacq X, Tjian R. Overcoming the bottleneck to widespread testing: a rapid review of nucleic acid testing approaches for COVID-19 detection. RNA. 2020;26(7):771-83.

4. World Health O. WHO Director-General's opening remarks at the media briefing on COVID-19 - 11 March 2020 2020 [Available from: https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020.

5. Amiel-Tison C. [Some aspects of prevention in perinatology (author's transl)]. J Gynecol Obstet Biol Reprod (Paris). 1978;7(3 Pt 2):596-604.

6. Seth-Smith HMB, Bonfiglio F, Cuenod A, Reist J, Egli A, Wuthrich D. Evaluation of Rapid Library Preparation Protocols for Whole Genome Sequencing Based Outbreak Investigation. Front Public Health. 2019;7:241.

7. St Hilaire BG, Durand NC, Mitra N, Pulido SG, Mahajan R, Blackburn A, et al. A rapid, low cost, and highly sensitive SARS-CoV-2 diagnostic based on whole genome sequencing. bioRxiv. 2020:2020.04.25.061499.

8. Liu L, Li Y, Li S, Hu N, He Y, Pong R, et al. Comparison of next-generation sequencing systems. J Biomed Biotechnol. 2012;2012:251364.

9. GISAID. Pandemic coronavirus causing COVID-19 2022 [Available from: https://www.gisaid.org/.

10. Chantal Babb de Villiers LB, Sarah Cook, Joanna Janus, Emma Johnson, Mark Kroese. Next generation sequencing for SARS-CoV-2. PHG Foundation; 2021.

11. Resende PC, Motta FC, Roy S, Appolinario L, Fabri A, Xavier J, et al. SARS-CoV-2 genomes recovered by long amplicon tiling multiplex approach using nanopore sequencing and applicable to other sequencing platforms. bioRxiv. 2020:2020.04.30.069039.

12. Gohl DM, Garbe J, Grady P, Daniel J, Watson RHB, Auch B, et al. A rapid, cost-effective tailed amplicon method for sequencing SARS-CoV-2. BMC Genomics. 2020;21(1):863.

13. van Dijk EL, Jaszczyszyn Y, Naquin D, Thermes C. The Third Revolution in Sequencing Technology. Trends Genet. 2018;34(9):666-81.

14. Xu Y, Lewandowski K, Jeffery K, Downs LO, Foster D, Sanderson ND, et al. Nanopore metagenomic sequencing to investigate nosocomial transmission of human metapneumovirus from a unique genetic group among haematology patients in the United Kingdom. J Infect. 2020;80(5):571-7.

15. Schadt EE, Turner S, Kasarskis A. A window into third-generation sequencing. Hum Mol Genet. 2010;19(R2):R227-40.

16. Wang M, Fu A, Hu B, Tong Y, Liu R, Gu J, et al. Nanopore target sequencing for accurate and comprehensive detection of SARS-CoV-2 and other respiratory viruses. medRxiv. 2020:2020.03.04.20029538.

17. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet. 2016;17(6):333-51.

18. Bull RA, Adikari TN, Ferguson JM, Hammond JM, Stevanovski I, Beukers AG, et al. Analytical validity of nanopore sequencing for rapid SARS-CoV-2 genome analysis. Nat Commun. 2020;11(1):6272.

19. Jayamohan H, Lambert CJ, Sant HJ, Jafek A, Patel D, Feng H, et al. SARS-CoV-2 pandemic: a review of molecular diagnostic tools including sample collection and commercial response with associated advantages and limitations. Anal Bioanal Chem. 2021;413(1):49-71.

20. Quick J. nCoV-2019 sequencing protocol. Protocols io[Google Scholar]. 2020.

21. Pillay S, Giandhari J, Tegally H, Wilkinson E, Chimukangara B, Lessells R, et al. Whole Genome Sequencing of SARS-CoV-2: Adapting Illumina Protocols for Quick and Accurate Outbreak Investigation during a Pandemic. Genes (Basel). 2020;11(8).

22. 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(11):3552-5.

23. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30(15):2114-20.

24. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nature methods. 2015;12(1):59-60.

25. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. Journal of computational biology. 2012;19(5):455-77.

26. Nick Loman WR, Andrew Rambaut. nCoV-2019 novel coronavirus bioinformatics protocol 2020-01-23 [Available from: https://artic.network/ncov-2019/ncov2019-bioinformatics-sop.html.

27. Aksamentov I, Roemer C, Hodcroft E, Neher R. Nextclade: clade assignment, mutation calling and quality control for viral genomes. Journal of Open Source Software. 2021;6:3773.

28. Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol. 2015;32(1):268-74.

29. Tegally H, Wilkinson E, Lessells RJ, Giandhari J, Pillay S, Msomi N, et al. Sixteen novel lineages of SARS-CoV-2 in South Africa. Nat Med. 2021;27(3):440-6.

30. James P, Stoddart D, Harrington ED, Beaulaurier J, Ly L, Reid SW, et al. LamPORE: rapid, accurate and highly scalable molecular screening for SARS-CoV-2 infection, based on nanopore sequencing. medRxiv. 2020:2020.08.07.20161737.

574 31. Kono N, Arakawa K. Nanopore sequencing: Review of potential applications in
575 functional genomics. Dev Growth Differ. 2019;61(5):316-26.

576 32. Jain M, Koren S, Miga KH, Quick J, Rand AC, Sasani TA, et al. Nanopore
577 sequencing and assembly of a human genome with ultra-long reads. Nat Biotechnol.
578 2018;36(4):338-45.

579 33. Hourdel V, Kwasiborski A, Baliere C, Matheus S, Batejat CF, Manuguerra JC, et al.
580 Rapid Genomic Characterization of SARS-CoV-2 by Direct Amplicon-Based Sequencing
581 Through Comparison of MinION and Illumina iSeq100(TM) System. Front Microbiol.
582 2020;11:571328.

583 34. McNaughton AL, Roberts HE, Bonsall D, de Cesare M, Mokaya J, Lumley SF, et al.
584 Illumina and Nanopore methods for whole genome sequencing of hepatitis B virus (HBV).
585 Sci Rep. 2019;9(1):7081.

586 35. Amarasinghe SL, Su S, Dong X, Zappia L, Ritchie ME, Gouil Q. Opportunities and
587 challenges in long-read sequencing data analysis. Genome Biol. 2020;21(1):30.

588 36. Sharon D, Tilgner H, Grubert F, Snyder M. A single-molecule long-read survey of the
589 human transcriptome. Nat Biotechnol. 2013;31(11):1009-14.

590 37. Loman NJ, Quick J, Simpson JT. A complete bacterial genome assembled de novo
591 using only nanopore sequencing data. Nat Methods. 2015;12(8):733-5.

592 38. Gleizes A, Laubscher F, Guex N, Iseli C, Junier T, Cordey S, et al. Virosaurus A
593 Reference to Explore and Capture Virus Genetic Diversity. Viruses. 2020;12(11):1248.

594 39. Tom MR, Mina MJ. To Interpret the SARS-CoV-2 Test, Consider the Cycle
595 Threshold Value. Clin Infect Dis. 2020;71(16):2252-4.

596 40. Wei X, Ju X, Yi X, Zhu Q, Qu N, Liu T, et al. Identification of sequence variants in
597 genetic disease-causing genes using targeted next-generation sequencing. PLoS One.
598 2011;6(12):e29500.

599 41. Ardui S, Ameur A, Vermeesch JR, Hestand MS. Single molecule real-time (SMRT)
600 sequencing comes of age: applications and utilities for medical diagnostics. Nucleic Acids
601 Res. 2018;46(5):2159-68.

602 42. Weirather JL, de Cesare M, Wang Y, Piazza P, Sebastiano V, Wang X-J, et al.
603 Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and
604 their applications to transcriptome analysis. F1000Research. 2017;6.

605 43. Weston S, Frieman MB. COVID-19: Knowns, Unknowns, and Questions. mSphere.
606 2020;5(2).

607 44. illumina. Advantages of paired-end and single-read sequencing 2021 [updated 2021.
608 Available from: https://www.illumina.com/science/technology/next-generation-
609 sequencing/plan-experiments/paired-end-vs-single-read.html.

610 45. Rhoads A, Au KF. PacBio Sequencing and Its Applications. Genomics Proteomics
611 Bioinformatics. 2015;13(5):278-89.

612 46. Tegally H, Wilkinson E, Giovanetti M, Iranzadeh A, Fonseca V, Giandhari J, et al.
613 Detection of a SARS-CoV-2 variant of concern in South Africa. Nature.
614 2021;592(7854):438-43.

615 47. Morel B, Barbera P, Czech L, Bettisworth B, Hübner L, Lutteropp S, et al.
616 Phylogenetic analysis of SARS-CoV-2 data is difficult. bioRxiv. 2020:2020.08.05.239046.

617 48. Rambaut A, Holmes EC, O'Toole A, Hill V, McCrone JT, Ruis C, et al. A dynamic
618 nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. Nat
619 Microbiol. 2020;5(11):1403-7.

620 49. Abernathy E, Glaunsinger B. Emerging roles for RNA degradation in viral replication
621 and antiviral defense. Virology. 2015;479-480:600-8.

622

Bridging Chapters 2 and 3

Chapter 2 compares sequences produced by the ONT GridION to those produced by the Illumina MiSeq to determine whether nanopore sequencing can be used in SARS-CoV-2 genomic surveillance. Chapter 3 is a preprint published in medrxiv.org and it shows the mutations identified in the SARS-CoV-2 spike protein from genomic surveillance of travelers in Africa. In this study, I assisted in SARS-CoV-2 sequence assembly, whole-genome analysis, and data curation. This helped to identify how the SARS-CoV-2 virus was spreading across Africa, thus allowing healthcare facilities to isolate infected individuals and prevent the spread of the different variants. Consensus genomes analyzed in this study were from both the ONT GridION and the Illumina MiSeq.

CHAPTER 3

A NOVEL VARIANT OF INTEREST OF SARS-COV-2 WITH MULTIPLE SPIKE MUTATIONS DETECTED THROUGH TRAVEL SURVEILLANCE IN AFRICA

A novel variant of interest of SARS-CoV-2 with multiple spike mutations detected through travel surveillance in Africa.

Tulio de Oliveira[1]*, Silvia Lutucuta[2]*, John Nkengasong[3], Joana Morais[2], Joana Paula Paixão[2], Zoraima Neto[2], Pedro Afonso[2], Julio Miranda[2], Kumbelembe David[2], Luzia Inglês[2], Amilton Pereira Agostinho Paulo Raisa Rivas Carralero[2] , Helga Reis Freitas[2], Franco Mufinda[2], Sofonias Kifle Tessema[3], Houriiyah Tegally[1], Emmanuel James San[1], Eduan Wilkinson[1], Jennifer Giandhari[1], Sureshnee Pillay[1], Marta Giovanetti[4], Yeshnee Naidoo[1], Aris Katzourakis[5], Mahan Ghafari[5], Lavanya Singh[1], Derek Tshiabuila[1], Darren Martin[6], Richard J Lessells[1].

[1]KwaZulu-Natal Research Innovation and Sequencing Platform (KRISP), Nelson R Mandela School of Medicine, University of KwaZulu-Natal, Durban, South Africa.[2] Angola Ministry of Health, Luanda, Angola. [3] Africa Centres for Disease Control and Prevention (Africa CDC), Addis Ababa, Ethiopia. [4] Oswaldo Cruz Foundation (FioCruz), Rio de Janeiro, Brazil. [5]University of Oxford, Oxford, U.K. [6] University of Cape Town, Cape Town, South Africa.

Corresponding authors: Prof. Tulio de Oliveira (deoliveira@ukzn.ac.za), Minister Silvia Lutucuta.

Abstract:

At the end of 2020, the Network for Genomic Surveillance in South Africa (NGS-SA) detected a SARS-CoV-2 variant of concern (VOC) in South Africa (501Y.V2 or PANGO lineage B.1.351)1. 501Y.V2 is associated with increased transmissibility and resistance to neutralizing antibodies elicited by natural infection and vaccination2,3. 501Y.V2 has since spread to over 50 countries around the world and has contributed to a significant resurgence of the epidemic in southern Africa. In order to rapidly characterize the spread of this and other emerging VOCs and variants of interest (VOIs), NGS-SA partnered with the Africa Centres for Disease Control and Prevention and the African Society of Laboratory Medicine through the Africa Pathogen Genomics Initiative to strengthen SARS-CoV-2 genomic surveillance across the region.

Here, we report the first genomic surveillance results from Angola, which has had 21 500 reported cases and around 500 deaths from COVID-19 up to March 2021 (Supplemental Fig S1). On 15 January 2021, in response to the international spread of VOCs, the government instituted compulsory rapid antigen testing of all passengers arriving at the main international airport, in addition to the existing requirement to present a negative PCR test taken within 72 hours of travel. All individuals with a positive antigen test are isolated in a government facility for a minimum of 14 days and require two negative RT-PCR tests at least 48 hours apart for de-isolation, whilst all travelers with a negative test on arrival proceed to mandatory self-quarantine for 10 days followed by a repeat test.

In March 2021, we received 118 nasopharyngeal swab samples collected between June 2020 and February 2021, a number of which were from incoming air travelers (Supplemental Fig S1). From these, we produced 73 high quality genomes (>80% coverage), 14 of which were known VOCs/VOIs (seven 501Y.V2/B.1.351, six B.1.1.7, one B.1.525), 44 of which were C.16 (a common lineage circulating in Portugal), and twelve of which were other lineages (Supplemental Fig S2). In addition, we detected a new VOI in three incoming travelers from Tanzania who were tested together at the airport in mid-February. The three genomes from these passengers were almost identical and presented highly divergent sequences within the A lineage (Figure 1A & 1B). The GISAID database contains nine other sequences reported to be sampled from cases involving travel from Tanzania, two of which are basal to the three sampled in Angola (Figure 1A, Supplemental Table S1).

This new VOI, temporarily designated A.VOI.V2, has 31 amino acid substitutions (11 in spike) and three deletions (all in spike) (Figure 1C & 1D). The spike mutations include three substitutions in the receptor-binding domain (R346K, T478R and E484K); five substitutions and three deletions in the N-terminal domain, some of which are within the antigenic supersite (Y144Δ, R246M, SYL247-249Δ and W258L)4; and two substitutions adjacent to the S1/S2 cleavage site (H655Y and P681H). Several of these mutations are present in other VOCs/VOIs and are evolving under positive selection.

Main Body of the paper (may be some repetion from abstract, due to journal format).

At the end of 2020, the Network for Genomic Surveillance in South Africa (NGS-SA) detected a SARS-CoV-2 variant of concern (VOC) in South Africa (501Y.V2 or PANGO lineage B.1.351)[1]. 501Y.V2 is associated with increased transmissibility and resistance to neutralizing antibodies elicited by natural infection and vaccination[2,3]. 501Y.V2 has since spread to over 50 countries around the world and has contributed to a significant resurgence of the epidemic in southern Africa. In order to rapidly characterize the spread of this and other emerging VOCs and variants of interest (VOIs), NGS-SA partnered with the Africa Centres for Disease Control and Prevention and the African Society of Laboratory Medicine through the Africa Pathogen Genomics Initiative to strengthen SARS-CoV-2 genomic surveillance across the region.

Here, we report the first genomic surveillance results from Angola, which has had 21 500 reported cases and around 500 deaths from COVID-19 up to March 2021 (Supplemental Fig S1). On 15 January 2021, in response to the international spread of VOCs, the government instituted

compulsory rapid antigen testing of all passengers arriving at the main international airport, in addition to the existing requirement to present a negative PCR test taken within 72 hours of travel. All individuals with a positive antigen test are isolated in a government facility for a minimum of 14 days and require two negative RT-PCR tests at least 48 hours apart for de-isolation, whilst all travelers with a negative test on arrival proceed to mandatory self-quarantine for 10 days followed by a repeat test.

In March 2021, we received 118 nasopharyngeal swab samples collected between June 2020 and February 2021, a number of which were from incoming air travelers (Supplemental Fig S1). From these, we produced 73 high quality genomes (>80% coverage), 14 of which were known VOCs/VOIs (seven 501Y.V2/B.1.351, six B.1.1.7, one B.1.525), 44 of which were C.16 (a common lineage circulating in Portugal), and twelve of which were other lineages (Supplemental Fig S2). In addition, we detected a new VOI in three incoming travelers from Tanzania who were tested together at the airport in mid-February. The three genomes from these passengers were almost identical and presented highly divergent sequences within the A lineage (Figure 1A & 1B). The GISAID database contains nine other sequences reported to be sampled from cases involving travel from Tanzania, two of which are basal to the three sampled in Angola (Figure 1A, Supplemental Table S1).

This new VOI, temporarily designated A.VOI.V2, has 31 amino acid substitutions (11 in spike) and three deletions (all in spike) (Figure 1C & 1D). The spike mutations include three substitutions in the receptor-binding domain (R346K, T478R and E484K)**;** five substitutions and three deletions in the N-terminal domain, some of which are within the antigenic supersite (Y144Δ, R246M, SYL247-249Δ and W258L)[4]**;** and two substitutions adjacent to the S1/S2

cleavage site (H655Y and P681H). Several of these mutations are present in other VOCs/VOIs and are evolving under positive selection (Figure 1D and Supplemental Text, Fig S2)[5].

We decided to report this as a new VOI given the constellation of mutations with known or suspected biological significance, specifically resistance to neutralizing antibodies and potentially increased transmissibility (Supplemental Table S3). Whilst we have only detected three cases with this new VOI, this warrants urgent investigation as the source country has a largely undocumented epidemic and few public health measures in place to prevent spread within and out of the country.

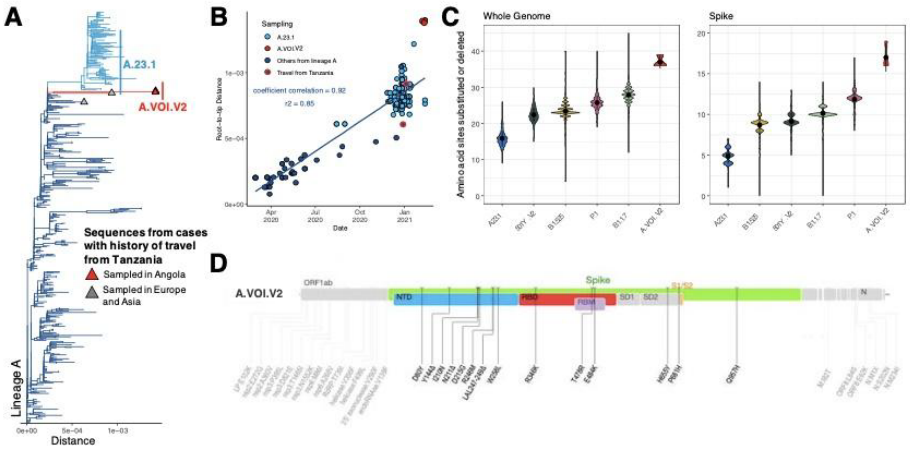


Figure 1: A) Phylogenetic tree of a subset of lineage A sequences (n=319) including five sequences from cases with history of travel of Tanzania, three of which are the A.VOI.V2 sampled in Angola (tips shown with a triangle); B) Regression of root-to-tip genetic distances against sampling dates, for sequences belonging to lineage A, showing the novel A.VOI.V2 (red), the known VOI A.23.1 (light blue), other sequences of lineage A (deep blue), two of which are documented to have travel history from Tanzania (red outline); C) Violin plot showing the number of amino acid mutations in the whole genome and spike glycoprotein in a subset of genomes from five known variants compared to the novel A.VOI.V2; D) Genome map showing the position of the 31 amino acid substitutions and three deletions (spike in color, NTD = N-

terminal domain, RBD = receptor-binding domain, RBM = receptor-binding motif, S1/S2 = S1/S2 cleavage site, and the rest of the genome in grey).

Authors:

Tulio de Oliveira*, Silvia Lutucuta*, John Nkengasong, Joana Morais, Joana Paula Paixão, Zoraima Neto, Pedro Afonso, Julio Miranda, Kumbelembe David, Luzia Inglês, Amilton Pereira Agostinho Paulo Raisa Rivas Carralero , Helga Reis Freitas, Franco Mufinda, Sofonias Kifle Tessema, Houriiyah Tegally, Emmanuel James San, Eduan Wilkinson, Jennifer Giandhari, Sureshnee Pillay, Marta Giovanetti, Yeshnee Naidoo, Aris Katzourakis, Mahan Ghafari, Lavanya Singh, Derek Tshiabuila, Darren Martin, Richard J Lessells.

*Corresponding authors

Angola Ministry of Health authors:
Silvia Lutucuta (Lutucuta, S)
Zoraima Neto (Neto, Z.)
Pedro Afonso (Afonso, P.)
Julio Miranda (Miranda, J.)
Kumbelembe David (David, K.)
Luzia Inglês (Inglês, L.)
Amilton Pereira (Pereira, A.)
Agostinho Paulo (Paulo, A.)
Raisa Rivas Carralero (Carralero, R. R.)
Joana Paula Paixão (Paixão. J. P.)
Helga Reis Freitas (Freitas R. H.)
Franco Mufinda (Mufinda M.)
Joana Morais (Morais, J.)

AFRICA CDC authors:
John N. Nkengasong (Nkengasong, J. N.)
Sofonias Kifle Tessema (Tessema, S.K.)

KRISP at UKZN authors:
Houriiyah Tegally (Tegally, H.)
Emmanuel James San (San, E.J.)
Eduan Wilkinson (Wilkinson, E.)
Jennifer Giandhari (Giandhari, J.)
Sureshnee Pillay (Pillay, S.)
Yeshnee Naidoo (Naidoo, Y.)
Lavanya Singh (Singh, L.)
Derek Tshiabuila (Tshiabuila, D.)
Richard J. Lessells (Lessells, R. K.)

Fiocruz author:
Marta Giovanetti (Giovanetti, M.)

University of Oxford authors:
Aris Katzourakis (Katzourakis, A.)
Mahan Ghafari (Ghafari, M.)

University of Cape Town author
Darren Martin (Martin, D.)

References

1. Tegally H, Wilkinson E, Giovanetti M, et al. Emergence of a SARS-CoV-2 variant of concern with mutations in spike glycoprotein. *Nature*. 2021 Mar 9
2. Cele S, Gazy I, Jackson L, et al. Escape of SARS-CoV-2 501Y.V2 variants from neutralization by convalescent plasma. *medRxiv*. 2021:2021.2001.2026.21250224v2
3. Madhi SA, Baillie V, Cutland CL, et al. Efficacy of the ChAdOx1 nCoV-19 Covid-19 Vaccine against the B.1.351 Variant. *N Engl J Med*. 2021 Mar 16
4. McCallum M, Marco A, Lempp F, et al. N-terminal domain antigenic mapping reveals a site of vulnerability for SARS-CoV-2. *Cell*. 2021: 184, 1-16
5. Garcia-Beltran WF, Lam EC, St. Denis K, et al. Multiple SARS-CoV-2 variants escape neutralization by vaccine-induced humoral immunity. *Cell*. 2021 Mar 12

Bridging Chapters 3 and 4

Chapter 3 shows how genomic surveillance helped identify the SARS-CoV-2 mutations that were producing VOIs and VOCs. As the pandemic continued through 2021, it was necessary to continue SARS-CoV-2 genomic surveillance to detect new variants and prevent their spread across communities. Chapter 4 highlights how continued genomic surveillance using Illumina and ONT sequencing platforms allowed transmission chains and new variants to be identified across Africa. The article was published in the journal *Science* and I assisted in SARS-CoV-2 sequence assembly, genome analysis, and data curation.

CHAPTER 4

A YEAR OF GENOMIC SURVEILLANCE REVEALS HOW THE SARS-COV-2 PANDEMIC UNFOLDED IN AFRICA

RESEARCH ARTICLE

CORONAVIRUS

A year of genomic surveillance reveals how the SARS-CoV-2 pandemic unfolded in Africa

Eduan Wilkinson[1,2]†, Marta Giovanetti[3,4]†, Houriiyah Tegally[1]†, James E. San[1]†, Richard Lessells[1], Diego Cuadros[5], Darren P. Martin[6,7], David A. Rasmussen[8,9], Abdel-Rahman N. Zekri[10], Abdoul K. Sangare[11], Abdoul-Salam Ouedraogo[12], Abdul K. Sesay[13], Abechi Priscilla[14], Adedotun-Sulaiman Kemi[14], Adewunmi M. Olubusuyi[15], Adeyemi O. O. Oluwapelumi[16], Adnène Hammami[17], Adrienne A. Amuri[18,19], Ahmad Sayed[20], Ahmed E. O. Ouma[21], Aida Elargoubi[22,23], Nnennaya A. Ajayi[24], Ajogbasile F. Victoria[14], Akano Kazeem[14], Akpede George[25], Alexander J. Trotter[26], Ali A. Yahaya[27], Alpha K. Keita[28,29], Amadou Diallo[30], Amadou Kone[31], Amal Souissi[32], Amel Chtourou[17], Ana V. Gutierrez[26], Andrew J. Page[26], Anika Vinze[33], Arash Iranzadeh[6,7], Arnold Lambisia[34], Arshad Ismail[35], Audu Rosemary[36], Augustina Sylverken[37], Ayoade Femi[14], Azeddine Ibrahimi[38], Baba Marycelin[39], Bamidele S. Oderinde[39], Bankole Bolajoko[14], Beatrice Dhaala[40], Belinda L. Herring[27], Berthe-Marie Njanpop-Lafourcade[27], Bronwyn Kleinhans[41], Bronwyn McInnis[10], Bryan Tegomoh[42], Cara Brook[43,44], Catherine B. Pratt[45], Cathrine Scheepers[35,46], Chantal G. Akoua-Koffi[47], Charles N. Agoti[34,48], Christophe Peyrefitte[30], Claudia Daubenberger[49], Collins M. Morang'a[50], D. James Nokes[34,51], Daniel G. Amoako[35], Daniel L. Bugembe[40], Danny Park[33], David Baker[26], Deelan Doolabh[7], Deogratius Ssemwanga[40,52], Derek Tshiabuila[1], Diarra Bassirou[30], Dominic S. Y. Amuzu[50], Dominique Goedhals[53], Donwilliams O. Omuoyo[34], Dorcas Maruapula[54], Ebenezer Foster-Nyarko[26], Eddy K. Lusamaki[18,19], Edgar Simulundu[55], Edidah M. Ong'era[34], Edith N. Ngabana[18,19], Edwin Shumba[56], Elmostafa El Fahime[57], Emmanuel Lokilo[18], Enatha Mukantwari[58], Eromon Philomena[14], Essia Belarbi[59], Etienne Simon-Loriere[60], Etilé A. Anoh[47], Fabian Leendertz[59], Faida Ajili[61], Fakayode O. Enoch[62], Fares Wasfi[63], Fatma Abdelmoula[32,64], Fausta S. Mosha[27], Faustinos T. Takawira[65], Fawzi Derrar[66], Feriel Bouzid[32], Folarin Onikepe[14], Fowotade Adeola[67], Francisca M. Muyembe[18,19], Frank Tanser[68,69,70], Fred A. Dratibi[27], Gabriel K. Mbunsu[19], Gaetan Thilliez[26], Gemma L. Kay[26], George Githinji[34,71], Gert van Zyl[41,72], Gordon A. Awandare[50], Grit Schubert[59], Gugu P. Maphalala[73], Hafaliana C. Ranaivoson[44], Hajar Lemriss[74], Happi Anise[14], Haruka Abe[75], Hela H. Karray[17], Hellen Nansumba[76], Hesham A. Elgahzaly[77], Hlanai Gumbo[65], Ibtihel Smeti[32], Ikhlas B. Ayed[32], Ikponmwosa Odia[25], Ilhem Boutiba Ben Boubaker[78,79], Imed Gaaloul[22], Inbal Gazy[80], Innocent Mudau[7], Isaac Ssewanyana[76], Iyaloo Konstantinus[81], Jean B. Lekana-Douk[82], Jean-Claude C. Makangara[18,19], Jean-Jacques M. Tamfum[18,19], Jean-Michel Heraud[30,44], Jeffrey G. Shaffer[83], Jennifer Giandhari[1], Jingjing Li[84], Jiro Yasuda[75], Joana Q. Mends[85], Jocelyn Kiconco[52], John M. Morobe[34], John O. Gyapong[85], Johnson C. Okolie[14], John T. Kayiwa[40], Johnathan A. Edwards[68,86], Jones Gyamfi[85], Jouali Farah[87], Joweria Nakaseegu[52], Joyce M. Ngoi[50], Joyce Namulondo[52], Julia C. Andeko[82], Julius J. Lutwama[40], Justin O'Grady[26], Katherine Siddle[33], Kayode T. Adeyemi[14], Kefentse A. Tumedi[88], Khadija M. Said[34], Kim Hae-Young[89], Kwabena O. Duedu[85], Lahcen Belyamani[38], Lamia Fki-Berrajah[17], Lavanya Singh[1], Leonardo de O. Martins[26], Lynn Tyers[7], Magalutcheemee Ramuth[91], Maha Mastouri[22,23], Mahjoub Aouni[22], Mahmoud el Hefnawi[92], Maitshwarelo I. Matsheka[88], Malebogo Kebabonye[93], Mamadou Diop[30], Manel Turki[32], Marietou Paye[33], Martin M. Nyaga[94], Mathabo Mareka[95], Matoke-Muhia Damaris[96], Maureen W. Mburu[34], Maximillian Mpina[49,97,98], Mba Nwando[99], Michael Owusu[100], Michael R. Wiley[45], Mirabeau T. Youtchou[101], Mitoha O. Ayekaba[97], Mohamed Abouelhoda[102,103], Mohamed G. Seadawy[104], Mohamed K. Khalifa[20], Mooko Sekhele[95], Mouna Ouadghiri[38], Moussa M. Diagne[30], Mulenga Mwenda[105], Mushal Allam[35], My V. T. Phan[40], Nabil Abid[79,106], Nadia Touil[107], Nadine Rujeni[108,109], Najla Kharrat[32], Nalia Ismael[110], Ndongo Dia[30], Nedio Mabunda[110], Nei-yuan Hsiao[7111], Nelson B. Silochi[97], Ngoy Nsenga[27], Nicksy Gumede[27], Nicola Mulder[112], Nnaemeka Ndodo[99], Norosoa H Razanajatovo[44], Nosamiefan Iguosadolo[14], Oguzie Judith[14], Ojide C. Kingsley[113], Okogbenin Sylvanus[25], Okokhere Peter[25], Oladiji Femi[114], Olawoye Idowu[14], Olumade Testimony[14], Omoruyi E. Chukwuma[67], Onwe E. Ogah[115], Chika K. Onwuamah[36,138], Oshomah Cyril[25], Ousmane Faye[30], Oyewale Tomori[14], Pascale Ondoa[56], Patrice Combe[116], Patrick Semanda[76], Paul E. Oluniyi[14], Paulo Arnaldo[110], Peter K. Quashie[50], Philippe Dussart[44], Phillip A. Bester[53], Placide K. Mbala[18,19], Reuben Ayivor-Djanie[85], Richard Njouom[117], Richard O. Phillips[118], Richmond Gorman[118], Robert A. Kingsley[26], Rosina A. A. Carr[85], Saâd El Kabbaj[119], Saba Gargouri[17], Saber Masmoudi[32], Safietou Sankhe[30], Salako B. Lawal[36], Samar Kassim[77], Sameh Trabelsi[120], Samar Metha[33], Sami Kammoun[121], Sanaâ Lemriss[122], Sara H. A. Agwa[77], Sébastien Calvignac-Spencer[59], Stephen F. Schaffner[33], Seydou Doumbia[31], Sheila M. Mandanda[18,19], Sherihane Aryeetey[123], Shymaa S. Ahmed[123], Siham Elhamoumi[33], Soafy Andriamandimby[44], Sobajo Tope[14], Sonia Lekana-Douki[82], Sophie Prosolek[26], Soumeya Ouangraoua[124,125], Steve A. Mundeke[18,19], Steven Rudder[26], Sumir Panji[112], Sureshnee Pillay[1], Susan Engelbrecht[41,72], Susan Nabadda[76], Sylvie Behillil[126], Sylvie L. Budiaki[95], Sylvie van der Werf[126], Tapfumanei Mashe[65], Tarik Aanniz[38], Thabo Mohale[35], Thanh Le-Viet[26], Tobias Schindler[49,97], Ugochukwu J. Anyaneji[1], Ugwu Chinedu[14], Upasana Ramphal[1,69,127], Uwanibe Jessica[14], Uwem George[14], Vagner Fonseca[1,4,128], Vincent Enouf[126], Vivianne Gorova[129,130], Wael H. Roshdy[123], William K. Ampofo[50], Wolfgang Preiser[41,72], Wonderful T. Choga[54,131], Yaw Bediako[50], Yeshnee Naidoo[1], Yvan Butera[108,132,133], Zaydah R. de Laurent[34], Amadou A. Sall[30], Ahmed Rebai[32], Anne von Gottberg[35,139], Bourema Kouriba[12], Carolyn Williamson[7,69,111], Daniel J. Bridges[105], Ihekweazu Chikwe[99], Jinal N. Bhiman[35,139], Madisa Mine[134], Matthew Cotten[40,135], Sikhulile Moyo[54,136], Simani Gaseitsiwe[54,136], Ngonda Saasa[55], Pardis C. Sabeti[33], Pontiano Kaleebu[40], Yenew K. Tebeje[21], Sofonias K. Tessema[21], Christian Happi[14], John Nkengasong[21], Tulio de Oliveira[1,2,69,137]*

The progression of the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) pandemic in Africa has so far been heterogeneous, and the full impact is not yet well understood. In this study, we describe the genomic epidemiology using a dataset of 8746 genomes from 33 African countries and two overseas territories. We show that the epidemics in most countries were initiated by importations predominantly from Europe, which diminished after the early introduction of international travel restrictions. As the pandemic progressed, ongoing transmission in many countries and increasing mobility led to the emergence and spread within the continent of many variants of concern and interest, such as B.1.351, B.1.525, A.23.1, and C.1.1. Although distorted by low sampling numbers and blind spots, the findings highlight that Africa must not be left behind in the global pandemic response, otherwise it could become a source for new variants.

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) emerged in late 2019 in Wuhan, China (*1*, *2*). Since then, the virus has spread to all corners of the world, causing almost 150 million cases of COVID-19 and more than 3 million deaths by the end of April 2021. Throughout the pandemic, it has been noted that Africa accounts for a relatively low proportion of reported cases and deaths—by the end of April 2021, there had been ~4.5 million cases and ~120,000 deaths on the continent, corresponding to less than 4% of the global burden. However, emerging data from seroprevalence surveys and autopsy studies in some African countries suggest that the true number of infections and deaths may be severalfold higher than reported (*3*, *4*). In addition, a recent analysis has shown that in

many African countries, the second wave of the pandemic was more severe than the first wave (*5*).

The first cases of COVID-19 on the African continent were reported in Nigeria, Egypt, and South Africa between mid-February and early March 2020, and most countries had reported cases by the end of March 2020 (*6*–*8*). These early cases were concentrated among airline travelers returning from regions of the world with high levels of community transmission. Many African countries introduced early public health and social measures, including international travel controls, quarantine for returning travelers, and internal lockdown measures, to limit the spread of the virus and give health services time to prepare (*5*, 9). The initial phase of the epidemic was then heterogeneous, with relatively high case numbers reported in North Africa and southern Africa, and fewer cases reported in other regions.

From the onset of the pandemic, genomic surveillance has been at the forefront of the COVID-19 response in Africa (*10*). Rapid implementation of SARS-CoV-2 sequencing by various laboratories in Africa enabled genomic data to be generated and shared from the early imported cases. In Nigeria, the first genome sequence was released just 3 days after the announcement of the first case (*6*). Similarly, in Uganda, a sequencing program was set up rapidly to facilitate virus tracing, and the collection of samples for sequencing began immediately upon confirmation of the first case (*11*). In South Africa, the Network for Genomic Surveillance in South Africa (NGS-SA) was established in March 2020, and within weeks, genomic analysis was helping to characterize outbreaks and community transmission (*12*).

Genomic surveillance has also been critical for monitoring ongoing SARS-CoV-2 evolution and detection of new SARS-CoV-2 variants in Africa. Intensified sampling by NGS-SA in the Eastern Cape Province of South Africa in November 2020, in response to a rapid resurgence of cases, led to the detection of B.1.351 (501Y.V2) (*13*). This variant was subsequently designated a variant of concern (VOC) by the World Health Organization (WHO), owing to evidence of increased transmissibility (*14*) and resistance to neutralizing antibodies elicited by natural infection and vaccines (*15*–*17*).

In this study, we performed phylogenetic and phylogeographic analyses of SARS-CoV-2 genomic data from 33 African countries and two overseas territories to help characterize the dynamics of the pandemic in Africa. We show that the early introductions were predominantly from Europe, but that as the pandemic progressed, there was increasing spread between African countries. We also describe the emergence and spread of a number of key SARS-CoV-2 variants in Africa and highlight how the spread of B.1.351 (501Y.V2) and other variants contributed to the more severe second wave of the pandemic in many countries.

All author affiliations are listed at the end of this paper.
***Corresponding author. Email: tulio@sun.ac.za**
†These authors contributed equally to this work.

SARS-CoV-2 genomic data

By 5 May 2021, 14,504 SARS-CoV-2 genomes had been submitted to the GISAID database (*18*) from 38 African countries and two overseas territories (Mayotte and Réunion) (Fig. 1A). Overall, this corresponds to approximately one sequence per ~300 reported cases. Almost half of the sequences were from South Africa (n = 5362), consistent with it being responsible for almost half of the reported cases in Africa. Overall, the number of sequences correlates closely with the number of reported cases per country (Fig. 1B). The countries and territories with the highest coverage of sequencing (defined as genomes per reported case) are Kenya (n = 856, one sequence per ~203 cases), Mayotte (n = 721, one sequence per ~21 cases), and Nigeria (n = 660, one sequence per ~250 cases). Although genomic surveillance started early in many countries, few have evidence of consistent sampling across the whole year. Half of all African genomes were deposited in the first 10 weeks of 2021, suggesting intensified surveillance in the second wave after the detection of B.1.351 (501Y.V2) and other variants (Fig. 1, C and D).

Genetic diversity and lineage dynamics in Africa

Of the 10,326 genomes retrieved from GISAID by the end of March 2021, 8746 genomes passed quality control and met the minimum metadata requirements. These genomes from Africa were compared in a phylogenetic framework with 11,891 representative genomes from around the world. Ancestral location state reconstruction of the dated phylogeny (hereafter referred to as discrete phylogeographic reconstruction) allowed us to infer the number of viral imports and exports between Africa and the rest of the world, and between individual African countries. African genomes in this study spanned the whole global genetic diversity of SARS-CoV-2, a pattern that largely reflects multiple introductions over time from the rest of the world (Fig. 2A).

In total, we detected at least 757 [95% confidence interval (CI): 728 to 786] viral introductions into African countries between the start of 2020 and February 2021, more than half of which occurred before the end of May 2020. Although the early phase of the pandemic was dominated by importations from outside Africa, predominantly from Europe, there was then a shift in the dynamics, with an increasing number of importations from other African countries as the pandemic progressed (Fig. 2, B and C). A rarefaction analysis in which we systematically subsampled genomes shows that vastly more introductions would have likely been identified with increased sampling in Africa or globally, suggesting that the introductions we identified are really just the "ears of the hippo," or a small part of a larger problem (fig. S1).

South Africa, Kenya, and Nigeria appear as major sources of importations into other African countries (Fig. 2D), although this is likely to be influenced by these three countries having the greatest number of deposited sequences. Particularly notable is the southern African region, where South Africa is the source for a large proportion (~80%) of the importations to other countries in the region. The North African region demonstrates a different pattern to the rest of the continent, with more viral introductions from Europe and Asia (particularly the Middle East) than from other African countries (fig. S2).

Africa has also contributed to the international spread of the virus, with at least 324 (95% CI: 295 to 353) exportation events from Africa to the rest of the world detected in this dataset. Consistent with the source of importations, most exports were to Europe (41%), Asia (26%), and North America (14%). As with the number of importations, exports were relatively evenly distributed over the 1-year period (fig. S3). However, an increase in the number of exportation events occurred between December 2020 and March 2021, which coincided with the second wave of infections in Africa and with some relaxations of travel restrictions around the world.

The early phase of the pandemic was characterized by the predominance of lineage B.1. This was introduced multiple times to African countries and has been detected in all but one of the countries included in this analysis. After its emergence in South Africa, B.1.351 became the most frequently detected SARS-CoV-2 lineage found in Africa (n = 1769, ~20%) (Fig. 1C). It was first sampled on 8 October 2020 in South Africa (*13*) and has since spread to 20 other African countries.

As air travel came to an almost complete halt in March and April 2020, the number(s) of detectable viral imports into Africa decreased and the pandemic entered a phase that was characterized in sub-Saharan Africa by sustained low levels of within-country movements and occasional international viral movements between neighboring countries, presumably via road and rail links between these. Though some border posts between countries were closed during the initial lockdown period (table S1), others remained open to allow trade to continue. Regional trade in southern Africa was only slightly affected by lockdown restrictions and quickly rebounded to prepandemic levels (fig. S4) after the relaxation of restrictions between June 2020 and December 2020.

Although lineage A viruses were imported into several African countries, they only account for 1.3% of genomes sampled in Africa. Despite lineage A viruses initially causing many localized clustered outbreaks, each the result of independent introductions to several countries (e.g., Burkina Faso, Côte d'Ivoire, and Nigeria), they were later largely replaced by lineage B viruses as the pandemic evolved. This is possibly due to the increased transmissibility of lineage B viruses by virtue of the D614G (Asp^{614}→Gly) mutation in the spike protein (*19*, *20*). However, there is evidence of an increasing prevalence of lineage A viruses in some African countries (*11*). In particular, A.23.1 emerged in East Africa and appears to be rapidly increasing in prevalence in Uganda and Rwanda (*11*). Furthermore, a highly divergent variant from lineage A was recently identified in Angola from individuals arriving from Tanzania (*21*).

Emergence and spread of new SARS-CoV-2 variants

To determine how some of the key SARS-CoV-2 variants are spreading within Africa, we performed phylogeographic analyses on the VOC B.1.351, the variant of interest (VOI) B.1.525, and two additional variants that emerged and that we designated as VOIs for this analysis (A.23.1 and C.1.1). These African VOCs and VOIs have multiple mutations on the spike glycoprotein, and a molecular clock analysis of these four datasets provided strong evidence that these four lineages are evolving in a clock-like manner (Fig. 3, A and B).

B.1.351 was first sampled in South Africa in October 2020, but phylogeographic analysis suggests that it emerged earlier, around August 2020. It is defined by 10 mutations in the spike protein, including K417N (Lys^{417}→Asn), E484K (Glu^{484}→Lys), and N501Y (Asn^{501}→Tyr) in the receptor binding domain (Fig. 3B). After its emergence in the Eastern Cape, it spread extensively within South Africa (Fig. 4A). By November 2020, the variant had spread into neighboring Botswana and Mozambique, and by December 2020, it had reached Zambia and Mayotte. Within the first 3 months of 2021,

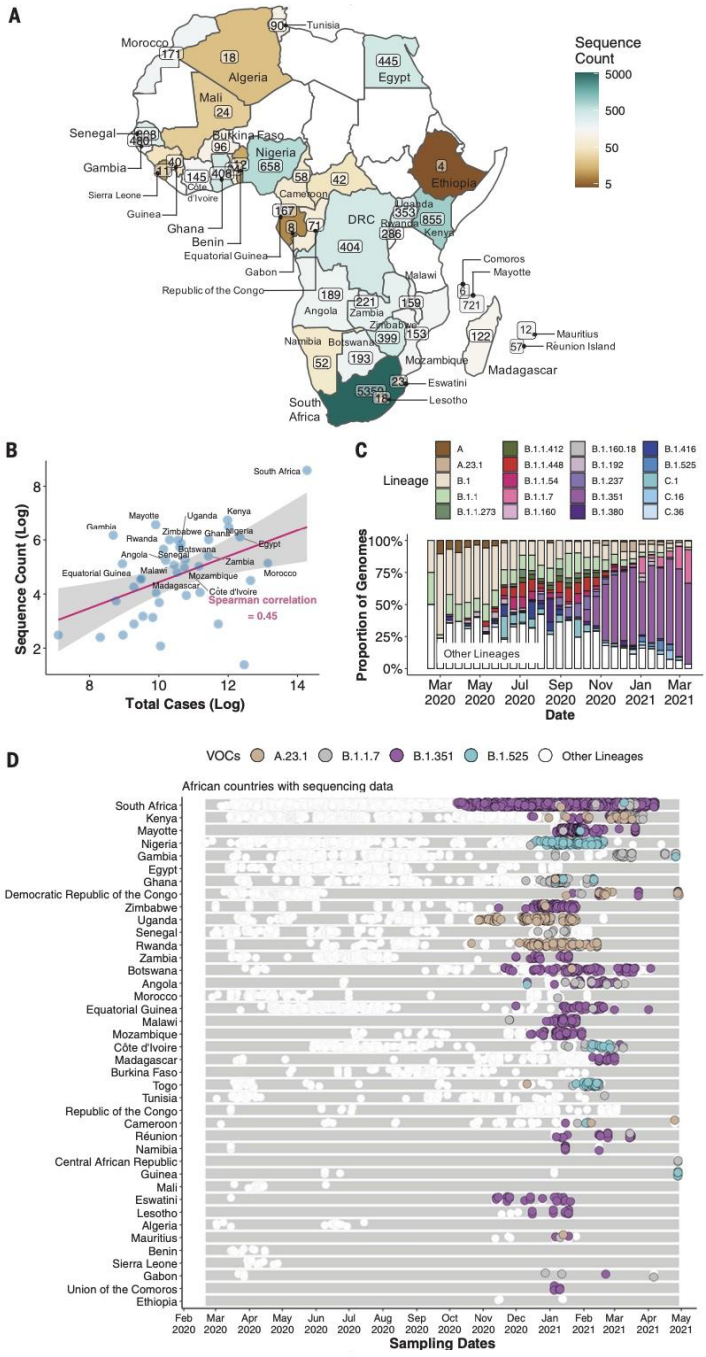


Fig. 1. SARS-CoV-2 sequences in Africa. (**A**) Map of the African continent with the number of SARS-CoV-2 sequences reflected in GISAID as of 5 May 2021. (**B**) Regression plot of the number of viral sequences versus the number of reported COVID-19 cases in various African countries as of 5 May 2021. Countries with >500 sequences are labeled. The shaded region indicates the 95% confidence interval. (**C**) Progressive distribution of the top 20 PANGO lineages on the African continent. (**D**) Temporal sampling of SARS-CoV-2 sequences in African countries (ordered by total number of sequences) through time, with VOCs of note highlighted and annotated according to their PANGO lineage assignment.

further exports from South Africa into Botswana, Zimbabwe, Mozambique, and Zambia occurred. By March 2021, B.1.351 had become the dominant lineage within most southern African countries as well as the overseas territories of Mayotte and Réunion (fig. S5). Our phylogeographic reconstruction also demonstrates movement of B.1.351 into East and Central Africa directly from southern Africa. Our discrete phylogeographic analysis of a wider sample of B.1.351 isolates demonstrates the spread of the lineage into West Africa. This patient from West Africa had a known travel history to Europe, so it is possible that the patient acquired the infection while in Europe or in transit and not from other African sources (fig. S6).

B.1.525 is a VOI defined by six substitutions in the spike protein [Q52R (Gln^{52}→Arg), A67V (Ala^{67}→Val), E484K, D614G, Q677H (Gln^{677}→His), and F888L (Phe^{888}→Leu)] and two deletions in the N-terminal domain [HV69-70Δ (deletion of His and Val at positions 69 and 70) and Y144Δ (deletion of Tyr at position 144)]. This was first sampled in the United Kingdom in mid-December 2020, but our phylogeographic reconstruction suggests that the variant originated in Nigeria in November 2020 [95% highest posterior density (HPD) 2020-11-01 to 2020-12-03] (Fig. 4B). Since then, it has spread throughout much of Nigeria and neighboring Ghana. Given sparse sampling from other neighboring countries within West and Central Africa (Fig. 1, A and C), the extent of the spread of this VOI in the region is not clear. Beyond Africa, this VOI has spread to Europe and the United States (fig. S6).

We designated A.23.1 and C.1.1 as VOIs for the purposes of this analysis because they present good examples of the continued evolution of the virus within Africa (*11*, *13*). Lineage A.23, characterized by three spike mutations [F157L (Phe^{157}→Leu), V367F (Val^{367}→Phe), and Q613H (Gln^{613}→His)], was first detected in a Ugandan prison in Amuru in July 2020 (95% HPD: 2020-07-15 to 2020-08-02). From there, the lineage was transmitted to Kitgum prison, possibly facilitated by the transfer of prisoners. Subsequently, the A.23 lineage spilled into the

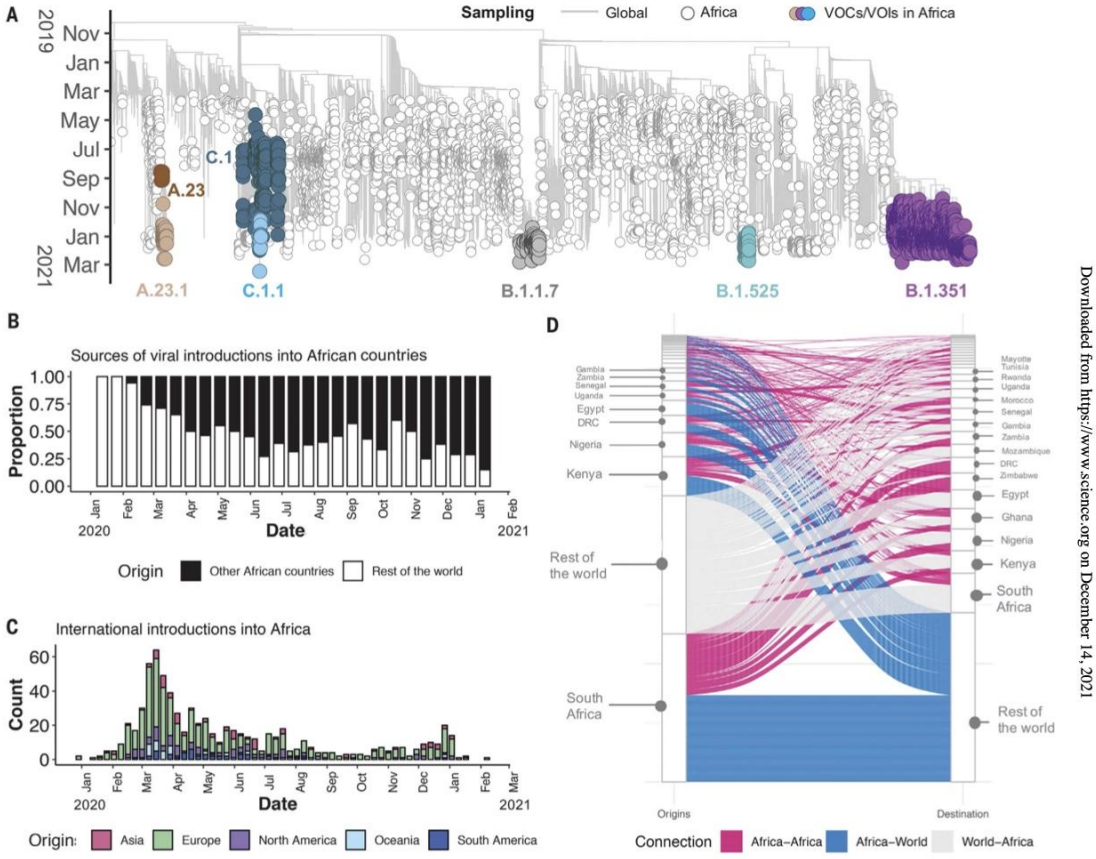


Fig. 2. Phylogenetic reconstruction of the SARS-CoV-2 pandemic on the continent of Africa. (**A**) Time-resolved maximum likelihood tree containing 8746 high-quality African SARS-CoV-2 near-full-genome sequences analyzed against a backdrop of global reference sequences. VOIs and VOCs are highlighted on the phylogeny. (**B**) Sources of viral introductions into African countries characterized as external introductions from the rest of the world versus internal introductions from other African countries. (**C**) Total external viral introductions over time into Africa. (**D**) The number of viral imports and exports into and out of various African countries depicted as internal (between African countries, in pink) or external (between African and non-African countries, in blue and gray).

general population and spread to Kampala, adding other spike mutations [R102I (Arg^{102}→Ile), L141F (Leu^{141}→Phe), E484K, and P681R (Pro^{681}→Arg)] along with additional mutations in nsp3, nsp6, ORF8, and ORF9, prompting a new lineage classification, A.23.1 (Fig. 3, A and B). Since the emergence of A.23.1 in September 2020 (95% HPD: 2020-09-02 to 2020-09-28), it has spread regionally into neighboring Rwanda and Kenya and has now also reached South Africa and Botswana in the south and Ghana in the west (Fig. 4C). However, our phylogeographic reconstruction of A.23.1 suggests that the introduction into Ghana may have occurred via Europe (fig. S6), whereas the introductions

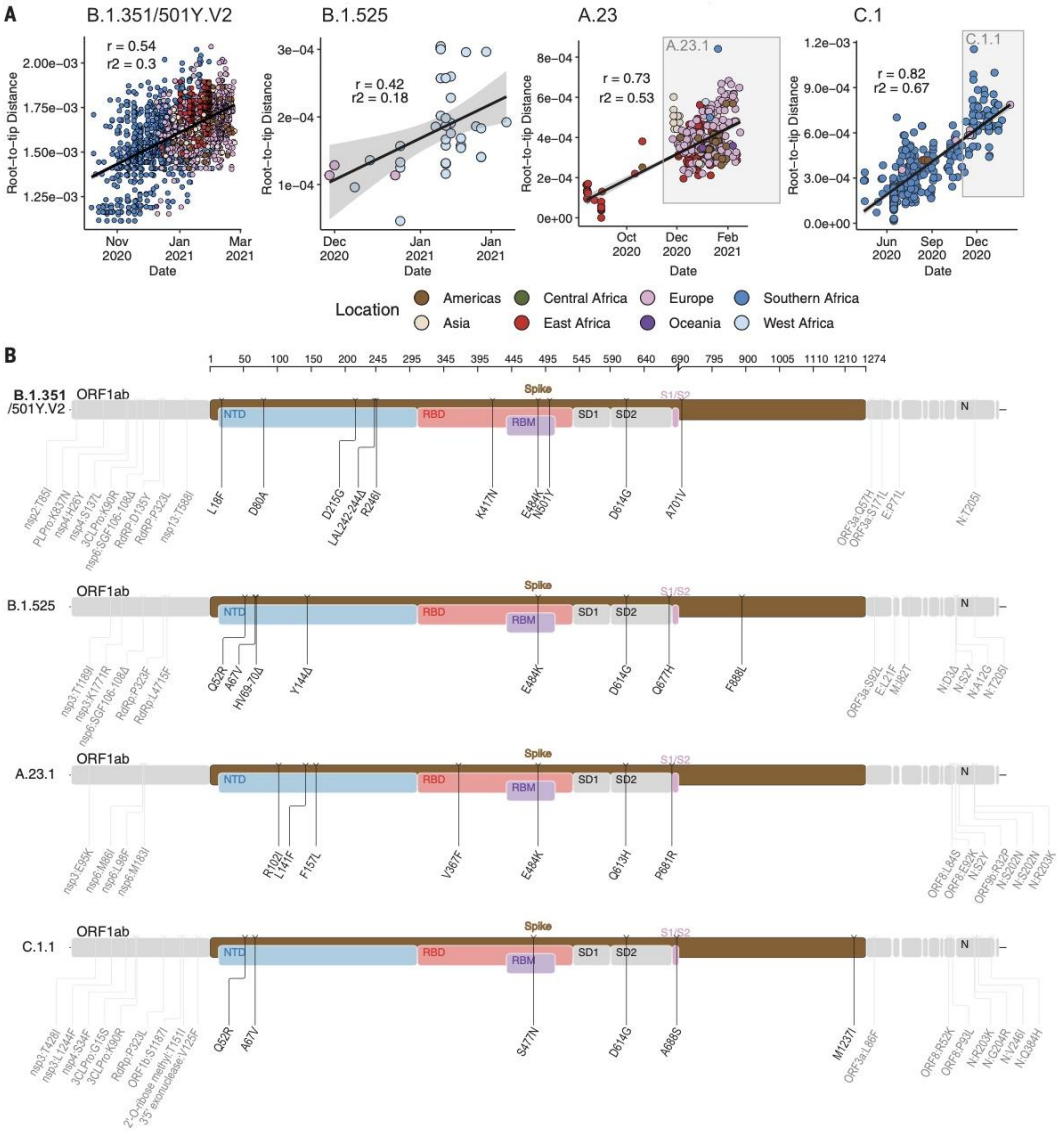


Fig. 3. Genetic profile of VOCs and VOIs under investigation. (**A**) Root-to-tip regression plots for four lineages of interest. C.1 and A.23 show continued evolution into VOIs C.1.1 and A.23.1, respectively. *r*, coefficient of correlation; r^2, coefficient of determination. (**B**) Genome maps of four VOCs and VOIs, where the spike region is shown in detail and in color and the rest of the genome is shown in gray. ORF, open reading frame; NTD, N-terminal domain; RBD, receptor binding domain; RBM, receptor binding motif; SD1, subdomain 1; SD2, subdomain 2.

into southern Africa likely occurred directly from East Africa. This is consistent with epidemiological data suggesting that the case detected in South Africa was a contact of an individual who had recently traveled to Kenya.

Lineage C.1 emerged in South Africa in March 2020 (95% HPD: 2020-03-13 to 2020-04-17) during a cluster outbreak before the first wave of the epidemic (*13*). C.1.1 is defined by the spike mutations S477N (Ser^{477}→Asn), A688S (Ala^{688}→Ser), and M1237I (Met^{1237}→Ile) and also contains the Q52R and A67V mutations similar to B.1.525 (Fig. 3B). A continuous trait phylogeographic reconstruction of the movement dynamics of these lineages suggests that C.1 emerged in the city of Johannesburg and spread within South Africa during the first wave (Fig. 4D). Independent exports of C.1 from South Africa led to regional spread to Zambia (June to July 2020) and Mozambique (July to August 2020), and the evolution to C.1.1 seems to have occurred in Mozambique around mid-September 2020 (95% HPD: 2020-09-07 to 2020-10-05). An in-depth analysis of SARS-CoV-2 genotypes from Mozambique suggests that the C.1.1 lineage was the most prevalent in the country until the introduction of B.1.351, which has dominated the epidemic since (fig. S5).

The VOC B.1.1.7, which was first sampled in Kent, England, in September 2020 (*22*), has also increased in prevalence in several African countries (fig. S5). To date, this VOC has been detected in 11 African countries, as well as the Indian Ocean islands of Mauritius and Mayotte (fig. S7). The time-resolved phylogeny suggests that this lineage was introduced into Africa on at least 16 occasions between November 2020 and February 2021, with evidence of local transmission in Nigeria and Ghana.

Conclusions

Our phylogeographic reconstruction of past viral dissemination patterns suggests a strong epidemiological linkage between Europe and Africa, with 64% of detectable viral imports into Africa originating in Europe and 41% of detectable viral exports from Africa landing in Europe (Fig. 1C). This phylogeographic analysis also

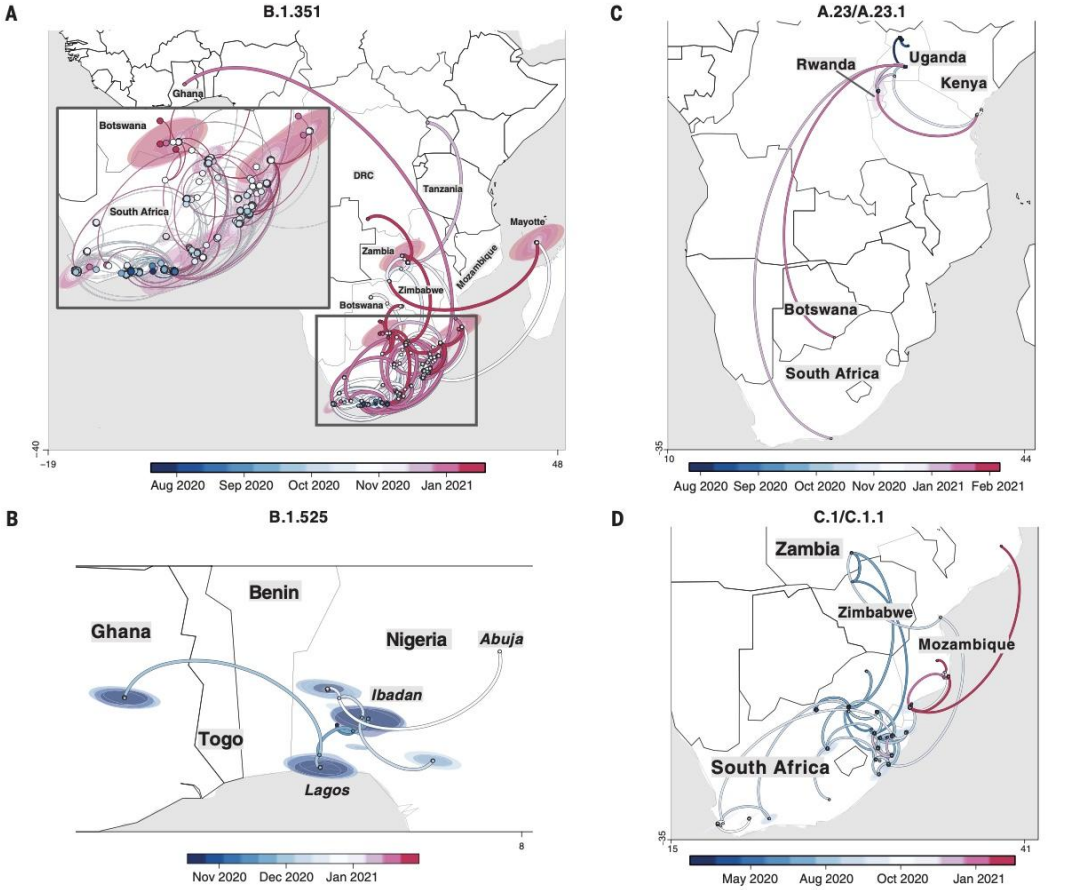


Fig. 4. Phylogeographic reconstruction of the spread of four VOCs and VOIs across the African continent. (**A** to **D**) Phylogeographic reconstruction of the spread of four VOCs and VOIs across the African continent using sequences showing strict continuous transmission across geographical regions: B.1.351 (A), B.1.525 (B), A.23/A.23.1 (C), and C.1/C.1.1 (D). Curved lines denote the direction of transmission in the counterclockwise direction. Solid lines show transmission paths as inferred by phylogeographic reconstruction and colored by date, whereas dashed lines show the known travel history of the particular case considered.

suggests a changing pattern of viral diffusion into and within Africa over the course of 2020. In almost all instances, the earliest introductions of SARS-CoV-2 into individual African countries were from countries outside Africa.

High rates of COVID-19 testing and consistent genomic surveillance in the south of the continent have led to the early identification of VOCs such as B.1.351 and VOIs such as C.1.1 (*13*). Since the discovery of these southern African variants, several other SARS-CoV-2 VOIs have emerged in different parts of the world, including elsewhere on the African continent, such as B.1.525 in West Africa and A.23.1 in East Africa. There is strong evidence that both of these VOIs are rising in frequency in the regions where they have been detected, which suggests that they may possess higher fitness than other variants in these regions. Although more-focused research on the biological properties of these VOIs is needed to confirm whether they should be considered VOCs, it would be prudent to assume the worst and focus on limiting their spread. It will be important to investigate how these different variants compete against one another if they occupy the same region.

Our focused phylogenetic analysis of the B.1.351 lineage revealed that in the final months of 2020, this variant spread from South Africa into neighboring countries, reaching as far north as the Democratic Republic of the Congo (DRC) by February 2021. This spread may have been facilitated through rail and road networks that form major transport arteries linking South Africa's ocean ports to commercial and industrial centres in Botswana, Zimbabwe, Zambia, and the southern parts of the DRC. The rapid, apparently unimpeded spread of B.1.351 into these countries suggests that current land-border controls that are intended to curb the international spread of the virus are ineffective. Perhaps targeted testing of cross-border travelers, genotyping of positive cases, and the focused tracking of frequent cross-border travelers, such as long distance truckers, would more effectively contain the spread of future VOCs and VOIs that emerge within this region.

The dominance of VOIs and VOCs in Africa has important implications for vaccine rollouts on the continent. For one, slow rollout of vaccines in most African countries creates an environment in which the virus can replicate and evolve: This will almost certainly produce additional VOCs, any of which could derail the global fight against COVID-19. Conversely, with the already widespread presence of known variants, difficult decisions about balancing reduced efficacy and availability of vaccines have to be made. This also highlights how crucial it is that trials are done. From a public health perspective, genomic surveillance is only one item in the toolkit of pandemic preparedness. It is important that such work is closely followed by genotype-to-phenotype research to determine the actual relevance of continued evolution of SARS-CoV-2 and other emerging pathogens.

The rollout of vaccines across Africa has been painfully slow (figs. S8 and S9). There have, however, been notable successes that suggest that the situation is not hopeless. The small island nation of the Seychelles had vaccinated 70% of its population by May 2021. Morocco has kept pace with many developed nations and, by mid-March, had vaccinated ~16% of its population. Rwanda, one of Africa's most resource-constrained countries, had, within 3 weeks of obtaining its first vaccine doses in early March, managed to provide first doses to ~2.5% of its population. For all other African countries, at the time of writing, vaccine coverage (first dose) was <1.0% of the general population.

The effectiveness of molecular surveillance as a tool for monitoring pandemics is largely dependent on continuous and consistent sampling through time, rapid virus genome sequencing, and rapid reporting. When this is achieved, molecular surveillance can ensure the early detection of changing pandemic characteristics. Further, when such changes are discovered, molecular surveillance data can also guide public health responses. In this regard, the molecular surveillance data that are being gathered by most African countries are less useful than they could be. For example, the time lag between when virus samples are taken and when sequences for these samples are deposited in sequence repositories is so great in some cases that the primary utility of genomic surveillance data is lost (fig. S10). This lag is driven by several factors, depending on the laboratory or country in question: (i) lack of reagents owing to disruptions in global supply chains, (ii) lack of equipment and infrastructure within the originating country, (iii) scarcity of technical skills in laboratory methods or bioinformatic support, and (iv) hesitancy by some health officials to release data. More-recent sampling and prompt reporting is crucial to reveal the genetic characteristics of currently circulating viruses in these countries.

The patchiness of African genomic surveillance data is therefore the main weakness of our study. However, there is evidence that the situation is improving, with ~50% of African SARS-CoV-2 genome sequences having been submitted to the GISAID database within the first 10 weeks of 2021. Although the precise factors underlying this surge in sequencing efforts are unclear, an important driver is almost certainly increased global interest in genomic surveillance after the discovery of multiple VOCs and VOIs since December 2020. We cannot reject that the observed increase in exports from Africa may be due to intensified sequencing activity after the detection of variants around the world. It is important to note here that phylogeographic reconstruction of viral spread is highly dependent on sampling where there is the caveat that the exact routes of viral movements between countries cannot be inferred if there is no sampling in connecting countries. Furthermore, our efforts to reconstruct the movement dynamics of SARS-CoV-2 across the continent are almost certainly biased by uneven sampling between different African countries. It is not a coincidence that we identified South Africa, Kenya, and Nigeria, which have sampled and sequenced the most SARS-CoV-2 genomes, as major sources of viral transmissions between sub-Saharan African countries. However, these countries also had the highest number of infections, which may decrease the sampling biases (Fig. 1A).

The reliability of genomic surveillance as a tool to prevent the emergence and spread of dangerous variants is dependent on the intensity with which it is embraced by national public health programs. As with most other parts of the world, the success of genomic surveillance in Africa requires that more samples are tested for COVID-19, higher proportions of positive samples are sequenced within days of sampling, and persistent analyses of these sequences are performed for concerning signals such as (i) the presence of novel nonsynonymous mutations at genomic sites associated with pathogenicity and immunogenicity, (ii) evidence of positive selection at codon sites where nonsynonymous mutations are observed, and (iii) evidence of lineage expansions. Despite limited sampling, Africa has identified many of the VOCs and VOIs that are being transmitted across the world. Detailed characterization of the variants and their impact on vaccine-induced immunity is of extreme importance. If the pandemic is not controlled in Africa, we may see the production of vaccine escape variants that may profoundly affect the population in Africa and across the world.

REFERENCES AND NOTES

1. C. Wang, P. W. Horby, F. G. Hayden, G. F. Gao, *Lancet* **395**, 470–473 (2020).
2. Q. Li *et al.*, *N. Engl. J. Med.* **382**, 1199–1207 (2020).
3. S. Uyoga *et al.*, *Science* **371**, 79–82 (2021).
4. L. Mwananyanda *et al.*, *BMJ* **372**, n334 (2021).
5. S. J. Salyer *et al.*, *Lancet* **397**, 1265–1275 (2021).
6. P. Oluniyi, First African SARS-CoV-2 genome sequence from Nigerian COVID-19 case. Virological (2020); https://virological.org/t/first-african-sars-cov-2-genome-sequence-from-nigerian-covid-19-case/421.
7. M. A. Medhat, M. El Kassas, *J. Glob. Health* **10**, 010368 (2020).
8. M. Allam *et al.*, *Microbiol. Resour. Announc.* **9**, e00572-e20 (2020).
9. N. Haider *et al.*, *BMJ Glob. Health* **5**, e003319 (2020).
10. S. C. Inzaule, S. K. Tessema, Y. Kebede, A. E. Ogwell Ouma, J. N. Nkengasong, *Lancet Infect. Dis.* **21**, e281–e289 (2021).
11. D. L. Bugembe *et al.*, *medRxiv* 2021.02.08.21251393 (2021); https://doi.org/10.1101/2021.02.08.21251393.
12. J. Giandhari *et al.*, *Int. J. Infect. Dis.* **103**, 234–241 (2021).
13. H. Tegally *et al.*, *Nat. Med.* **27**, 440–446 (2021).
14. C. A. Pearson, T. W. Russell, N. Davies, A. J. Kucharski, Estimates of severity and transmissibility of novel SARS-CoV-2 variant 501Y.V2 in South Africa. CMMID Repository (2021); https://cmmid.github.io/topics/covid19/sa-novel-variant.html.

15. S. Cele *et al.*, *medRxiv* 2021.01.26.21250224 (2021); https://doi.org/10.1101/2021.01.26.21250224.
16. S. A. Madhi *et al.*, *N. Engl. J. Med.* **384**, 1885–1898 (2021).
17. C. K. Wibmer *et al.*, *Nat. Med.* **27**, 622–625 (2021).
18. Y. Shu, J. McCauley, *Euro Surveill.* **22**, 30494 (2017).
19. E. Volz *et al.*, *Cell* **184**, 64–75.e11 (2021).
20. B. Korber *et al.*, *Cell* **182**, 812–827.e19 (2020).
21. T. de Oliveira *et al.*, *medRxiv* 2021.03.30.21254323 (2021); https://doi.org/10.1101/2021.03.30.21254323.
22. B. Meng *et al.*, *Cell Rep.* **35**, 109292 (2021).
23. S. E. James, T. Houriiyah, krisp-kwazulu-natal/africa-covid19-genomics: A year of genomic surveillance reveals how the SARS-CoV-2 pandemic unfolded in Africa - Code and scripts. Zenodo (2021); https://doi.org/10.5281/zenodo.5386379.

ACKNOWLEDGMENTS

We acknowledge the authors from the originating laboratories and the submitting laboratories, who generated and shared, via GISAID, the genetic sequence data on which this research is based (table S4). We also acknowledge the contribution of K. Maria from the NGS-SA platform for their contribution toward the sequencing effort in Cape Town, South Africa. Similarly, we thank A. M. Elsaame, S. M. Elsayed, and R. M. Darwish from the Faculty of Medicine Ain Shams Research Institute (MASRI) for their efforts toward sequencing in Egypt. We thank S. Bane, M. Sanogo, D. Diallo, A. Combo Georges Togo, and A. Coulibaly from the University Clinical Research Centre (UCRC) at the University of Sciences, Techniques, and Technologies of Bamako for the contribution they have made toward sequencing efforts in Mali. We acknowledge the contribution of M. Moeti and A. Salam Gueye from the WHO for their contribution toward combating SARS-CoV-2 on the African continent. We further wish to extend acknowledgment to S. Lutucuta and J. Morais from the Angolan Ministry of Health for their continued hard work with regards to SARS-CoV-2 sampling, sequencing, and pandemic response in Angola. From Malawi we wish to acknowledge the work of B. Chilima, B. Mvula, and M. Chitenje from the Malawian Ministry of Health for their work on the COVID-19 response within the country. **Funding:** The University of Ghana (WACCBIP) team was funded by a Wellcome/African Academy of Sciences Developing Excellence in Leadership Training and Science (DELTAS) grant (DEL-15-007 and 107755/Z/15/Z: Awandare); National Institute of Health Research (NIHR) (17.63.91) grants using UK aid from the UK government for a global health research group for genomic surveillance of malaria in West Africa (Wellcome Sanger Institute, UK) and the global research unit for Tackling Infections to Benefit Africa (TIBA partnership, University of Edinburgh); and a World Bank African Centres of Excellent grant (WACCBIP-NCDs: Awandare). Project ADAGE PRFCOV19-GP2 (2020-2022) includes 40 researchers from the Center of Biotechnology of Sfax, the University of Sfax, the University of Monastir, the University Hospital Hédi Chaker of Sfax, the Military Hospital of Tunis, and Dacima Consulting. Ministry of Higher Education and Scientific Research and Ministry of Health of the Republic of Tunisia. The Uganda contributions were funded by the UK Medical Research Council (MRC/UKRI) and the UK Department for International Development (DFID) under the MRC/DFID concordat agreement (grant agreement number NC_PC_19060) and by the Wellcome, DFID–Wellcome Epidemic Preparedness–Coronavirus (grant agreement number 220977/Z/20/Z) awarded to M.C. Work from Quadram Institute Bioscience was funded by The Biotechnology and Biological Sciences Research Council Institute Strategic Programme Microbes in the Food Chain BB/R012504/1 and its constituent projects BBS/E/F/000PR10348, BBS/E/F/000PR10349, BBS/E/F/000PR10351, and BBS/E/F/000PR10352 and by the Quadram Institute Bioscience BBSRC–funded Core Capability Grant (project number BB/CCG1860/1). The Africa Pathogen Genomics Initiative (Africa PGI) at the Africa CDC is supported by the Bill & Melinda Gates Foundation (INV018978 and INV018278), Illumina Inc, the US Centers for Disease Control and Prevention (CDC), and Oxford Nanopore Technologies. Sequences generated in Zambia through PATH were funded by the Bill & Melinda Gates Foundation. The findings and conclusions contained within are those of the authors and do not necessarily reflect positions or policies of the Bill & Melinda Gates Foundation. Funding for sequencing in Côte d'Ivoire, Burkina Faso, and part of the sequencing in the DRC was granted by the German Federal Ministry of Education and Research (BMBF). Sequencing efforts from Morocco have been supported by Academie Hassan II of Science and Technology, Morocco. Funding for surveillance, sampling, and testing in Madagasar was provided by the WHO, the CDC (grant U5/IP000812-05), the US Agency for International Development (USAID; cooperation agreement 72068719CA00001), and the Office of the Assistant Secretary for Preparedness and Response in the US Department of Health and Human Services (DHHS; grant number IDSEP190051-01-0200). Funding for sequencing was provided by the Bill & Melinda Gates Foundation (GCE/ID OPP1211841), Chan Zuckerberg Biohub, and the Innovative Genomics Institute at UC Berkeley. The Botswana Harvard AIDS Institute was supported by the following funding: H3ABioNet through funding from the National Institutes of Health Common Fund (U41HG006941)—H3ABioNet is an initiative of the Human Health and Heredity in Africa Consortium (H3Africa) program of the African Academy of Science (AAS); DHHS–NIH–National Institute of Allergy and Infectious Diseases (NIAID) (5K24AI131928-04 and 5K24AI131924-04); Sub-Saharan African Network for TB/HIV Research Excellence (SANTHE): a DELTAS Africa Initiative (grant DEL-15-0060—the DELTAS Africa Initiative is an independent funding scheme of the AAS's Alliance for Accelerating Excellence in Science in Africa (AESA) and supported by the New Partnership for Africa's Development Planning and Coordinating Agency (NEPAD Agency) with funding from the Wellcome Trust [grant 107752/Z/15/Z] and the UK government; and the South African Medical Research Council (SAMRC) and the Department of Technology and Innovation as part of the Network for Genomic Surveillance in South Africa (NGS-SA) and the Stellenbosch University Faculty of Medicine & Health Sciences, Strategic Equipment Fund. D.P.M. is funded by the Wellcome Trust (Wellcome Trust grant 222574/Z/21/Z). Sequencing activities at the NICD were 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 U.S. Centers for Disease Control and Prevention (grant number 5 U01IP001048-05-00); 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 grant number INV-018978; the UK Foreign, Commonwealth and Development Office and Wellcome (Grant no 221003/Z/20/Z); the South African Medical Research Council (Reference number SHIPNCD 76756); the UK Department of Health and Social Care, managed by the Fleming Fund and performed under the auspices of the SEQAFRICA project. Furthermore, pandemic surveillance in South Africa and Senegal was supported in part through NIH grant U01 AI151698 for the United World Antiviral Research Network (UWARN). Support for pandemic surveillance from the Tulio de Oliveira group to other African countries is funded by the Rockefeller Foundation. Sequencing efforts in the DRC were funded by the Bill & Melinda Gates Foundation under grant INV-018030 awarded to C.B.P. and further supported by funding from the Africa CDC through the ASLM (African Society of Laboratory Medicine) for Accelerating SARS-CoV-2 Genomic Surveillance in Africa. Sequencing efforts in Rwanda were commissioned by the NIHR Global Health Research program (16/136/33) using UK aid from the UK government (funding to E.M. and N. R. through TIBA partnership) and additional funds from the government of Rwanda through RBC/National Reference Laboratory in collaboration with the Belgian Development Agency (ENABEL) for additional genomic sequencing at GIGA Research Institute–Liege/Belgium. The sequencing effort in Equatorial Guinea was supported by a public-private partnership, the Bioko Island Malaria Elimination Project, composed of the government of Equatorial Guinea Ministries of Mines and Hydrocarbons, and Health and Social Welfare, Marathon EG Production Limited, Noble Energy, Atlantic Methanol Production Company, and EG LNG. Sample collection and typing in Mali were supported by Fondation Merieux–France, and sequence efforts have been supported by the Enable and Enhance Initiative of the German Federal Government's Security Cooperation against Biological Threats in the G5 Sahel Region. The Nigeria work was made possible by support from Flu Lab and a cohort of generous donors through TED's Audacious Project, including the ELMA Foundation, MacKenzie Scott, the Skoll Foundation, and Open Philanthropy. Further Nigeria funding came from grants from the NIAID (www.niaid.nih.gov), NIH-H3Africa (https://h3africa.org) (U01HG007480 and U54HG007480), and the World Bank grant (worldbank.org) (ACE IMPACT project) to C.H. Analysis for the Gabon strains was supported by the Science and Technology Research Partnership for Sustainable Development (SATREPS), Japan International Cooperation Agency (JICA), and Japan Agency for Medical Research and Development (AMED) (grant number JP20jm0110013) and a grant from AMED (grant number JP20wm0225003). Sequencing at KEMRI-Wellcome Trust Research Programme site in Kenya was supported by the National Institute for Health Research (NIHR) (project references 17/63/82 and 16/136/33), using UK aid from the UK Government to support global health research, and the UK Foreign, Commonwealth and Development Office and Wellcome Trust (grant# 102975; 220985). **Author contributions:** Conceptualization: E.W., H.T., J.G.S., J.O., J.O.G., K.O.D., R.A.D., R.A.K., R.L., S.K.T., S.Ma., T.d.O. Methodology: A.-R.N.Z., A.Re., C.N.A., D.P.M., D.A.R., E.W., H.T., J.A.E., J.Gy., J.G.S., K.H.Y., K.O.D., L.d.O.M., M.A.B., M.C., M.G., M.M.N., M.V.T.P., P.A., T.d.O., V.F. Investigation: A.E., A.Ir., A.-R.N.Z., A.Re., A.So., A.v.G., C.A.K., C.W., D.C., D.J.N., D.P.M., D.A.R., D.S., E.M., E.N.N., E.W., F.L., G.G., H.T., J.A.E., J.E.S., J.Gy., J.G.S., J.-J.M.T., J.L., J.Nk., J.Q.M., K.H.Y., M.-M.D., M.A.B., M.C., M.G., M.Mar., M.Mas., M.S., M.V.T.P., N.A., N.H.R., N.K., P.K., R.A.A.C., R.A.D., R.G., S.A.M., S.F.A., S.Ma., S.O., T.L.V., V.F., W.P. Sampling: A.-S.K., A.N.A., A.A.A., A.A.S., A.D., A.E., A.E.O.O., A.F., A.G., A.H., A.Ib., A.Ka., A.K.Sa., A.K.Se., A.L., A.M.O., A.O.O.O., A.P., A.Ro., A.Sy., A.F.V., A.V.G., A.v.G., B.B., B.L.H., B.Ko., B.Kl., B.Ma., B.N., B.S.O., B.T., C.N.A., C.D., C.P., C.S., C.W., D.Bas., D.C., D.D., D.G., D.G.A., D.J.N., D.S., E.K.L., E.M., E.M.O., E.N.N., E.P., E.Sh., F.Ab., F.Ad., F.O.E., F.M.M., F.S.M., F.O., F.T., F.T.T., G.A.A., G.G., G.K.M., G.P.M., G.T., G.v.Z., H.C., H.A.E., H.N., I.C., I.Gaa., I.Gaz., I.K., I.M., I.O., I.Ss., J.C.A., J.-C.M., J.B.L.-D., J.E.S., J.Gi., J.Gy., J.J.L., J.K., J.L., J.M.H., J.M.M., J.Nam., J.Nak., J.T.K., K.T.A., K.M.S., L.B., M.-M.D., M.Al., M.C., M.D., M.e.H., M.G.S., M.K.K., M.Mp., M.Mar., M.Mas., M.M.D., M.N., M.Ou., M.O.A., M.R., M.S., M.W.M., M.T.Y., N.A., N.G., N.H., N.H.R., N.Ig., N.K., N.Ma., N.Nd., N.B.S., N.S., O.Ch., O.E.C., O.Fa., O.Fe., O.I., O.J., O.C.K., O.E.O., O.P., O.S., O.Te., P.E.O., P.A.B., P.C., P.C.S., P.D., P.K., P.K.Q., P.O., P.S., R.A.D., R.G., R.N., S.A., S.S.A., S.B., S.B.L., S.D., S.El., S.E.K., S.F.A., S.Gas., S.Kam., S.L., S.L.D., S.Me., S.Mo., S.M.M., S.N., S.Pi., S.S., S.To., T.L.V., T.Ma., T.S., U.C., U.G., U.J., U.R., V.G., W.K.A., W.T.C., W.P., W.H.R., Y.Bu., Y.K.T., Y.N., Z.R.D. Sequencing: A.-S.Kam., A.N.A., A.A.A., A.A.S., A.C., A.D., A.E.O.O., A.F., A.Ib., A.Is., A.Ko., A.Ka., A.K.K., A.K.Sa., A.K.Se., A.L., A.-R.N.Z., A.P., A.Sa., A.So., A.Sy., A.S.O., A.J.T., A.F.V., A.V.G., A.v.G., A.A.Y., B.B., B.D., B.L.H., B.Ko., B.Kl., B.Mc., B.N., B.T., C.N.A., C.B., C.B.P., C.D., C.M.M., C.P., C.S., D.Bas., D.D., D.G., D.G.A., D.J.B., D.L.B., D.M., D.O.O., D.P., D.S.Y.A., D.T., E.E.F., E.F.N., E.K.L., E.L., E.M.O., E.P., E.Sh., E.Si., E.S.L., F.Ab., F.Aj., F.A.D., F.D., F.M.M., F.S.M., F.O., F.T.T., F.W., G.A.A., G.G., G.P.M., G.T., G.v.Z., H.Ab., H.An., H.C., H.C.R., H.A.E., H.G., H.H.K., H.N., I.B.-B.B., I.C., I.Gaa., I.Gaz., I.K., I.M., I.Ss., J.C.A., J.B., J.-C.M., J.B.L.-D., J.F., J.Gi., J.Gy., J.J.L., J.K., J.M.H., J.M.M., J.M.N., J.Nam., J.Nak., J.Q.M., J.T.K., J.Y., K.T.A., K.M.S., K.O.D., K.S., K.A.T., L.B., L.F., L.S., L.T., M.-M.D., M.Al., M.A.B., M.C., M.D., M.e.H., M.G.S., M.I.M., M.K.K., M.Mi., M.Mp., M.Mw., M.M.D., M.M.N., M.Ow., M.Ou., M.O.A., M.V.T.P., M.W.M., M.T.Y., N.D., N.G., N.H., N.Ig., N.Is., N.Ma., N.Nd., N.Ns., N.B.S., N.S., N.T., O.Cy., O.Ch., O.E.C., O.Fa., O.I., O.J., O.Te., P.A., P.A.B., P.C.S., P.D., P.E.O., P.K., P.K.Q., P.K.M., P.O., P.S., R.A.A.C., R.G., R.N., R.O.P., S.A., S.S.A., S.B., S.B.L., S.C.S., S.D., S.El., S.En., S.E.K., S.Gar., S.Gas., S.H.Ab., S.Kas., S.L., S.L.D., S.Me., S.Mo., S.Ma., S.M.M., S.N., S.Pi., S.Pr., S.R., S.S., S.To., S.Tr., S.v.d.W., T.A., T.Ma., T.Mo., T.S., U.C., U.G., U.J., U.J.A., U.R., V.G., W.K.A., W.T.C., W.H.R., Y.Bu., Y.Be., Y.K.T., Y.N., Z.R.D. Visualization: A.C., A.Is., A.Ko., A.K.K., A.Sa., A.So., A.A.Y., B.T., C.B., C.M.M., D.Bak., D.O.O., D.P., D.A.R., D.S.Y.A., E.A.A., E.B., E.S.L., E.W., F.A.D., F.B., F.D., F.W., G.S., H.Ab., H.An., H.G., H.L., H.T., I.B.A., I.Ss., J.A.E., J.B., J.F., J.Gy., J.M.N., J.Y., K.H.Y., K.S., L.F., L.S., L.T., M.Ao., M.Al., M.G., M.T., M.V.T.P., M.R.W., N.D., N.Is., N.K., N.Ns., N.T., O.Cy., O.To., P.A., P.C.S., P.E.O., R.A.A.C., S.B., S.F.S., S.H.A., S.Kas., S.Ma., T.A., T.Mo., V.E., Y.Be. Funding acquisition: A.J.P., A.Re., A.v.G., B.Ko., C.N.A., C.A.K., C.B.P., C.W., D.C., D.J.B., D.J.N., F.L., G.A.A., G.G., G.P.M., H.C., J.E.S., J.-J.M.T., J.L., J.M.H., J.Nk., J.O., K.O.D., M-M.D., M.C., M.I.M., M.Mas., M.V.T.P., N.A., P.C.S., P.K., P.K.M., R.A.K., S.A.M., S.El., S.Mo., S.v.d.W., T.d.O., W.P. Project administration: A.J.P., A.Re., A.F.V., A.v.G., B.Ko., C.W., D.J.B., D.J.N., E.W., F.Aj., F.T., G.A.A., G.P.M., G.S., G.T., H.C., J.C.O., J.-J.M.T., J.M.H., J.O., J.O.G., J.Y., K.O.D., M.C., M.K., M.Mar., M.P., M.V.T.P., M.R.W., N.R., O.To., P.C.S., P.K., P.K.M., R.A.K., S.A.M., S.El., S.F.S., S.Gas., S.Mo., T.d.O. Supervision: A.J.P., A.Re., B.Ko., C.W., D.J.N., E.N.N., E.W., F.T., G.A.A., G.L.K., H.C., J.B., J.M.H., J.Nk., J.O., J.O.G., K.O.D., M.Al., M.C., M.I.M., M.Mar., M.M.N., M.S., N.Mu., N.R., P.C.S., P.K., P.K.M., R.A.K., S.El., S.E.K., S.Gas., S.Me., S.Mo., S.Pa., T.d.O. Writing – original draft: A.K.Sa., A.-R.N.Z., B.Ko., D.P.M., E.W., F.T., G.L.K., H.T., J.B., J.-C.M., M.Al., M.A.B., M.C., M.G., M.Mi., N.Mu., R.L. Writing – review and editing: A.-R.N.Z., B.Ko., C.M.M., D.J.N., D.P.M., D.A.R., D.S.Y.A., D.T., E.K.L., E.L., E.S.L., E.W., H.T., J.E.S., J.G.S., L.d.O.M., M.A.B., M.C., M.e.H., P.K.Q., P.K.M., R.L., S.K.T., T.d.O., U.J.A. Author contributions are listed alphabetically. A full list of author abbreviations is included on the GitHub repository (https://github.com/krisp-kwazulu-natal/africa-covid19-genomics) (*23*). **Competing interests:** P.C.S. is a founder and shareholder of Sherlock Biosciences and is on the board and serves as shareholder of the Danaher Corporation. The authors declare no other conflicts of interest. **Data and materials availability:** All sequences that were used in the present study are listed in table S4 (accessible on the GitHub repository) along with their GISAID sequence IDs, dates of sampling, the originating and submitting laboratories, and main authors. All input files (e.g., alignments or XML files), all resulting output files, and scripts used in the study are shared publicly on GitHub (https://github.com/krisp-kwazulu-natal/africa-covid19-genomics) (*23*). This work is licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0) license, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. To view a copy of this license, visit https://creativecommons.org/licenses/by/4.0/. This license does not apply to

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.abj4336
Materials and Methods
Figs. S1 to S10
Tables S1 to S4
References (*24*–*38*)
MDAR Reproducibility Checklist

View/request a protocol for this paper from *Bio-protocol*.

[1]KwaZulu-Natal Research Innovation and Sequencing Platform (KRISP), Nelson R Mandela School of Medicine, University of KwaZulu-Natal, Durban, South Africa. [2]Centre for Epidemic Response and Innovation (CERI), School of Data Science and Computational Thinking, Stellenbosch University, Stellenbosch, South Africa. [3]Laboratorio de Flavivirus, Fundacao Oswaldo Cruz, Rio de Janeiro, Brazil. [4]Laboratório de Genética Celular e Molecular, Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, Brazil. [5]Department of Geography and GIS, University of Cincinnati, Cincinnati, OH, USA. [6]Institute of Infectious Diseases and Molecular Medicine, Department of Integrative Biomedical Sciences, Computational Biology Division, University of Cape Town, Cape Town, South Africa. [7]Division of Medical Virology, Wellcome Centre for Infectious Diseases in Africa, Institute of Infectious Disease and Molecular Medicine, University of Cape Town, Cape Town, South Africa. [8]Department of Entomology and Plant Pathology, North Carolina State University, Raleigh, NC, USA. [9]Bioinformatics Research Center, North Carolina State University, Raleigh, NC, USA. [10]Cancer Biology Department, Virology and Immunology Unit, National Cancer Institute, Cairo University, Cairo 11796, Egypt. [11]Centre d'Infectiologie Charles Mérieux-Mali (CICM-Mali), Bamako, Mali. [12]Bacteriology and Virology Department Souro Sanou University Hospital, Bobo-Dioulasso, Burkina Faso. [13]MRCG at LSHTM Genomics Lab, Fajara, Gambia. [14]African Centre of Excellence for Genomics of Infectious Diseases (ACEGID), Redeemer's University, Ede, Osun State, Nigeria. [15]Department of Virology, College of Medicine, University of Ibadan, Ibadan, Nigeria. [16]Department of Medical Microbiology and Parasitology, Faculty of Basic Clinical Sciences, College of Health Sciences, University of Ilorin, Ilorin, Kwara State, Nigeria. [17]CHU Habib Bourguiba, Laboratory of Microbiology, Faculty of Medicine of sFax, University of sFax, sFax, Tunisia. [18]Pathogen Sequencing Lab, Institut National de Recherche Biomédicale (INRB), Kinshasa, Democratic Republic of the Congo. [19]Université de Kinshasa (UNIKIN), Kinshasa, Democratic Republic of the Congo. [20]Genomics Research Program, Children's Cancer Hospital, Cairo, Egypt. [21]Institute of Pathogen Genomics, Africa Centres for Disease Control and Prevention (Africa CDC), Addis Ababa, Ethiopia. [22]Laboratory of Transmissible Diseases and Biological Active Substances (LR99ES27), Faculty of Pharmacy of Monastir, Monastir, Tunisia. [23]Laboratory of Microbiology, University Hospital of Monastir, Monastir, Tunisia. [24]Internal Medicine Department, Alex Ekwueme Federal University Teaching Hospital, Abakaliki, Nigeria. [25]Irrua Specialist Teaching Hospital, Irrua, Nigeria. [26]Quadram Institute Bioscience, Norwich, UK. [27]World Health Organization, Africa Region, Brazzaville Congo. [28]Centre de Recherche et de Formation en Infectiologie de Guinée (CERFIG), Université de Conakry, Conakry, Guinea. [29]TransVIHMI, Montpellier University/IRD/INSERM, Montpellier, France. [30]Virology Department, Institut Pasteur de Dakar, Dakar, Senegal. [31]Mali-University Clinical Research Center (UCRC), Bamako, Mali. [32]Laboratory of Molecular and Cellular Screening Processes, Centre of Biotechnology of Sfax, University of Sfax, Sfax, Tunisia. [33]Broad Insitute of Harvard and MIT, Cambridge, MA, USA. [34]KEMRI-Wellcome Trust Research Programme/KEMRI-CGMR-C, Kilifi, Kenya. [35]National Institute for Communicable Diseases (NICD) of the National Health Laboratory Service (NHLS), Johannesburg, South Africa. [36]The Nigerian Institute of Medical Research, Yaba, Lagos, Nigeria [37]Institute of Virology, Charité – Universitätsmedizin, Berlin, Germany. [38]Medical Biotechnology Laboratory, Rabat Medical and Pharmacy School, Mohammed V University, Rabat, Morocco. [39]Department of Immunology, University of Maiduguri Teaching Hospital, P.M.B. 1414, Maiduguri, Nigeria. [40]MRC/UVRI and LSHTM Uganda Research Unit, Entebbe, Uganda. [41]Division of Medical Virology, Faculty of Medicine and Health Sciences, Stellenbosch University, Tygerberg, Cape Town, South Africa. [42]The Biotechnology Center of the University of Yaoundé I, Cameroon and CDC Foundation, Yaounde, Cameroon. [43]Department of Ecology and Evolution, University of Chicago, Chicago, IL, USA. [44]Virology Unit, Institut Pasteur de Madagascar, Antananarivo, Madagascar. [45]University of Nebraska Medical Center (UNMC), Omaha, NE, USA. [46]Antibody Immunity Research Unit, School of Pathology, University of the Witwatersrand, Johannesburg, South Africa. [47]CHU de Bouaké, Laboratoire/Unité de Diagnostic des Virus des Fièvres Hémorragiques et Virus Émergents, Bouaké, Côte d'Ivoire. [48]School of Public Health, Pwani University, Kilifi, Kenya. [49]Swiss Tropical and Public Health Institute, Basel, Switzerland. [50]West African Centre for Cell Biology of Infectious Pathogens (WACCBIP), Department of Biochemistry, Cell and Molecular Biology, University of Ghana, Accra, Ghana. [51]School of Life Sciences and Zeeman Institute for Systems Biology and Infectious Disease Epidemiology Research (SBIDER), University of Warwick, Coventry, UK. [52]Uganda Virus Research Institute, Entebbe, Uganda. [53]Division of Virology, National Health Laboratory Service and University of the Free State, Bloemfontein, South Africa. [54]Botswana Harvard AIDS Institute Partnership and Botswana Harvard HIV Reference Laboratory, Gaborone, Botswana. [55]University of Zambia, School of Veterinary Medicine, Department of Disease Control, Lusaka, Zambia. [56]African Society for Laboratory Medicine, Addis Ababa, Ethiopia. [57]Functional Genomic Platform/National Centre for Scientific and Technical Research (CNRST), Rabat, Morocco. [58]Rwanda National Reference Laboratory, Kigali, Rwanda. [59]Robert Koch-Institute, Berlin, Germany. [60]G5 Evolutionary Genomics of RNA Viruses, Institut Pasteur, Paris, France. [61]Research Unit of Autoimmune Diseases UR17DN02, Military Hospital of Tunis, University of Tunis El Manar, Tunis, Tunisia. [62]Department of Public Health, Ministry of Health, Ilorin, Kwara State, Nigeria. [63]Laboratory of Clinical Virology, Institut Pasteur de Tunis, Tunis, Tunisia. [64]Faculty of Pharmacy of Monastir, Monastir, Tunisia. [65]National Microbiology Reference Laboratory, Harare, Zimbabwe. [66]National Influenza Centre, Viral Respiratory Laboratory, Algiers, Algeria. [67]Medical Microbiology and Parasitology Department, College of Medicine, University of Ibadan, Ibadan, Nigeria. [68]Lincoln International Institute for Rural Health, University of Lincoln, Lincoln, UK. [69]Centre for the AIDS Programme of Research in South Africa (CAPRISA), Durban, South Africa. [70]Africa Health Research Institute, KwaZulu-Natal, Durban, South Africa. [71]Department of Biochemistry and Biotechnology, Pwani University, Kilifi, Kenya. [72]National Health Laboratory Service (NHLS), Tygerberg, Cape Town, South Africa. [73]Institution and Department, Ministry Of Health, COVID-19 Testing Laboratory, Mbabane, Kingdom of Eswatini. [74]Laboratory of Health Sciences and Technologies, High Institute of Health Sciences, Hassan 1st University, Settat, Morocco. [75]Department of Emerging Infectious Diseases, Institute of Tropical Medicine, Nagasaki University, Nagasaki, Japan. [76]Central Public Health Laboratories (CPHL), Kampala, Uganda. [77]Faculty of Medicine Ain Shams Research institute (MASRI), Ain Shams University, Cairo, Egypt. [78]Charles Nicolle Hospital, Laboratory of Microbiology, National Influenza Center, 1006 Tunis, Tunisia. [79]Laboratory of Transmissible Diseases and Biological Active Substances (LR99ES27), Faculty of Pharmacy of Monastir, University of Monastir, Monastir, Tunisia. [80]Department of Biochemistry and Molecular Biology, The Institute for Medical Research Israel-Canada, Hadassah Medical School, The Hebrew University of Jerusalem, Jerusalem, Israel. [81]Namibia Institute of Pathology, Windhoek, Namibia. [82]Centre Interdisciplinaires de Recherches Medicales de Franceville (CIRMF), Franceville, Gabon. [83]Department of Biostatistics and Data Science, School of Public Health and Tropical Medicine, Tulane University, New Orleans, LA, USA. [84]Urban Health Collaborative, Dornsife School of Public Health, Drexel University, Philadelphia, PA, USA. [85]UHAS COVID-19 Testing and Research Centre, University of Health and Allied Sciences, Ho, Ghana. [86]Rollins School of Public Health, Emory University, Atlanta, GA, USA. [87]Anoual Laboratory, Casablanca, Morocco. [88]Botswana Institute for Technology Research and Innovation, Gaborone, Botswana. [89]New York University Grossman School of Medicine, New York City, NY, USA. [90]Centre de Recherches Medicales de Lambarene (CERMEL), Lambarene, Gabon. [91]Virology/Molecular Biology Department, Central Health Laboratory, Ministry of Health and Wellness, Mauritius. [92]Center of Scientific Excellence for Influenza Viruses, National Research Centre (NRC), Cairo Egypt. [93]Ministry of Health and Wellness, Gaborone, Botswana. [94]Next Generation Sequencing Unit and Division of Virology, Faculty of Health Sciences, University of the Free State, Bloemfontein 9300, South Africa. [95]National Reference Laboratory Lesotho, Maseru, Lesotho. [96]Centre for Biotechnology Research and Development, Kenya Medical Research Institute, Nairobi, Kenya. [97]Laboratorio de Investigaciones de Baney, Baney, Equatorial Guinea. [98]Ifakara Health Institute, Dar-es-Salaam, Tanzania. [99]Nigeria Centre for Disease Control, Abuja, Nigeria. [100]Department of Medical Diagnostics, Kumasi Centre for Collaborative Research in Tropical Medicine, Kwame Nkrumah University of Science and Technology, Kumasi, Ghana. [101]Department of Medical Laboratory Science, Niger Delta University, Bayelsa State, Nigeria. [102]Systems and Biomedical Engineering Department, Faculty of Engineering, Cairo University, Cairo 12613, Egypt. [103]King Faisal Specialist Hospital and Research Center, Riyadh, Kingdom of Saudi Arabia. [104]Biological Prevention Department, Main Chemical Laboratories, Egypt Army, Cairo, Egypt. [105]PATH, Lusaka, Zambia. [106]Department of Biotechnology, High Institute of Biotechnology of Sidi Thabet, University of Manouba, BP-66, 2020 Ariana-Tunis, Tunisia. [107]Genomic Center for Human Pathologies (GENOPATH), Faculty of Medicine and Pharmacy, Mohammed V University, Rabat, Morocco. [108]Rwanda National Joint Task Force COVID-19, Rwanda Biomedical Centre, Ministry of Health, Kigali, Rwanda. [109]School of Health Sciences, College of Medicine and Health Sciences, University of Rwanda, Kigali, Rwanda. [110]Instituto Nacional de Saude (INS), Maputo, Mozambique. [111]National Health Laboratory Service (NHLS), Cape Town, South Africa. [112]Computational Biology Division, Department of Integrative Biomedical Sciences, IDM, CIDRI Africa Wellcome Trust Centre, University of Cape Town, Cape Town, South Africa. [113]Virology Laboratory, Alex Ekwueme Federal University Teaching Hospital, Abakaliki, Nigeria. [114]Department of Epidemiology and Community Health, Faculty of Clinical Sciences, College of Health Sciences, University of Ilorin, Ilorin, Kwara State, Nigeria. [115]Alex Ekwueme Federal University Teaching Hospital, Abakaliki, Nigeria. [116]Mayotte Hospital Center, Mayotte, France. [117]Virology Service, Centre Pasteur of Cameroun, Yaounde, Cameroon. [118]Kumasi Centre for Collaborative Research in Tropical Medicine, Kwame Nkrumah University of Science and Technology, Kumasi, Ghana. [119]Laboratoire de Recherche et d'Analyses Médicales de la Gendarmerie Royale, Rabat, Morocco. [120]Clinical and Experimental Pharmacology Lab, LR16SP02, National Center of Pharmacovigilance, University of Tunis El Manar, Tunis, Tunisia. [121]CHU Hedi Chaker Sfax, Service de Pneumologie, Tunis, Tunisia. [122]Laboratoire de Recherche et d'Analyses Médicales de la Gendarmerie Royale, Rabat, Morocco. [123]Central Public Health Laboratories (CPHL), Cairo, Egypt. [124]Centre MURAZ, Ouagadougou, Burkina Faso. [125]National Institute of Public Health of Burkina Faso (INSP/BF), Ouagadougou, Burkina Faso. [126]National Reference Center for Respiratory Viruses, Molecular Genetics of RNA Viruses, UMR 3569 CNRS, University of Paris, Institut Pasteur, Paris, France. [127]Sub-Saharan African Network For TB/HIV Research Excellence (SANTHE), Durban, South Africa. [128]Coordenação Geral de Laboratórios de Saúde Pública/Secretaria de Vigilância em Saúde, Ministério da Saúde, Brasília, Distrito Federal, Brazil. [129]World Health Organization, WHO Lesotho, Maseru, Lesotho. [130]Med24 Medical Centre, Ruwa, Zimbabwe. [131]Division of Human Genetics, Department of Pathology, University of Cape Town, Cape Town, South Africa. [132]Center for Human Genetics, College of Medicine and Health Sciences, University of Rwanda, Kigali, Rwanda. [133]Laboratory of Human Genetics, GIGA Research Institute, Liège, Belgium. [134]National Health Laboratory, Gaborone, Botswana. [135]MRC-University of Glasgow Centre for Virus Research, Glasgow, UK.[136]Harvard T.H. Chan School of Public Health, Boston, MA, USA. [137]Department of Global Health, University of Washington, Seattle, WA, USA. [138]Centre for Human Virology and Genomics, Nigerian Institute of Medical Research, Yaba, Lagos, Nigeria. [139]School of Pathology, Faculty of Health Science, University of the Witwatersrand, Johannesburg, South Africa.

12 May 2021; accepted 3 September 2021
Published online 9 September 2021
10.1126/science.abj4336

Bridging Chapters 4 and 5

Chapter 4 shows how genomic surveillance helped to identify VOIs and VOCs from travelers in Africa. This is essential, as patients can then be isolated, which may help prevent transmission chains from occurring within communities. SARS-CoV-2 sequencing allows for the rapid identification of variants within communities. Chapter 5 shows how the sequencing of SARS-CoV-2 assisted in identifying the Delta variant in South Africa and is a preprint that can be obtained from medrxiv.org. In this study, I assisted in SARS-CoV-2 data production, sequence assembly, genome analysis, and data curation.

CHAPTER 5

RAPID REPLACEMENT OF THE BETA VARIANT BY THE DELTA VARIANT IN SOUTH AFRICA

Rapid replacement of the Beta variant by the Delta variant in South Africa

Authors: Houriiyah Tegally[1,2#], Eduan Wilkinson[1,2#], Christian L. Althaus[3], Marta Giovanetti[4,5], James Emmanuel San[2], Jennifer Giandhari[2], Sureshnee Pillay[2], Yeshnee Naidoo[2], Upasana Ramphal[2], Nokukhanya Msomi[6], Koleka Mlisana[7,22], Daniel G. Amoako[8], Josie Everatt[8], Thabo Mohale[8], Anele Nguni[8], Boitshoko Mahlangu[8], Noxolo Ntuli[8], Zamantungwa T. Khumalo[8], Zinhle Makatini[18,20], Nicole Wolter[8], Cathrine Scheepers[8,9], Arshad Ismail[8], Deelan Doolabh[10], Rageema Joseph[10], Amy Strydom[11], Adriano Mendes[11], Michaela Davis[11], Simnikiwe H. Mayaphi[11], Yajna Ramphal[1], Arisha Maharaj[1], Wasim Abdool Karim[2], Derek Tshiabuila[2], Ugochukwu J. Anyaneji[2], Lavanya Singh[2], Susan Engelbrecht[12], Vagner Fonseca[2,5], Kruger Marais[13], Stephen Korsman[13], Diana Hardie[13], Nei-yuan Hsiao[13], Tongai Maponga[12], Gert van Zyl[12], Gert Marais[13], Arash Iranzadeh[14,15], Darren Martin[14,15], Luiz Carlos Junior Alcantara[4,5], Phillip Armand Bester[16], Martin M. Nyaga[17], Kathleen Subramoney[18], Florette K. Treurnicht[18,20], Marietjie Venter[11], Dominique Goedhals[19], Wolfgang Preiser[12], Jinal N. Bhiman[8,20], Anne von Gottberg[8,20,21], Carolyn Williamson[13,15,22], Richard J. Lessells[2,22]*, and Tulio de Oliveira[1,2,22,23]*¶

[#]These authors jointly contributed to this work

*These authors jointly supervised this work

[¶]Corresponding Author: Tulio de Oliveira, tulio@sun.ac.za

Affiliations

[1]Centre for Epidemic Response and Innovation (CERI), School of Data Science and Computational Thinking, Stellenbosch University; Stellenbosch, South Africa

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

[3]Institute of Social and Preventive Medicine, University of Bern, Bern, Switzerland. Funding: CA received funding from the European Union's Horizon 2020 research and innovation programme - project EpiPose (No 101003688).

[4]Laboratorio de Flavivirus, Fundacao Oswaldo Cruz, Rio de Janeiro, Brazil

[5]Laboratório de Genética Celular e Molecular, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil

[6]Discipline of Virology, School of Laboratory Medicine and Medical Sciences and National Health Laboratory Service (NHLS), University of KwaZulu–Natal, Durban, South Africa

[7]NHLS, Johannesburg, South Africa

[8]National Institute for Communicable Diseases (NICD) of the National Health Laboratory Service (NHLS), Johannesburg, South Africa

[9]SA MRC Antibody Immunity Research Unit, School of Pathology, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa

[10]Wellcome CIDRI-Africa and Institute of Infectious Disease and Molecular Medicion, Division of Medical Virology, Department of Pathology, University of Cape Town and National Health Laboratory Service

[11] Department of Medical Virology, University of Pretoria and NHLS Tshwane Academic division, Pretoria, South Africa.

[12]Division of Medical Virology, Faculty of Medicine and Health Sciences, Stellenbosch University and NHLS Tygerberg Hospital, Cape Town, South Africa

[13]Division of Medical Virology, NHLS Groote Schuur Hospital, University of Cape Town, Cape Town, South Africa

[14]Computational Biology Division, Department of Integrative Biomedical Sciences, University of Cape Town, Cape Town, South Africa

[15]Division of Medical Virology, Institute of Infectious Disease and Molecular Medicine, University of Cape Town, Cape Town, South Africa

[16]Division of Virology, National Health Laboratory Service and University of the Free State, Bloemfontein, South Africa

[17]Next Generation Sequencing Unit and Division of Virology, University of the Free State, Bloemfontein, South Africa

[18]Department of Virology, National Health Laboratory Service, Charlotte Maxeke Johannesburg Academic Hospital, Johannesburg, South Africa

[19]PathCare Vermaak, Pretoria, South Africa and Division of Virology, University of the Free State, Bloemfontein, South Africa

[20]School of Pathology, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa

[21]Department of Pathology, Faculty of Health Sciences, University of Cape Town, Cape Town

[22]Centre for the AIDS Programme of Research in South Africa (CAPRISA), Durban, South Africa

[23]Department of Global Health, University of Washington, Seattle, WA, USA

Abstract

The Beta variant of the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) emerged in South Africa in late 2020 and rapidly became the dominant variant, causing over 95% of infections in the country during and after the second epidemic wave. Here we show rapid replacement of the Beta variant by the Delta variant, a highly transmissible variant of concern (VOC) that emerged in India and subsequently spread around the world. The Delta variant was imported to South Africa primarily from India, spread rapidly in large monophyletic clusters to all provinces, and became dominant within three months of introduction. This was associated with a resurgence in community transmission, leading to a third wave which was associated with a high number of deaths. We estimated a growth advantage for the Delta variant in South Africa of 0.089 (95% confidence interval [CI] 0.084-0.093) per day which

corresponds to a transmission advantage of 46% (95% CI 44-48) compared to the Beta variant. These data provide additional support for the increased transmissibility of the Delta variant relative to other VOC and highlight how dynamic shifts in the distribution of variants contribute to the ongoing public health threat.

Main text

By the beginning of 2021, the emergence and international spread of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) variants of concern (VOC) with increased transmissibility or partial immune evasion properties completely altered the dynamics of the coronavirus disease 2019 (COVID-19) pandemic. The detection of Alpha, Beta and Gamma variants in the UK, SA and Brazil respectively resulted in a renewed worldwide race to contain new waves of infection and mortality and re-affirmed the need for global access to vaccines[1–3]. In these VOC, the accumulation of amino acid changes in the spike glycoprotein and other viral proteins have significant phenotypic impact, most notably increased transmissibility[4]. The Beta variant was also associated with resistance to neutralizing antibodies, contributing to reduced vaccine protection against symptomatic COVID-19[5–7].

On 11 May 2021, the World Health Organization (WHO) designated a fourth VOC as Delta (initial Pango lineage B.1.617.2). The Delta variant was first sampled in October 2020 in India[8] and was later associated with a massive resurgence of infections throughout that country from March 2021[9]. The Delta variant has since spread globally (sampled in 180 countries as of 14 September 2021), where in many cases it has competitively replaced previously circulating variants. This has been demonstrated most clearly with the replacement of the previously dominant Alpha variant in countries such as the United Kingdom[10]. Here, we describe in the context of South Africa how the Delta variant rapidly replaced the previously dominant Beta variant, fueling a third wave in South Africa associated with a high number of cases and deaths.

The South African epidemic has been characterized by three distinct waves, causing almost 250 000 excess deaths by the end of August 2021, of which 85-95% are estimated to be COVID-19 deaths[11–13].

The first wave from March 2020 to September 2020 was characterized by multiple co-circulating SARS-CoV-2 lineages[14], whereas the second wave between November 2020 and February 2021 was dominated by the Beta variant[1]. Initial estimates suggested that the Beta variant was associated with a transmission advantage of 23-50% compared to ancestral lineages, although there remains uncertainty around whether this was predominantly related to an inherent increase in the transmissibility of the Beta variant, or partly to reduced cross-protection from natural immunity following prior infection with a different lineage[4,15,16]. After the second wave, estimates of seroprevalence from blood donor surveys ranged from 32% to 62% across the nine provinces of South Africa, with a weighted national estimate of 47%[17]. Between March and May 2021, restrictions were eased (adjusted level 1 lockdown[18]), daily case counts stabilized at below 1500 per day and the effective reproduction number (R_e) hovered around 1[19]. The national vaccination rollout in South Africa only began on 17 May 2021 (around 480 000 health care workers had been vaccinated before then as part of a phase 3b clinical trial[20]). Starting in May, with minimal restrictions still in place (adjusted level 1 lockdown[18]), a resurgence in cases began in the inland provinces such as the Northern Cape, Free State and, Gauteng, which was followed by all other provinces. Despite the re-introduction of higher-level restrictions at the end of May[18], incidence continued to increase, peaking at around 20 000 cases per day in early July (Fig. 1a). By the end of August, there had been over 90 000 excess deaths in the third wave, similar to the number observed in the second wave[11] (Extended Data Fig. 1).

Figure 1b shows the distribution of variants over time at a national level based on genomic surveillance throughout the second and third waves. Beta remained dominant from November 2020 through the second wave and the post-wave period of lower-level transmission. We detected several additional variants of interest and variants of concern sporadically during this post wave period, including the Alpha variant from January 2021 onwards, but it remained at a low frequency nationally and was not associated with a significant resurgence of cases in any province (Fig 1c). In April 2021, through intensified sampling of cases associated with recent travel from India, we started to detect the Delta variant.

However, retrospective sequencing of samples collected earlier in the Northern Cape demonstrated the presence of Delta in that province from as early as 10 March 2021, although Beta remained dominant as the incidence of SARS-CoV-2 increased in that province in April and May (Fig 1c). Through May and June 2021, Delta rapidly displaced Beta and this shift was associated with a rapid increase in SARS-CoV-2 incidence (Fig. 1a, b). The dominance of Delta was consistently observed through increased genomic surveillance during the third wave, with the detection of Beta drastically decreasing to almost none in the last weeks (Extended Data Fig. 2a). We estimated that Delta has a growth advantage of 0.089 (95% confidence interval [CI] 0.084-0.093) per day compared to Beta (Fig. 1d). Assuming the same generation time of 5.2 days for both variants[21], this corresponds to a growth advantage of 46% (95% CI 44-48%) per generation of viral transmission. Even though Delta has been shown to exhibit earlier viral growth and high viral load in vivo compared to earlier variants[22], the serial interval of Delta (and consequently the viral generation time) does not seem to differ substantially[23]. However, the degree to which the growth advantage of Delta is mediated by inherent increased transmissibility and/or immune evasion cannot be determined[16].

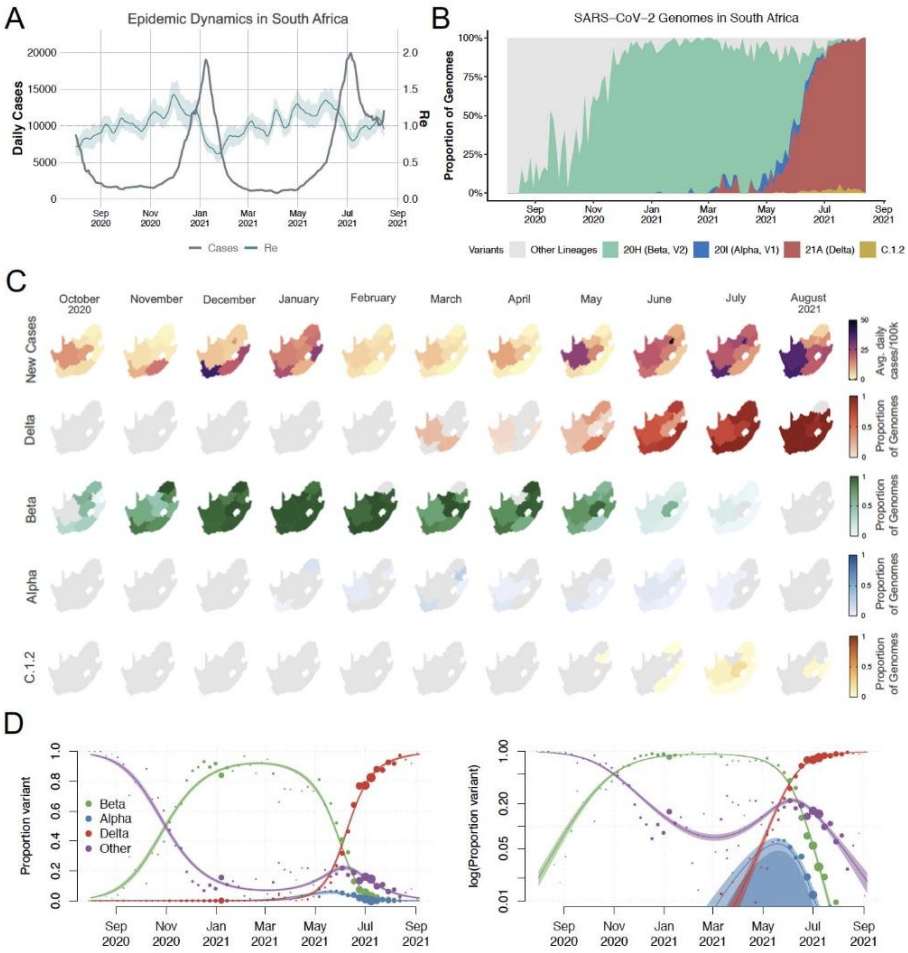


Figure 1 – Replacement of Beta by Delta in South Africa during two distinct epidemic waves. A) Dynamics of the SARS-CoV-2 epidemic in South Africa showing the number of daily COVID-19 cases and estimates of the effective reproduction number (R_e). B) Progression in the proportion of circulating variants in South Africa over the second and third waves of infection, showing the rapid replacement of Beta by Delta. C) Prevalence maps following the progression in the monthly average of daily number of

cases and proportions of relevant variants per province in South Africa from October 2020 to August 2021. D) Modeled proportion of SARS-CoV-2 variants over time in South Africa in linear (left) scale and logarithmic (right) scales, showing that Beta became the dominant variant in South Africa by the end of 2020 and was rapidly outcompeted by Delta from May to August 2021. Model fits are based on a multinomial logistic regression with splines. The size of dots corresponds to the weekly sample size.

Phylogenetic analysis of 5602 Delta genomes from South Africa (sampled between 10 March and 20 August 2021) against a globally representative set of other Delta genomes (n=4983) revealed at least 72 introductions of the Delta variant into South Africa (Fig 2b), with just under half (43%) originating from India (Extended Data Fig. 4). Between January and May 2021, India was in the top five leading countries in terms of overseas international travelers entering South Africa, with an average of 1000 travelers entering per month [24]. Following introductions into South Africa, the Delta variant appears to have spread in several monophyletic clusters spanning multiple provinces (Fig. 2a and Extended Data Fig. 3a), including one particularly large transmission cluster at the top of the tree consisting of 2680 sequences (Fig. 2a). In fact, cluster analysis revealed that 38.5% (2161/5602) of Delta genomes from South Africa belonged to monophyletic clusters of five or more sequences (Extended Data Fig. 3a). Phylogeographic reconstruction of the largest cluster (Fig. 2a, Cluster A) shows its origin in the Gauteng province dated back to 8 May 2021 (95% highest posterior density ranging from 1-12 May 2021) and followed by dissemination to all provinces (Fig. 2d). This illustrates nationwide monophyletic transmission of the Delta variant, characteristic of high transmissibility. Delta sequences in South Africa comprised various sub-lineages with the most represented ones being B.1.617.2, AY.4 and AY.12 (Extended Data Fig. 2b).

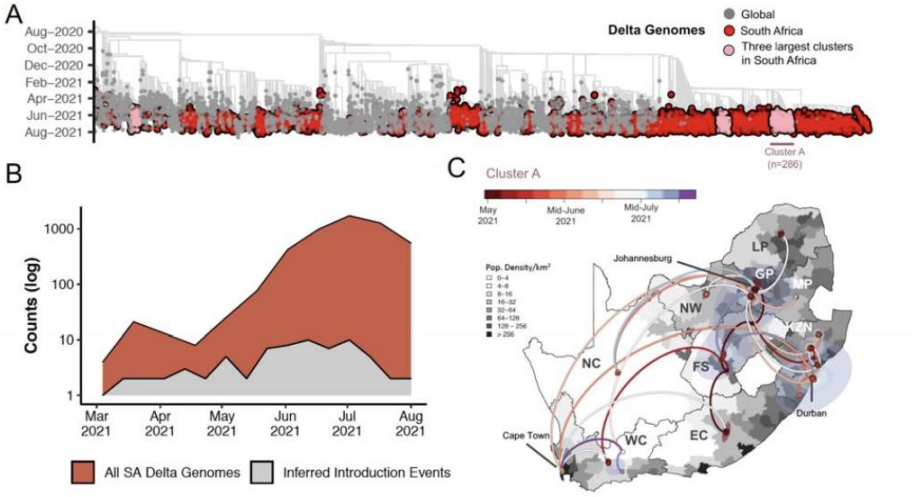


Figure 2: Phylogenetic reconstruction of the Delta variant in South Africa. A) Time-resolved maximum clade credibility phylogeny of 5602 SARS-CoV-2 Delta sequences from South Africa (red, pink) along with 4983 global Delta sequence (grey). The largest identified monophyletic Delta cluster (n=286) is highlighted in pink (Cluster A). B) Inference of viral introduction events from ancestral state reconstruction mapped to the number of Delta genomes sequenced in South Africa. C) Spatiotemporal reconstruction of the spread of Delta cluster A in South Africa. Circles represent nodes of the maximum clade credibility phylogeny, coloured according to their inferred time of occurrence (colour scale in top-left). Shaded areas around the nodes represent the 80% highest posterior density interval and depict the uncertainty of the phylogeographic estimates for each node. Solid curved lines denote the links between nodes and the directionality of movement anticlockwise along the curve. The map additionally shows population density per square kilometer in South Africa.

These findings provide more support that, at this stage in the pandemic, the inherent increased transmissibility of the Delta variant compared to ancestral strains and other VOC, gives it a transmission

fitness advantage over other VOC. Although genomic surveillance in Africa remains heterogeneous[25,26], there is good evidence that Delta has also rapidly replaced Beta in many other African countries[27]. This has also been observed in other countries outside Africa such as Bangladesh, where Beta dominated a second wave and then Delta replaced Beta, driving a more severe third wave[27]. This also suggests that regions where other VOC or variants of interest (VOI) have dominated, especially South America with the dominance of Gamma and Lambda, could remain susceptible to further resurgences driven by the Delta variant.

Whether partial immune evasion also contributed to the transmission fitness advantage of Delta is difficult to establish at the moment. We previously showed that infection with the Beta variant elicited humoral immune responses that offered strong neutralizing antibody cross-protection against ancestral lineages and some other VOC, but this was prior to the emergence of the Delta variant[5,28]. Beta and Delta are antigenically distinct[29], and it remains possible that, in the context of the South African epidemic, reduced cross-protection may have contributed to the transmission fitness advantage of Delta.

In conclusion, the Delta variant rapidly replaced the Beta variant in South Africa and fueled a third wave throughout the country which was associated with a high number of deaths. This again highlights the importance of strengthening genomic surveillance to support the ongoing pandemic response. In this phase of the epidemic, as population immunity will have increased further (from a mix of natural infection and vaccination), we are now closely monitoring for the emergence of variants that combine partial immune evasion with enhanced transmissibility[30]. To this end, we recently reported the emergence and functional characterization one such lineage (C.1.2) and we are monitoring this closely to assess how it competes with the dominant Delta variant[31]. The continued evolution of SARS-CoV-2 is a reminder of the ongoing public health threat and highlights the importance of addressing vaccine inequity and accelerating vaccine delivery in all parts of the world.

Data availability

All of the SARS-CoV-2 Delta genomes generated by NGS-SA and presented in this article are publicly accessible through the GISAID platform (https://www.gisaid.org/). The GISAID accession identifiers of the Delta sequences analysed in this study are provided as part of Extended Data Table 1. Other raw data for this study are provided as a supplementary dataset at https://github.com/krisp-kwazulu-natal/SARSCoV2_Delta_South_Africa. The reference SARS-CoV-2 genome (MN908947.3) was downloaded from the NCBI database (https://www.ncbi.nlm.nih.gov/).

Code availability

All custom scripts to reproduce the analyses and figures presented in this article are available at https://github.com/krisp-kwazulu-natal/SARSCoV2_Delta_South_Africa.

Acknowledgements

We thank the global laboratories that generated and made public the SARS-CoV-2 sequences (through GISAID) used as reference dataset in this study (a complete list of individual contributors of sequences is provided in Extended Data Table 1). This research reported in this publication was supported by the Strategic Health Innovation Partnerships Unit of the South African Medical Research Council, with funds received from the South African Department of Science and Innovation (DSI). Genomics Surveillance in South Africa was supported in part through National Institutes of Health USA grant U01 AI151698 for the United World Antiviral Research Network (UWARN) and by the Rockefeller Foundation (Prof. Tulio de Oliveira and Dr. Eduan Wilkinson). CERI and KRISP have received donations from Chan Soon-Shiong Family Foundation (CSSFF) and Illumina . CW is funded by the South African MRC; Wellcome Trust (2222574/Z/21/Z) and the EDCTP (RADIATES Consortium; RIA2020EF-3030)

Author Contributions

Produced SARS-CoV-2 genomic data: J.G., S.P., Y.N., U.R., H.T., J.E.S, D.G.A, J.E., T.M., A.N., B.M., N.N., Z.T.K., Z.M., N.W., C.S., A.Ismail, D.D., R.J., A.S., A.M., M.D., S.H.M., Y.R., A.M., W.A.K., D.T., U.J.A, L.S., S.E., T.M., G.v.Z, G.M., A.Iranzadeh, P.A.B., M.M.N., K.S.

Collected samples and curated metadata: N.M., K.Mlisana, K.Marais, S.K., D.H., N-y.H.,

Analysed the data: H.T., E.W., C.L.A., M.G., J.E.S, V.F., R.J.L, and T.dO.

Helped with study design and data interpretation. H.T., E.W., D.M., L.C.J.A, F.T., M.V., D.G., W.P., J.N.B., A.v.G., C.W., R.J.L and T.dO

Wrote the initial manuscript, which was reviewed by all authors. H.T., E.W., J.E.S, R.J.L, and T.dO

References

1. Tegally, H. *et al*. Detection of a SARS-CoV-2 variant of concern in South Africa. *Nature* **592**, 438–443 (2021).
2. Faria, N. R. *et al*. Genomics and epidemiology of the P.1 SARS-CoV-2 lineage in Manaus, Brazil. *Science* **372**, 815–821 (2021).
3. Meng, B. *et al*. Recurrent emergence of SARS-CoV-2 spike deletion H69/V70 and its role in the Alpha variant B.1.1.7. *Cell Rep*. **35**, 109292 (2021).
4. Campbell, F. *et al*. Increased transmissibility and global spread of SARS-CoV-2 variants of concern as at June 2021. *Euro Surveill*. **26**, (2021).
5. Cele, S. *et al*. Escape of SARS-CoV-2 501Y.V2 from neutralization by convalescent plasma. *Nature* **593**, 142–146 (2021).

6. Wibmer, C. K. *et al*. SARS-CoV-2 501Y.V2 escapes neutralization by South African COVID-19 donor plasma. *Nat. Med.* **27**, 622–625 (2021).
7. Madhi, S. A. *et al*. Efficacy of the ChAdOx1 nCoV-19 Covid-19 Vaccine against the B.1.351 Variant. *N. Engl. J. Med.* **384**, 1885–1898 (2021).
8. Cherian, S. *et al*. Convergent evolution of SARS-CoV-2 spike mutations, L452R, E484Q and P681R, in the second wave of COVID-19 in Maharashtra, India. *BioRxiv* (2021) doi:10.1101/2021.04.22.440932.
9. India: WHO Coronavirus Disease (COVID-19) Dashboard With Vaccination Data | WHO Coronavirus (COVID-19) Dashboard With Vaccination Data. https://covid19.who.int/region/searo/country/in.
10. Mishra, S. *et al*. Changing composition of SARS-CoV-2 lineages and rise of Delta variant in England. *EClinicalMedicine* **39**, 101064 (2021).
11. Bradshaw D, Laubscher R, Dorrington R, Groenewald P, Moultrie T. Report on Weekly Deaths in South Africa | South African Medical Research Council. https://www.samrc.ac.za/reports/report-weekly-deaths-south-africa.
12. Dorrington, R. E., Moultrie, T. A., Laubscher, R., Groenewald, P. J. & Bradshaw, D. Rapid mortality surveillance using a national population register to monitor excess deaths during SARS-CoV-2 pandemic in South Africa. *Genus* **77**, 19 (2021).
13. Bradshaw, D., Dorrington, R. E., Laubscher, R., Moultrie, T. A. & Groenewald, P. Tracking mortality in near to real time provides essential information about the impact of the COVID-19 pandemic in South Africa in 2020. *South African Medical Journal* (2021).
14. Tegally, H. *et al*. Sixteen novel lineages of SARS-CoV-2 in South Africa. *Nat. Med.* **27**, 440–446 (2021).
15. Pearson, C. A., Russell, T. W., Davies, N. & Kucharski, A. J. Estimates of severity and transmissibility of novel SARS-CoV-2 variant 501Y.V2 in South Africa | CMMID Repository. https://cmmid.github.io/topics/covid19/sa-novel-variant.html (2021).

16. Althaus, C. L. *et al*. A tale of two variants: Spread of SARS-CoV-2 variants Alpha in Geneva, Switzerland, and Beta in South Africa. *medRxiv* (2021) doi:10.1101/2021.06.10.21258468.
17. Vermeulen, M. *et al*. Prevalence of anti-SARS-CoV-2 antibodies among blood donors in South Africa during the period January-May 2021. *Res. Sq*. (2021) doi:10.21203/rs.3.rs-690372/v1.
18. COVID-19 Risk Adjusted Strategy - SA Corona Virus Online Portal. https://sacoronavirus.co.za/covid-19-risk-adjusted-strategy/.
19. Huisman, J. S. *et al*. Estimation and worldwide monitoring of the effective reproductive number of SARS-CoV-2. *medRxiv* (2020) doi:10.1101/2020.11.26.20239368.
20. Sisonke (Together): OPEN LABEL TRIAL COVID-19 - Full Text View - ClinicalTrials.gov. https://clinicaltrials.gov/ct2/show/NCT04838795.
21. Ganyani, T. *et al*. Estimating the generation interval for coronavirus disease (COVID-19) based on symptom onset data, March 2020. *Euro Surveill*. **25**, (2020).
22. Li, B. *et al*. Viral infection and transmission in a large well-traced outbreak caused by the Delta SARS-CoV-2 variant. *medRxiv* (2021) doi:10.1101/2021.07.07.21260122.
23. Pung, R., Mak, T. M., CMMID COVID-19 working group, Kucharski, A. J. & Lee, V. J. Serial intervals in SARS-CoV-2 B.1.617.2 variant cases. *Lancet* **398**, 837–838 (2021).
24. Statistics South Africa. Statistical Release P0351: Tourism and Migration [Monthly reports]. . http://www.statssa.gov.za/?page_id=1859.
25. Brito, A. F. *et al*. Global disparities in SARS-CoV-2 genomic surveillance. *medRxiv* (2021) doi:10.1101/2021.08.21.21262393.
26. Wilkinson Eduan *et al*. A year of genomic surveillance reveals how the SARS-CoV-2 pandemic unfolded in Africa. *Science* **0**, eabj4336.
27. Julia L. Mullen, Ginger Tsueng, Alaa Abdel Latif, Manar Alkuzweny, Marco Cano, Emily Haag, Jerry Zhou, Mark Zeller, Emory Hufbauer, Nate Matteson, Kristian G. Andersen, Chunlei Wu, Andrew I. Su, Karthik Gangavarapu, Laura D. Hughes, and the Center for Viral Systems Biology outbreak.info. outbreak.info.

28. Moyo-Gwete, T. *et al*. Cross-Reactive Neutralizing Antibody Responses Elicited by SARS-CoV-2 501Y.V2 (B.1.351). *N. Engl. J. Med*. **384**, 2161–2163 (2021).
29. Liu, C. *et al*. Reduced neutralization of SARS-CoV-2 B.1.617 by vaccine and convalescent serum. *Cell* **184**, 4220-4236.e13 (2021).
30. Bushman, M., Taylor, B. P., Kahn, R., Lipsitch, M. & Hanage, W. P. Population impact of SARS-CoV-2 variants with enhanced transmissibility and/or partial immune escape. *medRxiv* (2021) doi:10.1101/2021.08.26.21262579.
31. Scheepers, C. *et al*. The continuous evolution of SARS-CoV-2 in South Africa: a new lineage with rapid accumulation of mutations of concern and global detection. *medRxiv* (2021) doi:10.1101/2021.08.20.21262342.

Materials and Methods

Ethics statement

Delta SARS-CoV-2 sequences that were used in the present study were generated by the six laboratory hubs in the Network for Genomic Surveillance in South Africa (NGS-SA - https://www.ngs-sa.org/)[32]. The genomic surveillance was approved by the University of KwaZulu–Natal Biomedical Research Ethics Committee (ref. BREC/00001510/2020), the University of the Witwatersrand Human Research Ethics Committee (HREC, ref. M180832, M210159, M210752), Stellenbosch University HREC (ref. N20/04/008_COVID-19), and the University of Cape Town HREC (ref. 383/2020), the University of Pretoria HREC (100/2017), and the University of Free State Health Sciences Research Ethics Committee (ref. UFS-HSD2020/1860/2710). Individual participant consent was not required for genomic surveillance - this requirement was waived by the Research Ethics Committees. Following sequencing, all consensus sequences are uploaded to GISAID[33] and their use is subject to the database terms and conditions.

Epidemiological data

We analyzed daily cases of SARS-CoV-2 in South Africa up to 2nd September 2021, from publicly released data provided by the National Department of Health and the National Institute for Communicable Diseases. This was accessible through the repository of the Data Science for Social Impact Research Group at the University of Pretoria (https://github.com/dsfsi/covid19za)[34,35]. The National Department of Health releases daily updates on the number of confirmed new cases and reported COVID-19 deaths with a breakdown by province. Population size estimated for each province was based on 2021 mid-year population estimates [36]. We retrieved estimates of the effective reproduction number (R_e) of SARS-CoV-2 in South Africa from the 'covid-19-Re' data repository (https://github.com/covid-19-Re/dailyRe-Data), as of 2 September 2021[19]. Data on weekly excess deaths were retrieved from the South African Medical Research Council Burden of Disease Research Unit report on weekly deaths in South Africa[11–13].

Genomic data

NGS-SA partner laboratories primarily use two main sequencing technologies, namely Illumina (Illumina, San Diego, CA) and Oxford Nanopore (Oxford Nanopore Technologies, United Kingdom).

Tiling-based Polymerase Chain Reaction

Ribonucleic acid (RNA) is extracted from residual nasopharyngeal and oropharyngeal specimens of quantitative polymerase chain reaction (qPCR)-confirmed COVID-19 cases using either manual or automated extraction methods. Complementary DNA synthesis was performed on extracted RNA using either SuperScript IV reverse transcriptase (Life Technologies, Carslbad, CA) or LunaScript RT SuperMix Kit (New England Biolabs), followed by gene-specific multiplex PCR, using the ARTIC protocol, as described previously[37,38]. In summary, SARS-CoV-2 whole-genome amplification by multiplex PCR was performed using primers designed on Primal Scheme (http://primal.zibraproject. org/) to generate 400 base pair (bp) amplicons with 70bp overlaps, covering the 30 kilobase SARS-CoV-2 genome. Some labs used the Illumina COVIDSeq Kit (Illumina, San Diego, CA) to generate amplicons as

per the manufacturer's recommendations. PCR products were purified in a 1:1 ratio, using AmpureXP purification beads (Beckman Coulter, High Wycombe, UK), and were quantified using the Qubit double strand DNA (dsDNA) High Sensitivity assay kit on a Qubit fluorometer (Life Technologies, Carslbad, CA).

Library Preparation and Next-Generation Sequencing

Depending on the partner institution, library preparation and sequencing was done either on the Illumina or Oxford Nanopore Platform.

Illumina Nextera Flex DNA Library Preparation and Sequencing

Depending on the lab, libraries were prepared using either the Nextera DNA Flex Library Prep kit with Nextera CD indexes (Illumina, San Diego) or the Illumina COVIDSeq kit (Illumina, San Diego), according to the manufacturer's instructions. The libraries were then cleaned using 0.9x sample purification beads (Beckman Coulter, Brea, CA) and eluted in 32 μL resuspension buffer. Libraries were quantified by using the Qubit dsDNA High Sensitivity assay kit on a Qubit fluorometer (Life Technologies, Carslbad, CA). Each sample library was normalized to 4 nM concentration, and denatured with 5 μL of 0.2 N NaOH. The final library was diluted to 8 pM and spiked with 1% PhiX Control v3 (Illumina, San Diego) prior to sequencing on an Illumina MiSeq platform (Illumina), using a MiSeq Reagent Kit v2 (500 cycles). All primer sequences are provided in Table S1. This protocol is available at protocols.io coronavirus-method-development community website (https://www.protocols.io/view/illumina-nextera-dna-flex-library-construction-and-bhjgj4jw) since 17 June 2020[39].

Oxford Nanopore Library Preparation and Sequencing

Sequencing libraries were generated from the barcoded products using the Genomic DNA Sequencing Kit SQK-LSK109 (Oxford Nanopore Technologies). Briefly, tiling PCR amplicons were generated as described above. Ligation was carried out using UltraII End Prep Reaction Mix and UltraII End Prep Enzyme Mix and barcoded using the Native Barcoding Kit (Oxford Nanopore Technologies, Oxford, UK). Ninety-six barcodes were used in each run. This included 94 samples and two controls. The libraries were cleaned up using AmpureXP purification beads (Beckman Coulter, High Wycombe, UK) in a 1:1 ratio and eluted in 15μl of elution buffer. Quantification was done using the Qubit dsDNA High Sensitivity assay on the Qubit flourometer (Life Technologies, Carslbad, CA). Sequencing libraries were loaded onto a R9.4 flow cell and data were collected for up to 21 hours.

Genome assembly and Quality control

Sequences generated on the Illumina platform were assembled using Genome Detective 1.132/3 while sequences generated using the Nanopore sequencing technology were assembled with the Artic-nCoV2019 SARS-CoV-2 assembly pipeline (https://github.com/connor-lab/ncov2019-artic-nf).

Estimating relative transmission advantage

We analyzed 16,471 SARS-CoV-2 South African sequences from GISAID (with sample collection dates from 1 August 2020 to 5 September 2021[33]. We used a multinomial logistic regression model to estimate the growth advantage of Delta compared to Beta in South Africa[4,40]. We added splines to account for time-varying growth rates in the model fit and estimated the overall growth advantage of Delta compared to Beta at the time point where the proportion of Delta reached 50%. We fitted the model using the *multinom* function of the *nnet* package[41] and estimated the growth advantage using the package *emmeans*[42] in R.

Phylogenetic analysis

We analyzed 5602 Delta variants from South Africa, publicly available on GISAID[33] as of 2 September 2021, in a phylogenetic context against a globally representative (n=4983) set of other SARS-CoV-2 Delta variants from around the world (selection obtained from Nextstrain publicly available Delta build at the time of analysis: https://nextstrain.org/groups/neherlab/ncov/21A.Delta). The full set of sequences were aligned with NextAlign[43] to obtain a good codon quality alignment against the Wuhan-Hu-1 universal reference. The subsequent alignment was then used to infer a Maximum Likelihood tree topology in IQTREE2[44] (-m GTR, -b 100). Transfer bootstrap support for splits in the topology was inferred using Booster[45]. The resulting consensus ML tree topology was assessed for molecular clock signal in TempEst[46] (Extended Data Fig. 3b). Potential outlier sequences or sequences lacking required metadata (e.g. date and location of sampling) were pruned off the topology with the ape package[47] in R prior to dating. The branches in the ML-tree topology were then converted into units of calendar time in TreeTime[48] using a constant rate of 0.0008 substitutions/site/year with a clock standard deviation of 0.0004 substitutions/site/year. Following the dating of the phylogeny, we annotated the tips and internal nodes using the "mugration" package, an extension of TreeTime and then counted the state changes from one country to another and their inferred time points. This gave us the number and timing of SARS-CoV-2 Delta viral exchanges between South Africa and the rest of the world. To infer some measure of confidence in the time and source of viral transitions, we performed the discrete ancestral state reconstruction on 10 bootstrap replicate trees.

Cluster analysis was performed with the phylotype software package[49] to identify monophyletic clades of South African Delta sequences within the ML-tree. Sequences of the largest monophyletic clade (n=286) were extracted to infer continuous phylogeography histories in BEAST v1.10[50]. Briefly, sequences from the selected cluster were aligned in NextAlign and ML-tree topologies were inferred in IQTREE2, as previously described. The temporal signal of each cluster was assessed in TempEst v1.5.3[46] followed by the removal of potential outlier sequences that may violate the molecular clock assumption. Linear regression of root-to-tip genetic distances against sampling dates indicated that the SARS-CoV-2

sequences in that cluster evolved in a relatively strong clock-like manner (correlation coefficient = 0.37, R2 = 0.13) (Extended Data Fig. 3c). Duplicate Markov Chain Monte Carlo (MCMC) analyses for each cluster were executed in BEAST v1.10.4 for 100 million iterations with sampling every 10000 steps in the chain. Convergence of runs was assessed in Tracer v.1.7.1[51] based on high effective sample sizes and good mixing. Maximum clade credibility trees for each run were summarized using TreeAnnotator after discarding the first 10% of the chain as burn-in. The R package "seraphim"[52] was used to extract and map the spatiotemporal information embedded in the MCC trees.

Additional References

32.Msomi, N., Mlisana, K., de Oliveira, T. & Network for Genomic Surveillance in South Africa writing group. A genomics network established to respond rapidly to public health threats in South Africa. *Lancet Microbe* **1**, e229–e230 (2020).

33.Shu, Y. & McCauley, J. GISAID: Global initiative on sharing all influenza data - from vision to reality. *Euro Surveill.* **22**, 30494 (2017).

34.Marivate, V. *et al*. Coronavirus disease (COVID-19) case data - South Africa. *Zenodo* (2020) doi:10.5281/zenodo.3819126.

35.Marivate, V. & Combrink, H. M. Use of Available Data To Inform The COVID-19 Outbreak in South Africa: A Case Study. *Data Sci. J.* **19**, (2020).

36.Statistics South Africa. Statistical Release P0302: Mid-year population estimates 2021. http://www.statssa.gov.za/publications/P0302/P03022021.pdf.

37.Quick, J. nCoV-2019 sequencing protocol v3 (LoCost). (2020).

38.Pillay, S. *et al*. Whole Genome Sequencing of SARS-CoV-2: Adapting Illumina Protocols for Quick and Accurate Outbreak Investigation during a Pandemic. *Genes (Basel)* **11**, (2020).

39.Pillay, S. Illumina Nextera DNA Flex library construction and sequencing for SARS-CoV-2: Adapting COVID-19 ARTIC protocol. (2020).

40.Davies, N. G. *et al*. Estimated transmissibility and impact of SARS-CoV-2 lineage B.1.1.7 in

England. *Science* **372**, (2021).

41.Venables, W. N. & Ripley, B. D. *Modern Applied Statistics with S (Statistics and Computing)*. 510 (Springer, 2002).

42.Lenth RV. *emmeans: Estimated Marginal Means, aka Least-Squares Means,R package version 1.6.1*. (2021).

43.GitHub - nextstrain/nextclade: Viral genome alignment, mutation calling, clade assignment, quality checks and phylogenetic placement. https://github.com/nextstrain/nextclade.

44.Nguyen, L.-T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).

45.Lemoine, F. *et al*. Renewing Felsenstein's phylogenetic bootstrap in the era of big data. *Nature* **556**, 452–456 (2018).

46.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**, vew007 (2016).

47.Popescu, A.-A., Huber, K. T. & Paradis, E. ape 3.0: New tools for distance-based phylogenetics and evolutionary analysis in R. *Bioinformatics* **28**, 1536–1537 (2012).

48.Sagulenko, P., Puller, V. & Neher, R. A. TreeTime: Maximum-likelihood phylodynamic analysis. *Virus Evol.* **4**, vex042 (2018).

49.Chevenet, F., Jung, M., Peeters, M., de Oliveira, T. & Gascuel, O. Searching for virus phylotypes. *Bioinformatics* **29**, 561–570 (2013).

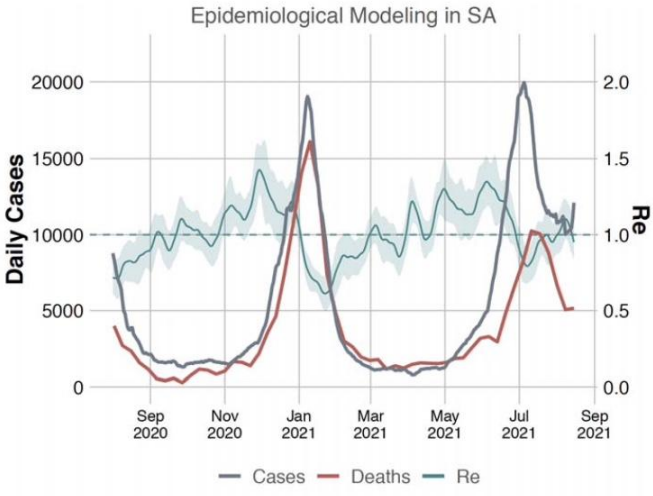
50.Drummond, A. J., Suchard, M. A., Xie, D. & Rambaut, A. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol. Biol. Evol.* **29**, 1969–1973 (2012).

51.Rambaut, A., Drummond, A. J., Xie, D., Baele, G. & Suchard, M. A. Posterior summarization in bayesian phylogenetics using tracer 1.7. *Syst. Biol.* **67**, 901–904 (2018).

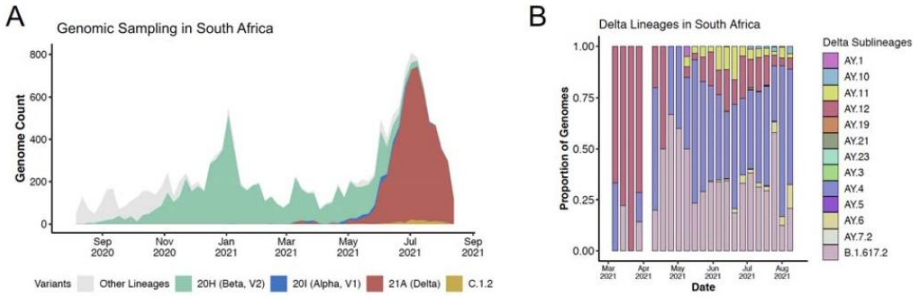
52.Dellicour, S., Rose, R. & Pybus, O. G. Explaining the geographic spread of emerging epidemics: a framework for comparing viral phylogenies and environmental landscape data. *BMC*

Bioinformatics **17**, 82 (2016).

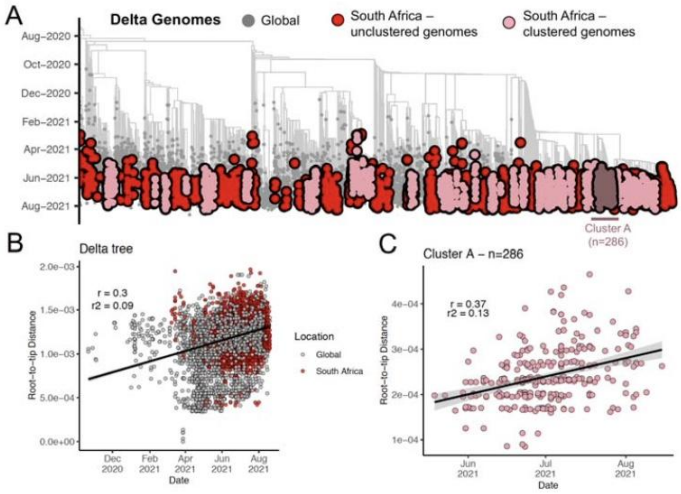
Extended Data Figures



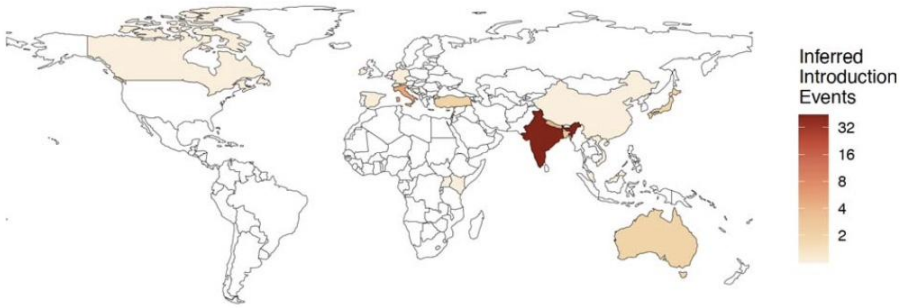
Extended Data Figure 1. Dynamic of the SARS-CoV-2 epidemic in South Africa showing the progression of daily cases (grey), weekly excess deaths (red), and Re value (blue).



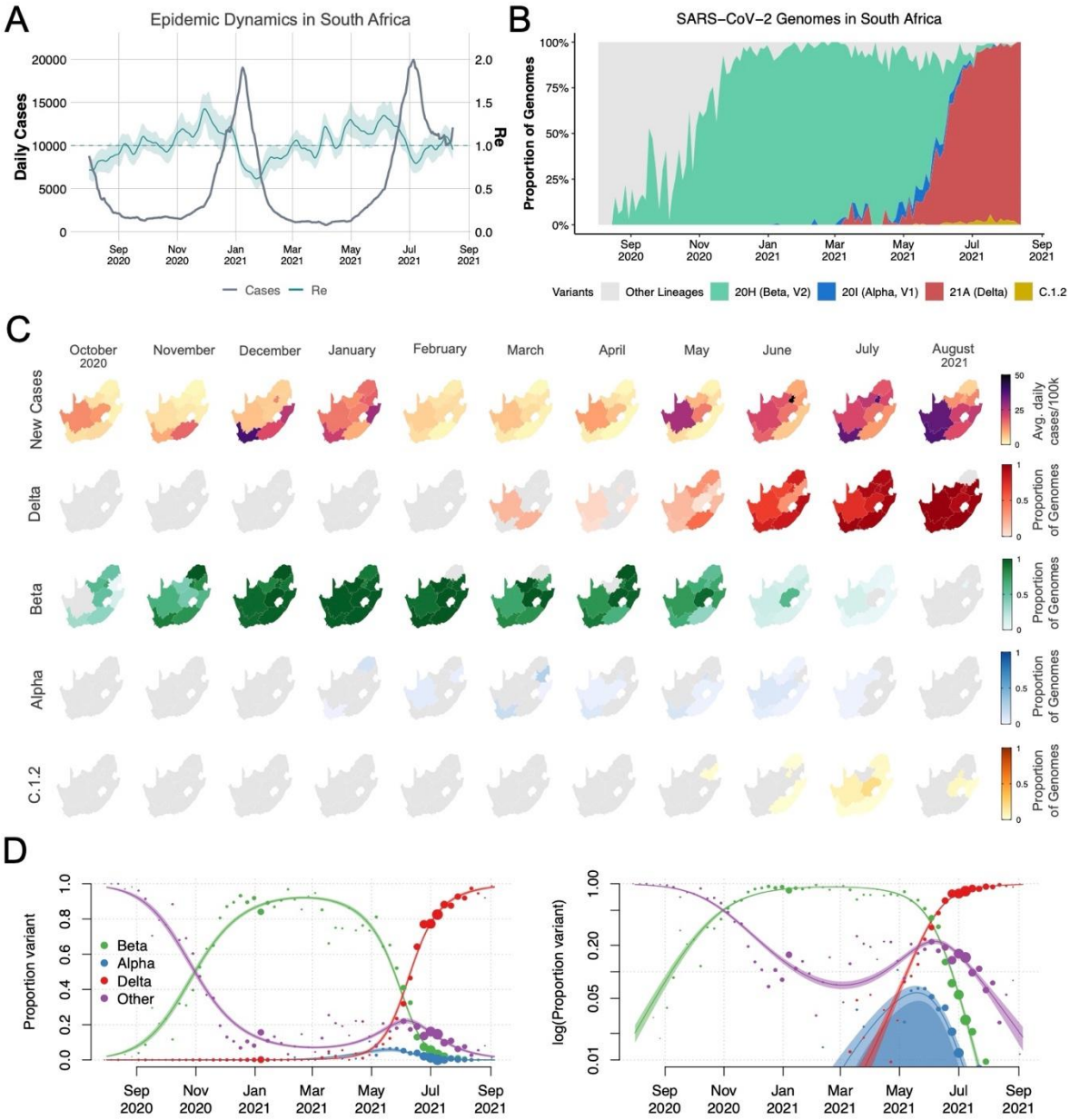
Extended Data Figure 2. Genomic Surveillance in South Africa. A) Number of genomes sampled and sequenced in South Africa from September 2020 to August 2021 classified by variants. B) Proportion of Delta genomes classified in Delta sublineages per time.

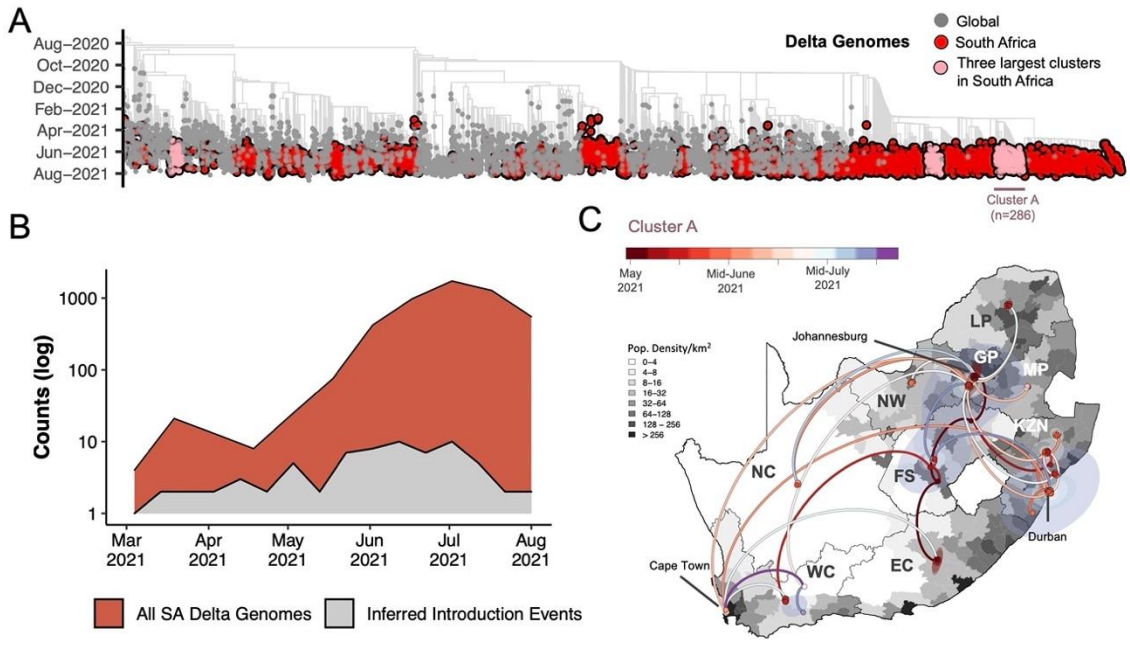


Extended Data Figure 3. Phylogenetic reconstruction of Delta sequences in South Africa. A) Timed Maximum-Likelihood tree of South Africa Delta genomes in the context of global Delta sequences, highlighted by monophyletic clusters identified with Phylotype, including Cluster A, the largest monophyletic cluster identified in South Africa. B) Testing molecular signal of the Delta phylogenetic tree in A in Tempest. C) Testing molecular signal of Cluster A in Tempest.



Extended Data Figure 4. Inferred locations of importations of the Delta variant into South Africa


A
Epidemic Dynamics in South Africa
Daily Cases
Re
Cases
Re
B
SARS−CoV−2 Genomes in South Africa
Proportion of Genomes
Variants
Other Lineages
20H (Beta, V2)
20I (Alpha, V1)
21A (Delta)
C.1.2
C
New Cases
Delta
Beta
Alpha
C.1.2
Avg. daily cases/100k
Proportion of Genomes
D
Proportion variant
Beta
Alpha
Delta
Other
log(Proportion variant)


A
Aug–2020
Oct–2020
Dec–2020
Feb–2021
Apr–2021
Jun–2021
Aug–2021
Delta Genomes
Global
South Africa
Three largest clusters in South Africa
Cluster A (n=286)
B
Counts (log)
1000
100
10
1
Mar 2021
Apr 2021
May 2021
Jun 2021
Jul 2021
Aug 2021
All SA Delta Genomes
Inferred Introduction Events
C
Cluster A
May 2021
Mid-June 2021
Mid-July 2021
Pop. Density/km²
0–4
4–8
8–16
16–32
32–64
64–128
128 – 256
> 256
Johannesburg
LP
GP
MP
NW
KZN
NC
FS
EC
WC
Cape Town
Durban

CHAPTER 6: GENERAL DISCUSSION

Synthesis

COVID-19 is a continuing pandemic that has seen over 440 000 000 confirmed cases globally. The pandemic requires continued surveillance and tests to be conducted repeatedly on a large proportion of the population to detect outbreaks before they begin to spread (Esbin et al., 2020). Rapid, accurate SARS-CoV-2 sequencing may help to identify infected individuals, encourage isolation, and thus reduce the spread of the virus (Bajaj and Purohit, 2020, Maurier et al., 2019).

Although patients may not be showing overt COVID-19 symptoms, testing is a major recommended part of the hospital's infection control policies and require patient biological samples to be provided via appropriate means. Nasopharyngeal specimens are the sample of choice for SARS-CoV-2 testing but nasal and oropharyngeal samples are also accepted (Wölfel et al., 2020). Saliva samples have also been used, but with highly variable results while various studies are evaluating the use of other types of samples such as stool samples (Szymczak et al., 2020). Patients at healthcare facilities have refused to test for SARS-CoV-2 via nasopharyngeal swabs as this collection method has been reported to be very uncomfortable. The less-invasive methods, such as the nasal, throat, saliva, and self-administered swabs have assisted in detecting SARS-CoV-2 but with limitations that have to be addressed (Johnson et al., 2021). Testing of individuals with consistent COVID-19 symptoms would be more cost-effective than restricting testing to individuals with symptoms severe enough to warrant hospitalization. Furthermore, expanding testing to asymptomatic individuals may decrease infections, death, and hospitalization as infected individuals can be identified and isolated thus preventing further spread of the virus (Schuetz et al., 2020).

SARS-CoV-2 respiratory specimens are to be stored between 2 – 8 °C for up to 72 hours after collection and at or below -70 °C if samples are to be transported or there is a delay in analysis (Prevention, 2020). Sequencing may be affected by sample quality. Thus, samples must be sequenced as soon as possible after collection or stored in cold temperatures to maintain RNA integrity (Yilmaz Gulec et al., 2021). Many laboratory settings require SARS-CoV-2 samples to be shipped in or inactivated before sequencing. This, however, is not always possible as some hospitals lack the necessary biosafety conditions. Sample inactivation time, storage temperature, and storage time have led to at least 10.2 % of positive cases being called negative (Prevention, 2020, Yilmaz Gulec et al., 2021).

The complexity and expensive nature of molecular diagnostics and NGS have resulted in backlogs in laboratory facilities as the personnel who are competent in using specialized lab equipment and reagents are required (Esbin et al., 2020). Although diagnostic tests such as RT-PCR are advantageous as they are easy to use and are more tolerant of variable DNA quality than NGS, they are limited by the

multiplex capabilities seen with NGS platforms (1). Several initiatives have been established to help with this matter namely, COVID-19 Genomics UK Consortium (COG-UK), The Indian SARS-CoV-2 Genomics Consortium (INSA-COG), NGS-SA, and SARS-CoV-2 Sequencing for Public Health Emergency Response, Epidemiology, and Surveillance (SPHERES).

Many lineages of SARS-CoV-2 have emerged since the beginning of the COVID-19 pandemic and are still circulating (Tegally et al., 2021a). Sequencing of SARS-CoV-2 has assisted in the identification of new variants due to the mutations that are identified with NGS (Rambaut et al., 2020). SARS-CoV-2 genomic surveillance makes it easy for new variants, causing outbreaks, to be identified and has allowed the development of vaccines and rapid diagnostic methods (Seth-Smith et al., 2019). This information can be used to better understand how this impacts health. The Network for Genomic Surveillance in South Africa (NGS-SA), which includes the National Institute for Communicable Diseases (NICD), KRISP, University of Cape Town, Stellenbosch University, the University of the Free State, the University of Pretoria, the University of the Witwatersrand, and the National Health Laboratory Service (NHLS), continues to monitor and assess the evolution of SARS-CoV-2.

Identification of new variants is important as it helps us determine viral transmissibility and resistance, which are important factors for understanding how to manage the virus (Prevention, 2021). To help prevent the transmission of variants and possible deaths occurring due to infection, real-time polymerase chain reaction (RT-PCR) tests are performed by diagnostic laboratories. RT-PCR is currently a gold standard within diagnostic laboratories for SARS-CoV-2 detection, as it is extremely sensitive and can amplify and identify a single copy of the genomic sequence using PCR. RT-PCR is, however, limited by high costs and the relatively long analysis times that have created bottlenecks within diagnostic laboratories. Furthermore, the genetic diversity of the SARS-CoV-2 virus affects the sensitivity and specificity of detection via PCR as primers specific to all strains need to be constructed to amplify all strains present within the population whilst still excluding all other viruses (Afzal, 2020, Bezier et al., 2020). Sequencing of SARS-CoV-2 overcomes these challenges as NGS technologies use platforms that are capable of sequencing millions of small DNA fragments simultaneously and bioinformatics software is then used to analyze and piece these fragments together by mapping them against a reference genome (Niedringhaus et al., 2011, Rambaut et al., 2020).

LamPORE is a novel diagnostic platform used for the detection of SARS-CoV-2 RNA as it combines loop-mediated isothermal amplification with nanopore sequencing (James et al., 2020). This has the potential to be used to analyze thousands of samples per day on a single instrument thus helping remove the bottlenecks occurring within laboratories. LamPORE has a detection limit of 10 genome copies/µl of extracted RNA, which is higher than the limit achieved by RT-PCR but was not associated with a

significant reduction in clinical samples. LamPORE has a similar performance as RT-PCR and promises high-throughput testing (Peto et al., 2020).

NGS is associated with deep sequencing, which increases the detection of unique variants. NGS also has a high sensitivity for the detection of low-frequency variants, faster turnaround time for high sample volumes, genomic coverage that is comprehensive, lower limit of detection, higher throughput as a result of sample multiplexing, and the possibility of sequencing thousands of genes or gene regions simultaneously (Jamuar et al., 2014, Konig et al., 2015, Rivas et al., 2011, Schuster, 2008, Shendure and Ji, 2008). Illumina sequencing is the most widely used sequencing technology for SARS-CoV-2, as it sequences from both ends resulting in higher coverage, a higher number of reads, and more data when compared to single-end sequencing systems (Ambardar et al., 2016, GISAID, 2022). Furthermore, all genomic changes which result in new variants, such as insertions, deletions, inversion, repetitive sequence elements, new transcripts, and gene fusions, are detected more easily than single-end sequencing technologies, resulting in the identification of new VOIs and VOCs (Ambardar et al., 2016).

Many VOIs and VOCs, such as B.1.351, B.1.525, A.23.1, and C.1.1, spread across Africa as a result of the mobility of infected individuals across countries (Cele et al., 2021, Wibmer et al., 2021). SARS-CoV-2 genomic surveillance has been at the forefront of Africa's response to the COVID-19 pandemic and lack of genomic surveillance may result in Africa becoming a source of new variants. The first wave of the pandemic in Africa saw introductions generally from Europe, but as the pandemic progressed, there was an increase in the spread of the virus within and between African countries (Wilkinson et al., 2021). Phylogenetic analysis of VOIs and VOCs helps to determine how some of the key SARS-CoV-2 variants are spreading within Africa (Rambaut et al., 2020). Using molecular surveillance to monitor pandemics depends on continuous and consistent sampling through time, rapid virus genome sequencing, and rapid reporting. The success of genomic surveillance requires high COVID-19 testing, high sequencing of positive samples within days of sampling, and persistent analyses of the consensus genomes to identify mutational changes (Wilkinson et al., 2021).

Sanger sequencing, with over 99 % accuracy, remains the "gold standard" in clinical and basic research applications and is commonly used to validate gene variants identified using NGS (Slatko et al., 2018). Sanger sequencing is ideal for the sequencing of single genes and single nucleotide variants, targeted sequencing of up to 100 amplicons, sequencing of up to 96 samples without the need for barcoding, and sequencing of regions with high GC content, identification of microbes, microsatellite analysis, and plasmid sequencing. Furthermore, Sanger sequencing does not require expensive equipment, when compared to NGS platforms, and can produce quality data for samples with low viral loads that yield low genome coverage for some NGS technologies. However, NGS technologies are more common in

research laboratories due to their higher throughput capabilities which are necessary when dealing with pandemics (Scientific, 2022).

Illumina Platforms have been used to sequence over 90 % of NGS data worldwide with data outputs varying between 1.2 Gb and 6 Tb. The Ion Torrent technology is available in several models and has a data output of between 30 Mb and 25 Gb per chip and is also commonly used for microbial targeted amplicon and WGS (Loman et al., 2012). Library preparation kits and choice of sequencing platform have an impact on the breadth of genome coverage and accuracy of consensus genomes produced by the two technologies. The cost per sample for high throughput sequencing using the Ion Torrent and Illumina MiSeq are comparable and therefore the two technologies are viable options for genomic sequencing of RNA viruses (Marine et al., 2019). AmpliSeq is a SARS-CoV-2 sequencing workflow from Ion Torrent that is very easily automated with the Ion Chef and S5 instruments and does not require as much training and experience with NGS sample preparation as the Illumina workflow (Plitnick et al., 2021).

Sequencing can be long, laborious, and expensive and laboratories are generally not located at the same site as sample collection sites (Quick, 2020). Portability and cost are therefore important considerations for SARS-CoV-2 sequencing as they determine the number of available sequencing sites and how rapidly they can acquire samples (Cleemput et al., 2020). ONT has developed a range of sequencers capable of producing large amounts of data within relatively short sequencing times. With sequencing platforms such as the flongle and minion, ONT has overcome the portability challenge and has facilitated sequencing mobility allowing for SARS-CoV-2 sequencing to occur in different regions (Nick Loman, 2020). Although the GridION is not classified as a portable sequencer, it is still small enough to be mobile and can be set up at various locations, therefore allowing for rapid, high throughput sequencing at mobile healthcare facilities.

ONT promises relatively quick, real-time, SARS-CoV-2 long-read sequencing that is portable and relatively cheap (Nick Loman, 2020). The quality of the consensus genomes produced, however, has yet to be investigated. Here, we compared the sequences produced by the ONT GridION X5 with those of the Illumina MiSeq, which has currently been the gold standard for SARS-CoV-2 NGS sequencing, as it is a high throughput and high accuracy sequencing platform that allows for high levels of multiplexing (Ambardar et al., 2016, Rambaut et al., 2020). The sequencing platform is, however, large and expensive and is characterized by long sequencing times (Illumina, 2022, Quick, 2020). Consensus genomes produced by both the Illumina MiSeq and ONT GridION X5 were compared by looking at the sequence coverage, the number of mutations, and the type of mutations identified by both platforms. The MiSeq had an overall better sequence quality compared to the GridION, as determined by Nextclade online tool, and had significantly higher sequence coverage. Analysis of sequence coverage

is important as genomes contain genes, noncoding DNA, repetitive sequences, and other elements that may alter the alignment of the sequence to a reference genome (van Dijk et al., 2018).

Sequence coverage is, however, not uniform and may be affected by factors such as sample quality, sample input, homologous regions, regions of low complexity, hypervariable regions, and high GC content (Chiara et al., 2021). The number and type of mutations identified by the MiSeq and the GridION differed significantly and may thus impact the information obtained from genomic surveillance. To correctly identify genetic mutations, a sufficient number of correctly mapped reads mapping a particular region is required (Wei et al., 2011). The MiSeq had an overall higher sequence coverage, which is associated with more accurate identification of variants. The Illumina MiSeq makes use of paired-end sequencing that sequences from both ends and produces more reads than single-end sequencing utilized by the ONT GridION (illumina, 2021). The higher number of reads allows for more accurate identification of mutations for the correct assignment of viral strains. Genomic surveillance may, therefore, be more accurate using Illumina MiSeq sequencing than ONT GridION sequencing.

The nanopore technology is prone to high error rates in basecalling that may result in false-positive results in variant calling and identification of mutations (Wang et al., 2021). This, however, has not stopped researchers from using the technology for SARS-CoV-2 sequencing as 25 % of the viral consensus genome on GISAID was obtained using ONT (Chantal Babb de Villiers, 2021). ONT sequences in real-time and the GridION X5 allows for up to 5 flow cells to be sequenced simultaneously. Thus, a total of 470 samples can be sequenced in 24 hours using the ONT rapid library prep kit compared to the Illumina MiSeq's 96 samples every 36 hrs (Jain et al., 2018, Kono and Arakawa, 2019). The difference in throughput may be a major consideration when deciding on which platform to use for genomic surveillance as a fast turnaround time is required to help laboratories that are generally flooded with samples (Esbin et al., 2020). Sequencing for a longer period and pooling fewer samples per run may increase the quality of consensus genomes produced by the GridION.

ONT has continuously refined the nanopore and motor protein having released eight versions of the system by 2020. These include the R6 (June 2014), R7 (July 2014), R7.3 (October 2014), R9 (May 2016), R9.4 (October 2016), R9.5 (May 2017), R10 (March 2019), and R10.3 (January 2020) (Wang et al., 2021). ONT sequencing applications can be classified into 3 major groups; basic research, clinical usage, and on-site applications. The "on-site applications" include metagenomics and surveillance that are used for strain characterization and rapid microbial and pathogen detection due to their fast, real-time sequencing capabilities and small size (Wang et al., 2021).

Conclusion

High throughput sequencing of SARS-CoV-2, using the ONT GridION, is achievable at a relatively fast rate, as up to 470 samples can be sequenced in approximately 21hrs. As the GridION is relatively small and can be set up within portable laboratories, it can be used to remove the bottlenecks occurring within diagnostic laboratories and thus improve sequencing efficiency. It has been shown that a GridION flowcell can be reused up to 10 times allowing many facilities to perform SARS-CoV-2 sequencing at a relatively low cost (Petersen et al., 2019). Nanopore sequencing, however, still shows lower sequencing coverage and inaccurate identification of mutations when compared to Illumina sequencing and so SARS-CoV-2 genomic surveillance, using the GridION, should be performed with caution.

Recommendation

With the number of COVID-19 infections still rising dangerously, a fast, reliable, and cost-effective sequencing method must be established to assist in rapidly diagnosing patients and identifying transmission chains. This may assist in preventing the further spread of the virus. Sample quality has been shown to affect sequence coverage and thus is an important factor to consider when comparing sequencing technologies. This study was limited by the availability of samples and the quality of collected swab samples. Samples arriving from various institutions in several African countries may have been damaged during transportation and storage. Due to the large number of samples that had to be sequenced and the time required for sequencing, only one run was sequenced on both platforms. Additional analysis should be performed to determine whether the relatively low sequence quality noted from the GridION is due to the quality of the sample or the sequencing technology. Saliva samples have been reported to be more sensitive than nasopharyngeal swabs and are easier to obtain. It is therefore necessary to further investigate whether sequencing of saliva samples on the GridION will result in higher SARS-CoV-2 sequence coverage and better-quality consensus genomes.

REFERENCES

1. Khan S, Siddique R, Shereen MA, Ali A, Liu J, Bai Q, et al. Emergence of a Novel Coronavirus, Severe Acute Respiratory Syndrome Coronavirus 2: Biology and Therapeutic Options. J Clin Microbiol. 2020;58(5).

2. Weston S, Frieman MB. COVID-19: Knowns, Unknowns, and Questions. mSphere. 2020;5(2).

3. Fiorillo L, Cervino G, Matarese M, D'Amico C, Surace G, Paduano V, et al. COVID-19 Surface Persistence: A Recent Data Summary and Its Importance for Medical and Dental Settings. Int J Environ Res Public Health. 2020;17(9).

4. World Health O. WHO Director-General's opening remarks at the media briefing on COVID-19 - 11 March 2020 2020 [Available from: https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020.

5. Seemann T, Lane C, Sherry N, Duchene S, Goncalves da Silva A, Caly L, et al. Tracking the COVID-19 pandemic in Australia using genomics. medRxiv. 2020:2020.05.12.20099929.

6. St Hilaire BG, Durand NC, Mitra N, Pulido SG, Mahajan R, Blackburn A, et al. A rapid, low cost, and highly sensitive SARS-CoV-2 diagnostic based on whole-genome sequencing. bioRxiv. 2020:2020.04.25.061499.

7. Seth-Smith HMB, Bonfiglio F, Cuenod A, Reist J, Egli A, Wuthrich D. Evaluation of Rapid Library Preparation Protocols for Whole Genome Sequencing Based Outbreak Investigation. Front Public Health. 2019;7:241.

8. Resende PC, Motta FC, Roy S, Appolinario L, Fabri A, Xavier J, et al. SARS-CoV-2 genomes recovered by long amplicon tiling multiplex approach using nanopore sequencing and applicable to other sequencing platforms. bioRxiv. 2020:2020.04.30.069039.

9. Liu L, Li Y, Li S, Hu N, He Y, Pong R, et al. Comparison of next-generation sequencing systems. J Biomed Biotechnol. 2012;2012:251364.

10. GISAID. Pandemic coronavirus causing COVID-19 2022 [Available from: https://www.gisaid.org/.

11. Gohl DM, Garbe J, Grady P, Daniel J, Watson RHB, Auch B, et al. A rapid, cost-effective tailed amplicon method for sequencing SARS-CoV-2. BMC Genomics. 2020;21(1):863.

12. Wang M, Fu A, Hu B, Tong Y, Liu R, Gu J, et al. Nanopore target sequencing for accurate and comprehensive detection of SARS-CoV-2 and other respiratory viruses. medRxiv. 2020:2020.03.04.20029538.

13. Quick J. nCoV-2019 sequencing protocol. Protocols io[Google Scholar]. 2020.

14. Pillay S, Giandhari J, Tegally H, Wilkinson E, Chimukangara B, Lessells R, et al. Whole Genome Sequencing of SARS-CoV-2: Adapting Illumina Protocols for Quick and Accurate Outbreak Investigation during a Pandemic. Genes (Basel). 2020;11(8).

15. 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(11):3552-5.
16. Nick Loman WR, Andrew Rambaut. nCoV-2019 novel coronavirus bioinformatics protocol 2020-01-23 [Available from: https://artic.network/ncov-2019/ncov2019-bioinformatics-sop.html.
17. Zhu N, Zhang D, Wang W, Li X, Yang B, Song J, et al. A Novel Coronavirus from Patients with Pneumonia in China, 2019. N Engl J Med. 2020;382(8):727-33.
18. Walls AC, Park Y-J, Tortorici MA, Wall A, McGuire AT, Veesler D. Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. Cell. 2020;181(2):281-92.e6.
19. Scudellari M. The sprint to solve coronavirus protein structures - and disarm them with drugs. Nature. 2020;581(7808):252-5.
20. Kim D, Lee J-Y, Yang J-S, Kim JW, Kim VN, Chang H. The Architecture of SARS-CoV-2 Transcriptome. Cell. 2020;181(4):914-21.e10.
21. Santos IA, Grosche VR, Bergamini FRG, Sabino-Silva R, Jardim ACG. Antivirals Against Coronaviruses: Candidate Drugs for SARS-CoV-2 Treatment? Front Microbiol. 2020;11(1818):1818.
22. Walls AC, Park YJ, Tortorici MA, Wall A, McGuire AT, Veesler D. Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. Cell. 2020;181(2):281-92 e6.
23. Wan Y, Shang J, Graham R, Baric RS, Li F. Receptor Recognition by the Novel Coronavirus from Wuhan: an Analysis Based on Decade-Long Structural Studies of SARS Coronavirus. J Virol. 2020;94(7).
24. Zhang L, Lin D, Sun X, Curth U, Drosten C, Sauerhering L, et al. Crystal structure of SARS-CoV-2 main protease provides a basis for design of improved alpha-ketoamide inhibitors. Science. 2020;368(6489):409-12.
25. Candido DS, Claro IM, de Jesus JG, Souza WM, Moreira FRR, Dellicour S, et al. Evolution and epidemic spread of SARS-CoV-2 in Brazil. Science. 2020;369(6508):1255-60.
26. Laamarti M, Alouane T, Kartti S, Chemao-Elfihri MW, Hakmi M, Essabbar A, et al. Large scale genomic analysis of 3067 SARS-CoV-2 genomes reveals a clonal geo-distribution and a rich genetic variations of hotspots mutations. PLoS One. 2020;15(11):e0240345.
27. Wang C, Liu Z, Chen Z, Huang X, Xu M, He T, et al. The establishment of reference sequence for SARS-CoV-2 and variation analysis. J Med Virol. 2020;92(6):667-74.
28. Rambaut A, Holmes EC, O'Toole A, Hill V, McCrone JT, Ruis C, et al. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. Nat Microbiol. 2020;5(11):1403-7.
29. Tegally H, Wilkinson E, Lessells RJ, Giandhari J, Pillay S, Msomi N, et al. Sixteen novel lineages of SARS-CoV-2 in South Africa. Nat Med. 2021;27(3):440-6.
30. Prevention CfDCa. SARS-CoV-2 Variant Classifications and Definitions 2021 [Available from: https://www.cdc.gov/coronavirus/2019-ncov/variants/variant-info.html.

31. Behjati S, Tarpey PS. What is next generation sequencing? Arch Dis Child Educ Pract Ed. 2013;98(6):236-8.
32. Slatko BE, Gardner AF, Ausubel FM. Overview of Next-Generation Sequencing Technologies. Curr Protoc Mol Biol. 2018;122(1):e59.
33. Heather JM, Chain B. The sequence of sequencers: The history of sequencing DNA. Genomics. 2016;107(1):1-8.
34. World Health O. Public health surveillance for COVID-19: interim guidance, 16 December 2020. Geneva: World Health Organization; 2020 2020. Contract No.: WHO/2019-nCoV/SurveillanceGuidance/2020.8.
35. Xu Y, Lewandowski K, Jeffery K, Downs LO, Foster D, Sanderson ND, et al. Nanopore metagenomic sequencing to investigate nosocomial transmission of human metapneumovirus from a unique genetic group among haematology patients in the United Kingdom. J Infect. 2020;80(5):571-7.
36. Esbin MN, Whitney ON, Chong S, Maurer A, Darzacq X, Tjian R. Overcoming the bottleneck to widespread testing: a rapid review of nucleic acid testing approaches for COVID-19 detection. RNA. 2020;26(7):771-83.
37. Amiel-Tison C. [Some aspects of prevention in perinatology (author's transl)]. J Gynecol Obstet Biol Reprod (Paris). 1978;7(3 Pt 2):596-604.
38. Bajaj A, Purohit HJ. Understanding SARS-CoV-2: Genetic Diversity, Transmission and Cure in Human. Indian Journal of Microbiology. 2020;60(3):398-401.
39. Maurier F, Beury D, Fléchon L, Varré JS, Touzet H, Goffard A, et al. A complete protocol for whole-genome sequencing of virus from clinical samples: Application to coronavirus OC43. Virology. 2019;531:141-8.
40. Tegally H, Wilkinson E, Giovanetti M, Iranzadeh A, Fonseca V, Giandhari J, et al. Detection of a SARS-CoV-2 variant of concern in South Africa. Nature. 2021;592(7854):438-43.
41. Morel B, Barbera P, Czech L, Bettisworth B, Hübner L, Lutteropp S, et al. Phylogenetic analysis of SARS-CoV-2 data is difficult. bioRxiv. 2020:2020.08.05.239046.
42. Niedringhaus TP, Milanova D, Kerby MB, Snyder MP, Barron AE. Landscape of next-generation sequencing technologies. Anal Chem. 2011;83(12):4327-41.
43. Bustin SA, Nolan T. RT-qPCR Testing of SARS-CoV-2: A Primer. Int J Mol Sci. 2020;21(8):3004.
44. Afzal A. Molecular diagnostic technologies for COVID-19: Limitations and challenges. J Adv Res. 2020;26:149-59.
45. Bezier C, Anthoine G, Charki A. Reliability of real-time RT-PCR tests to detect SARS-Cov-2: A literature review. Int J Metrol Qual Eng. 2020;11:13.
46. Jamuar SS, Lam AT, Kircher M, D'Gama AM, Wang J, Barry BJ, et al. Somatic mutations in cerebral cortical malformations. N Engl J Med. 2014;371(8):733-43.

47. Konig K, Peifer M, Fassunke J, Ihle MA, Kunstlinger H, Heydt C, et al. Implementation of Amplicon Parallel Sequencing Leads to Improvement of Diagnosis and Therapy of Lung Cancer Patients. J Thorac Oncol. 2015;10(7):1049-57.
48. Rivas MA, Beaudoin M, Gardet A, Stevens C, Sharma Y, Zhang CK, et al. Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. Nat Genet. 2011;43(11):1066-73.
49. Schuster SC. Next-generation sequencing transforms today's biology. Nat Methods. 2008;5(1):16-8.
50. Shendure J, Ji H. Next-generation DNA sequencing. Nat Biotechnol. 2008;26(10):1135-45.
51. Ambardar S, Gupta R, Trakroo D, Lal R, Vakhlu J. High Throughput Sequencing: An Overview of Sequencing Chemistry. Indian J Microbiol. 2016;56(4):394-404.
52. Cele S, Gazy I, Jackson L, Hwa S-H, Tegally H, Lustig G, et al. Escape of SARS-CoV-2 501Y.V2 variants from neutralization by convalescent plasma. medRxiv. 2021:2021.01.26.21250224.
53. Wibmer CK, Ayres F, Hermanus T, Madzivhandila M, Kgagudi P, Oosthuysen B, et al. SARS-CoV-2 501Y.V2 escapes neutralization by South African COVID-19 donor plasma. Nat Med. 2021;27(4):622-5.
54. Wilkinson E, Giovanetti M, Tegally H, San JE, Lessells R, Cuadros D, et al. A year of genomic surveillance reveals how the SARS-CoV-2 pandemic unfolded in Africa. Science. 2021;374(6566):423-31.
55. van Dijk EL, Jaszczyszyn Y, Naquin D, Thermes C. The Third Revolution in Sequencing Technology. Trends Genet. 2018;34(9):666-81.
56. Chiara M, D'Erchia AM, Gissi C, Manzari C, Parisi A, Resta N, et al. Next generation sequencing of SARS-CoV-2 genomes: challenges, applications and opportunities. Brief Bioinform. 2021;22(2):616-30.
57. Wei X, Ju X, Yi X, Zhu Q, Qu N, Liu T, et al. Identification of sequence variants in genetic disease-causing genes using targeted next-generation sequencing. PLoS One. 2011;6(12):e29500.
58. illumina. Advantages of paired-end and single-read sequencing 2021 [updated 2021. Available from: https://www.illumina.com/science/technology/next-generation-sequencing/plan-experiments/paired-end-vs-single-read.html.
59. Petersen LM, Martin IW, Moschetti WE, Kershaw CM, Tsongalis GJ. Third-Generation Sequencing in the Clinical Laboratory: Exploring the Advantages and Challenges of Nanopore Sequencing. J Clin Microbiol. 2019;58(1):e01315-19.

ANNEXURE 1

BREC Approval Letter (BREC/00001195/2020)

UNIVERSITY OF KWAZULU-NATAL
INYUVESI YAKWAZULU-NATALI

16 March 2021

Prof Salim Safurdeen Abdool Karim (93336)
CAPRISA
MEDICAL SCHOOL

Dear Prof Karim,

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
Degree: Non-degree

RECERTIFICATION APPLICATION APPROVAL NOTICE

Approved:	**31 March 2021**
Expiration of Ethical Approval:	**30 March 2022**

I wish to advise you that your application for recertification received on 08 March 2021 for the above study has been **noted and approved** by a subcommittee of the Biomedical Research Ethics Committee (BREC). The start and end dates of this period are indicated above.

If any modifications or adverse events occur in the project before your next scheduled review, you must submit them to BREC for review. Except in emergency situations, no change to the protocol may be implemented until you have received written BREC approval for the change.

The committee will be notified of the above approval at its next meeting to be held on 13 April 2021.

Yours sincerely

[redacted]

.....
Ms A Marimuthu
(for) Prof D Wassenaar
Chair: Biomedical Research Ethics Committee

Biomedical Research Ethics Committee
Chair: Professor D R Wassenaar
UKZN Research Ethics Office Westville Campus, Govan Mbeki Building
Postal Address: Private Bag X54001, Durban 4000
Email: BREC@ukzn.ac.za
Website: http://research.ukzn.ac.za/Research-Ethics/Biomedical-Research-Ethics.aspx

ANNEXURE 2

BREC Approval Letter (BREC/00002764/2021)

UNIVERSITY OF KWAZULU-NATAL
INYUVESI YAKWAZULU-NATALI

04 October 2021

Mr Derek Kalala Tshiabuila (215024685)
School of Lab Med & Medical Sc
Medical School

Dear Mr Tshiabuila,

Protocol reference number: BREC/00002764/2021
Project title: Comparison of diagnostic methods for Severe Acute Respiratory Syndrome Corona virus 2 using Illumina sequencing, Real Time PCR, and Oxford Nanopore sequencing
Degree: MMedSc

EXPEDITED APPLICATION: APPROVAL LETTER

A sub-committee of the Biomedical Research Ethics Committee has considered and noted your application.

The conditions have been met and the study is given full ethics approval and may begin as from 04 October 2021. Please ensure that outstanding site permissions are obtained and forwarded to BREC for approval before commencing research at a site.

This approval is subject to national and UKZN lockdown regulations, see (http://research.ukzn.ac.za/Libraries/BREC/BREC_Amended_Lockdown_Level_1_Guidelines.sflb.ashx). Based on feedback from some sites, we urge PIs to show sensitivity and exercise appropriate consideration at sites where personnel and service users appear stressed or overloaded.

This approval is valid for one year from 04 October 2021. To ensure uninterrupted approval of this study beyond the approval expiry date, an application for recertification must be submitted to BREC on the appropriate BREC form 2-3 months before the expiry date.

Any amendments to this study, unless urgently required to ensure safety of participants, must be approved by BREC prior to implementation.

Your acceptance of this approval denotes your compliance with South African National Research Ethics Guidelines (2015), South African National Good Clinical Practice Guidelines (2020) (if applicable) and with UKZN BREC ethics requirements as contained in the UKZN BREC Terms of Reference and Standard Operating Procedures, all available at http://research.ukzn.ac.za/Research-Ethics/Biomedical-Research-Ethics.aspx.

BREC is registered with the South African National Health Research Ethics Council (REC-290408-009). BREC has US Office for Human Research Protections (OHRP) Federal-wide Assurance (FWA 678).

The sub-committee's decision will be noted by a full Committee at its next meeting taking place on 09 November 2021.

Yours sincerely,

[redacted]

Prof D Wassenaar
Chair: Biomedical Research Ethics Committee

Biomedical Research Ethics Committee
Chair: Professor D R Wassenaar
UKZN Research Ethics Office Westville Campus, Govan Mbeki Building
Postal Address: Private Bag X54001, Durban 4000
Email: BREC@ukzn.ac.za
Website: http://research.ukzn.ac.za/Research-Ethics/Biomedical-Research-Ethics.aspx

ANNEXURE 3

BREC Approval Letter (HREC 100/2017)

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Health Sciences

Institution: The Research Ethics Committee, Faculty Health Sciences, University of Pretoria complies with ICH-GCP guidelines and has US Federal wide Assurance.

- FWA 00002567, Approved dd 22 May 2002 and Expires 03/20/2022.
- IORG #: IORG0001762 OMB No. 0990-0279 Approved for use through February 28, 2022 and Expires: 03/04/2023.

Faculty of Health Sciences **Research Ethics Committee**

16 July 2021

Approval Certificate
Amendment

Dear Prof M Venter

Ethics Reference No.: 100/2017
Title: INVESTIGATION OF RESPIRATORY, EMERGING AND VECTOR BORNE VIRUSES AS THE CAUSE OF UNEXPLAINED ACUTE FEBRILE DISEASE WITH OR WITHOUT NEUROLOGICAL SIGNS AS WELL AS BIRTH DEFECTS INCLUSIVE OF STILL BIRTHS IN HUMANS IN SOUTH AFRICA

The **Amendment** as supported by documents received between 2021-06-30 and 2021-07-14 for your research, was approved by the Faculty of Health Sciences Research Ethics Committee on 2021-07-14 as resolved by its quorate meeting.

Please note the following about your ethics approval:

- Please remember to use your protocol number (100/2017) on any documents or correspondence with the Research Ethics Committee regarding your research.
- Please note that the Research Ethics Committee may ask further questions, seek additional information, require further modification, monitor the conduct of your research, or suspend or withdraw ethics approval.

Ethics approval is subject to the following:

- The ethics approval is conditional on the research being conducted as stipulated by the details of all documents submitted to the Committee. In the event that a further need arises to change who the investigators are, the methods or any other aspect, such changes must be submitted as an Amendment for approval by the Committee.

We wish you the best with your research.

Yours sincerely

[redacted]

On behalf of the FHS REC, Dr R Sommers
MBChB, MMed (Int), MPharmMed, PhD
Deputy Chairperson *of the Faculty of Health Sciences Research Ethics Committee, University of Pretoria*

The Faculty of Health Sciences Research Ethics Committee complies with the SA National Act 61 of 2003 as it pertains to health research and the United States Code of Federal Regulations Title 45 and 46. This committee abides by the ethical norms and principles for research, established by the Declaration of Helsinki, the South African Medical Research Council Guidelines as well as the Guidelines for Ethical Research: Principles Structures and Processes, Second Edition 2015 (Department of Health).

Research Ethics Committee
Room 4-60, Level 4, Tswelopele Building
University of Pretoria, Private Bag x323
Gezina 0031, South Africa
Tel +27 (0)12 356 3084
Email: deepeka.behari@up.ac.za
www.up.ac.za

Fakulteit Gesondheidswetenskappe
Lefapha la Disaense tša Maphelo