Unravelling the Genomic and Epidemiological Dynamics of SARS-CoV-2 in South Africa

Upasana Ramphal

Submitted in fulfilment of the academic requirements for the degree of Doctor of Philosophy (Medicine) in Virology, College of Health Sciences, Nelson R Mandela School of Medicine, School of Laboratory Medicine and Medical Sciences, University of KwaZulu-Natal

2025

"The more comfortable we become with being stupid, the deeper we will wade into the unknown and the more likely we are to make big discoveries"

(Martin A. Schwartz, 2008. *Journal of Cell Science*. doi:10.1242/jcs.033340)

Preface

The work described in this dissertation was carried out at the following laboratories: KwaZulu-Natal Research Innovation and Sequencing Platform (KRISP) located within the Nelson R Mandela School of Medicine at University of KwaZulu-Natal in Durban; Centre for Epidemic Response and Innovation (CERI) based at the BMRI building, Tygerberg Campus, Stellenbosch University and Central Analytical Facilities located within Stellenbosch University. Several laboratories also contributed to the generation of SARS-CoV-2 data for South Africa and are acknowledged as members of the Network for Genomic Surveillance in South Africa (NGS-SA). This study was conducted from March 2021 to May 2024 under the supervision of Prof. Tulio de Oliveira.

This study falls under the approval of the Biomedical Research Ethics Committee of the University of KwaZulu-Natal, South Africa, with protocol reference number BREC/00004745/2022. As no human subjects were involved, no informed consent was required. The details of data used throughout this study is available in the supplementary section. Acknowledgment is given to all sequencing laboratories as part of supplementary files S1 and S2.

The study represents original work by the author and has not been submitted in any form for any degree or diploma to any other tertiary institution. Where use has been made of the work of others, it is duly acknowledged in the text.

Signed: _[redacted]_ Date: 24 February 2025

Candidate: Upasana Ramphal

General Declaration

I, Upasana Ramphal, declare that:

(i) The research reported in this dissertation, except where otherwise indicated, is my original work.

(ii) This dissertation has not been submitted for any degree or examination at any other university.

(iii) This dissertation does not contain other persons' data, pictures, graphs or other information, unless specifically acknowledged as being sourced from other persons or resource.

(iv) 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:

a) their words have been re-written, but the general information attributed to them has been referenced;

b) where their exact words have been used, their writing has been placed inside quotation marks and referenced.

(v) Where I have reproduced a publication of which I am an author, co-author or editor, I have indicated in detail which part of the publication was actually written by myself alone and have fully referenced such publications.

(vi) This dissertation does not contain text, graphics, or tables copied and pasted from the Internet, unless specifically acknowledged, and the source being detailed in the dissertation and in the Reference sections.

Signed: _[redacted]_ Date: 24 February 2025

Candidate: Upasana Ramphal

As the candidate's Supervisor, in addition to the above, I confirm and agree to the submission of this dissertation.

Signed: [redacted] Date: 27 February 2025

Supervisor: Prof Tulio de Oliveira

Artificial Intelligence Declaration

In accordance with the UKZN Artificial Intelligence (AI) policy, Principles and Guidelines on the Use of Generative Artificial Intelligence Tools in Academic Work (Document Reference: SE/01/0312/24), I hereby declare the following regarding the submission of my PhD thesis:

This thesis utilised traditional AI tools (Grammarly and QuillBot), specialised design tools (Biorender) and generative AI tools (OpenAI products such as ChatGPT-4 and DALL-E) to support the refinement of writing and language, the creation of visual content (Figure 1.4 using Biorender, and Figure 4.1 using DALL-E), and the clarification and resolution of errors encountered during R programming. Refer to Appendix D for all prompts used within this thesis. In strict adherence to the UKZN AI Academic Guidelines, all AI-generated content was thoroughly evaluated for its relevance, accuracy, and originality. Proper attribution and citations were provided wherever necessary to acknowledge the contributions of original authors, and AI-generated content was transparently disclosed in the relevant sections of the thesis. Human oversight and critical evaluation were consistently maintained throughout all stages of the thesis development. The use of generative AI was fully aligned with the principles of academic integrity, ensuring that human critical thinking, originality, and creativity remained central to the research and its outcomes.

Signed:  Date: 24 February 2025

Candidate: Upasana Ramphal

Presentations

Part of the findings observed in this study were presented at the following conferences or symposia:

1. Poster Presentation at the **AHRI Research Day - incorporating SANTHE RSA Research day**, 18-19 October 2023, AHRI, Durban.
2. Poster presentation at **SANTHE Annual Consortium Meeting**, 10-15 July 2023, in Lusaka, Zambia.
3. Poster presentation at **The Global Health Network Conference, "Enabling Health Research in Every Healthcare Setting**", 24-25 November 2022, in Cape Town, South Africa.
4. Poster presentation at the **1st South African Workshop on SARS-CoV-2 Variants and Evolution**, 31 October - 3 November 2022, in St Lucia, South Africa.
5. Oral presentation at the **Virus Evolution and Molecular Epidemiology (COVEME)**, 30 August - 3 September 2021. Online Event.
6. Oral presentation at Virtual **SANTHE Annual Consortium Meeting**, 25-27 August 2021. Online event.
7. Oral presentation at the Virtual **SANTHE Research day**, 21-24 May 2021. Online event.
8. Oral presentation at the Virtual **SANTHE Research day**, 18-19 June 2020. Online event.
9. Oral and poster presentation at the **9th IDA Symposium**, 7-12 October 2019, River Club, Mowbray, Cape Town, South Africa.
10. Poster Presentation at the **SANTHE Annual Consortium Meeting**, 2-5 October 2019, Nairobi, Kenya.
11. Poster Presentation at the **ICGEB Workshop - Epigenetics of infectious and non-communicable diseases,** 16-19 September 2019, Cape Town, South Africa.
12. Oral presentation at the Virtual **SANTHE Research day**, 3-4 June 2019, Westville Country Club, South Africa.

Publications

Publications/Manuscripts from this thesis:

1. Reflective evaluation of Next-Generation Sequencing data during early phase detection of the Delta variant. **Ramphal U,** Tshiabuila D, Ramphal Y, Giandhari J, van Heerden C, Baxter C, van Wyk S, Pillay S, Laguda-Akingba O, Wilkinson E, Lessells R, de Oliveira T. *OBM Genetics* (2024); 8(2): 239; doi: https://10.21926/obm.genet.2402239 (30 May 2024).
2. Retrospective Evaluation of SARS-CoV-2 Genomic Data From KwaZulu-Natal during 2020 to 2023. **Ramphal U**, Giandhari J, Baxter C, Wilkinson E, Lessells R, de Oliveira T. Manuscript in preparation.
3. Pioneering Genomic Surveillance: Unveiling the Triumphs of South Africa in the SARS-CoV-2 Pandemic against a Global Framework. **Ramphal U**, Giandhari J, Ramphal Y, Baxter C, Wilkinson E, Lessells R, de Oliveira T. Manuscript in preparation.

*Collaborative Publications/Manuscripts during PhD

4. A year of genomic surveillance reveals how the SARS-CoV-2 pandemic unfolded in Africa. Wilkinson E, Giovanetti M, Tegally H, San JE, … , **Ramphal U**, ….., Nkengasong J, de Oliveira T, Science (2021). https://doi.org/10.1126/science.abj4336
5. Persistent SARS-CoV-2 infection and intra-host evolution in association with advanced HIV infection. Karim F, Moosa MYS, Gosnell BI, Cele S, Giandhari J, Pillay S, Tegally H, Wilkinson E, San JE, Msomi N, Mlisana K, Khan K, Bernstein M, Manickchund N, Singh L, **Ramphal U**, COMMIT-KZN Team, Hanekom W, Lessells RJ, Sigal A, de Oliveira T. 2021. https://doi.org/10.1101/2021.06.03.21258228
6. Rapid replacement of the Beta variant by the Delta variant in South Africa. Tegally H, Wilkinson E, Althaus C, Giovanetti M, San J, Giandhari J, Pillay S, Naidoo Y, **Ramphal U**, ….., Lessells R, de Oliveira T, medRxiv (2021), MEDRXIV-2021-264018v1:.
7. Management and containment of a SARS-CoV-2 Beta variant outbreak at the Malawi University of Science and Technology. Bandawe G, …., **Ramphal U**, San JE, Tegally H, Wilkinson E, de Oliveira T, Malata A. *MedRxiv* 2021. https://doi.org/10.1101/2021.11.25.21266298
8. Deciphering DNA Methylation in HIV Infection. Arumugam T, **Ramphal U**, Adimulam T, Chinniah R, Ramsuran V. *Frontiers Microbial Immunology* (2021). https://doi.org/10.3389/fimmu.2021.795121 PMID: 34925380; PMCID: PMC8674454.
9. Emergence and phenotypic characterization of the global SARS-CoV-2 C.1.2 lineage. Scheepers C, Everatt J, Amoako DG, Tegally H, Wibmer CK, Mnguni A, Ismail A, Mahlangu B, Lambson BE, Martin DP, Wilkinson E, San JE, Giandhari J, Manamela N, Ntuli N, Kgagudi P, Cele S, Richardson

SI, Pillay S, Mohale T, **Ramphal U**, ……, de Oliveira T, von Gottberg A, Bhiman JN, *Nature Commununications* (2022), 13(1):1976. https://doi.org/10.1038/s41467-022-29579-9

10. Comparison of SARS-CoV-2 sequencing using the ONT GridION and the Illumina MiSeq. Tshiabuila D, Giandhari J, Pillay S, **Ramphal U**, Ramphal Y, Maharaj A, Anyaneji UJ, Naidoo Y, Tegally H, San EJ, Wilkinson E, Lessells R, de Oliveira T, *BMC genomics* (2022). https://doi.org/10.1186/s12864-022-08541-5
11. SARS-CoV-2 variants from COVID-19 positive cases in the Free State province, South Africa from July 2020 to December 2021. Mwangi P, Okendo J, ….., **Ramphal U**, de Oliveira T, Bester A, Nyaga M. *Frontiers Virology* (2022). https://doi.org/10.3389/fviro.2022.935131
12. Rapid epidemic expansion of the SARS-CoV-2 Omicron variant in southern Africa. Viana R*, Moyo S*, Amoako DG*, Tegally H*, …., **Ramphal U**, San JE, Scott L, Shapiro R, Singh L, Smith-Lawrence P, Stevens W, Strydom A, Subramoney K, Tebeila N, Tshiabuila D, Tsui J, van Wyk S, Weaver S, Wibmer CK, Wilkinson E, Wolter N, Zarebski AE, Zuze B, Goedhals D, Preiser W, Treurnicht F, Venter M, Williamson C, Pybus OG, Bhiman J, Glass A, Martin DP, Rambaut A, Gaseitsiwe S, von Gottberg A, de Oliveira T, *Nature* (2022). https://doi.org/10.1038/s41586-022-04411-y
13. Rapid Replacement of SARS-CoV-2 Variants by Delta and Subsequent Arrival of Omicron, Uganda, 2021. Bbosa N, Ssemwanga D, Namagembe H, Kiiza R, Kiconco J, Kayiwa J, Lutalo T, Lutwama J, Ssekagiri A, Ssewanyana I, Nabadda S, Kyobe-Bbosa H, Giandhari J, Pillay S, **Ramphal U**, Ramphal Y, Naidoo Y, Tshiabuila D, Tegally H, San EJ, Wilkinson E, de Oliveira T, Kaleebu P. *Emerg Infect Dis* (2022). https://doi.org/10.3201/eid2805.220121
14. Identification of SARS-CoV-2 Omicron variant using spike gene target failure and genotyping assays, Gauteng, South Africa, 2021. Subramoney K, Mtileni N, Bharuthram A, Davis A, Kalenga B, Rikhotso M, Maphahlele M, Giandhari J, Naidoo Y, Pillay S, **Ramphal U**, …..., Treurnicht FK. *Journal of Medical Virology* (2022). https://doi.org/10.1002/jmv.27797
15. The evolving SARS-CoV-2 epidemic in Africa: Insights from rapidly expanding genomic surveillance. Tegally H, San JE, Cotten M, …, **Ramphal U**, …… de Oliveira T*, Happi C, Lessells R, Nkengasong J, Wilkinson E*, Science (2022). https://doi.org/10.1126/science.abq5358
16. Genomic epidemiology of SARS CoV-2 during the first four waves in Mozambique. Ismael N, van Wyk S, Tegally H, Giandhari J, San JE, Moir M, Pillay S, Utpatel C, Singh, L, Naidoo, Y, **Ramphal U**, …., de Oliveira, T. *Plos Global public Health*. https://doi.org/10.1371/journal.pgph.0001593.
17. Global disparities in SARS-CoV-2 genomic surveillance. Brito, AF et al., 2022. *Nature Communications 13* (1):7003. http://dx.doi.org/10.1038/s41467-022-33713-y
18. SARS-CoV-2 spike protein diversity at an intra-host level, among SARS-CoV-2 infected individuals in South Africa, 2020 to 2022. Subramoney K, Mtileni N, Davis, A, Giandhari J, Tegally H, Wilkinson, E, Naidoo Y, **Ramphal U**, …., Treunicht. *Plos ONE* (2023) https://doi.org/10.1371/journal.pone.0286373

19. Molecular Epidemiology and Diversity of SARS-CoV-2 in Ethiopia, 2020–2022. Sisay A, ….., **Ramphal U**, Chabuka L, Wilkinson E, de Oliveira T, Desta AF, San JE. *Genes* (2023). https://doi.org/10.3390/genes14030705
20. Molecular epidemiology of SARS-CoV-2 during five COVID-19 waves and the significance of low frequency lineages. Subramoney K, Mtileni N, Giandhari J, Naidoo Y, Ramphal Y, Pillay S, **Ramphal U**, …., Treurnicht FK. *Viruses* (2023). https://doi.org/10.3390/v15051194
21. Identification of SARS-CoV-2 Omicron variant using spike gene target failure and genotyping assays, Gauteng, South Africa, 2021. Subramoney K, Mtileni N, Bharuthram A, Davis A, Kalenga B, Rikhotso M, Maphahlele M, Giandhari J, Naidoo Y, Pillay S, **Ramphal U,** Ramphal Y, Tegally H, Wilkinson E, Mohale T, Ismail A, Mashishi B, Mbenenge N, de Oliveira T, Makatini Z, Fielding BC, Treurnicht FK, J Med Virol. (2022), 4(8):3676-3684. https://doi.org/10.1002/jmv.27797.

* I contributed to the studies investigation, methodology, data generation, drafting and review of the manuscripts.

Dedication

I dedicate this degree and thesis to myself:

My unwavering commitment, resilience, and determination.

For recognising my self-worth and investing in my personal growth.

For overcoming obstacles, regardless of their shape or magnitude.

For persevering through challenges, both expected and unforeseen.

For making sacrifices throughout these years in pursuit of my goals.

For rising each day with renewed determination.

For enduring the difficult moments and pushing forward.

For believing in my inner strength and never wavering in faith in myself.

I did it!

A new chapter begins.

~

To my Father

To my late Ma - you always saw me as Dr Upasana Ramphal

♡

Acknowledgements

This work would not have been possible without the contribution of several individuals / teams:

CAPRISA and KRISP for the financial and unconditional support throughout the journey of this PhD. Sub-Saharan African Network for TB/HIV Research Excellence (SANTHE) team, for the ongoing motivation and support. This dissertation was partially funded by the Bill & Melinda Gates Foundation, through a grant to SANTHE.

I want to thank Professor Tulio de Oliveira for the opportunity to be part of the KRISP and CERI team. I am thankful for the experience.

I am deeply grateful for the invaluable guidance and unwavering support provided by Dr Jennifer Giandhari, Dr Richard Lessells, Dr Cheryl Baxter and Prof Bongani Nkambule throughout my PhD journey.

I am truly grateful to the students and staff at KRISP and CERI for their support, assistance and guidance.

I am deeply thankful to my father, sisters and friends for their unwavering love, encouragement and support throughout my studies. Without your support, I would not have reached this point in my life. Heartfelt appreciation for my pets in bringing me tranquillity and peace when I needed it most. Your presence soothes me in ways nothing else can. To my partner, Amith, your care and strength have been invaluable on this journey. I am grateful for your love, support, patience and understanding.

Finally, a nod to the forces that have fuelled my inner confidence, drive and ambition, helping me achieve one of my dreams and embark on this incredible adventure. This journey has been truly memorable.

Table of contents

List of Figures

List of Tables

List of Abbreviations

AI	Artificial Intelligence
AMR	antimicrobial resistance
bp	base pair
CAPRISA	Centre for the Aids Programme of Research in South Africa
CAF	Central Analytical Facilities
CDC	Centers for Disease Control and Prevention
cDNA	Complementary Deoxyribonucleic Acid
CERI	Centre for Epidemic Response and Innovation
CFR	Case Fatality Rate
CoVs	Coronaviruses
COVID-19	Coronavirus Disease 2019
Ct	Cycle threshold
DNA	Deoxyribonucleic Acid
ENA	European Nucleotide Archive
GISAID	Global Initiative on Sharing Avian Influenza Data
GSS	Genomic Surveillance System
ICTV	International Committee on Taxonomy of Viruses
km^2	Kilometre squared (area)
KRISP	KwaZulu-Natal Research Innovation and Sequencing Platform
KZN	KwaZulu-Natal
MERS-CoV	Middle East Respiratory Syndrome Coronavirus
NCBI	National Center for Biotechnology Information
NGS	Next-generation Sequencing
NGS-SA	Network for Genomic Surveillance of South Africa
NHLS	National Health Laboratory Services
Ns	Non-significant
ONT	Oxford Nanopore Technology
OWID	Our World in Data
PCR	Polymerase Chain Reaction
PHEIC	Public Health Emergency of International Concern
qPCR	Quantitative Polymerase Chain Reaction (real-time PCR)
QC	Quality Control
RNA	Ribonucleic Acid
SGTF	S-gene Target Failure
S-gene	Spike Gene

SA	South Africa
SANTHE	Sub-Saharan African Network for TB/HIV Research Excellence
SARS-CoV	Severe Acute Respiratory Syndrome Coronavirus
SARS-CoV-2	Severe Acute Respiratory Syndrome Coronavirus 2
SNPs	Single Nucleotide Polymorphisms
SR	Sequencing Rates
SRA	Sequencing Read Archive
ssRNA	Single-stranded Ribonucleic Acid
TAT	Turnaround Time
TNA	Total Nucleic Acid
UKZN	University of KwaZulu-Natal
VL	Viral Load
VOC	Variant of Concern
VOI	Variant of Interest
VUM	Variant Under Monitoring
WoS	Web of Science
WGS	Whole-Genome Sequencing
WHO	World Health Organisation

Abstract

The Coronavirus Disease 2019 (COVID-19) pandemic necessitated a swift and coordinated global response, with genomic surveillance playing a crucial role in tracking the evolution and spread of Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2). This study presents a comprehensive retrospective analysis of genomic surveillance data, drawing from multiple sources, including the Global Initiative on Sharing Avian Influenza Data (GISAID), Our World in Data (OWID), and national datasets. By leveraging genomic data, this research highlights the importance of surveillance in strengthening outbreak preparedness and public health responses. The thesis is structured into five key chapters. Chapter One presents a detailed literature review, contextualising the pandemic and highlighting the significance of genomic surveillance. Chapter Two, a published manuscript, retrospectively compares the performance of the Illumina MiSeq and Ion Torrent S5 sequencing platforms in detecting the Delta variant during its early emergence. It assesses their efficiency, accuracy, and suitability for genomic surveillance, while addressing challenges in comparing data generated by different sequencing technologies. Chapter Three conducts a retrospective trend analysis of 6,089 SARS-CoV-2 genomes from KwaZulu-Natal, South Africa. It examines genomic surveillance trends across urban and rural districts, focusing on lineage distribution and transmission dynamics. The study aims to identify surveillance gaps and propose strategies to enhance epidemic preparedness in localised settings. Chapter Four expands on this analysis by comparing South Africa's genomic surveillance strategies with global initiatives, identifying strengths and areas for improvement. It evaluates data-sharing practices and pandemic response frameworks, highlighting successes and challenges relative to countries with similar or higher socioeconomic contexts and their impact on timely public health interventions. The final chapter synthesises the key findings, emphasising the need for sustained investment in genomic monitoring, technological advancements, and global collaboration to enhance preparedness for future public health threats. The study reveals that, while South Africa has made significant progress in genomic surveillance, disparities in sequencing coverage, particularly in rural areas and low-income countries, pose challenges to comprehensive outbreak monitoring. The comparative analysis of sequencing platforms offers critical insights into the reliability and efficiency of different methodologies, essential for optimising genomic surveillance systems. Additionally, the global comparison demonstrates the importance of international collaboration in genomic data sharing, providing strategic recommendations to strengthen surveillance efforts. In conclusion, this study demonstrates that genomic surveillance is indispensable for real-time pandemic response, variant tracking, and public health decision-making. By advocating for improvements in sequencing infrastructure, equitable data access, and enhanced global cooperation, this research contributes to building a more resilient and effective approach to managing future infectious disease outbreaks.

Keywords: COVID-19, SARS-CoV-2, genomic surveillance, GISAID, next-generation sequencing, KwaZulu-Natal, South Africa, public health, pandemic preparedness.

Thesis Outline

This thesis comprises five chapters.

Chapter One presents a review of relevant literature, establishing the background and significance of the study. It offers an in-depth overview of the SARS-CoV-2 pandemic from global to local perspectives, addressing achievements, challenges and limitations in genomic surveillance.

Chapter two is a publication titled “Reflective Evaluation of Next-Generation Sequencing Data during Early Phase Detection of the Delta Variant”. This chapter retrospectively analyses data from Illumina MiSeq and Ion Torrent S5 platforms used in early Delta variant detection. It emphasises Genome Detective and NextClade tools in assessing data variability and discusses considerations for comparing data from different sequencing technologies (Ramphal *et al.*, 2024).

Chapter Three is a manuscript titled “Retrospective Analysis of SARS-CoV-2 Genomic Surveillance Trends in KwaZulu-Natal from 2020 to 2023,” and examines genomic surveillance trends in urban and rural districts within KwaZulu-Natal. It examines SARS-CoV-2 lineage distribution and transmission trends within a well-documented province. The goal is to identify surveillance gaps and suggest improvements to existing strategies for future epidemic preparedness in localised settings.

Chapter Four is a manuscript titled “Unveiling South Africa's Triumphs: An Exploration of SARS-CoV-2 Genome Sharing Trends and Impacts, 2020–2024”. This chapter examines South Africa's genomic data sharing in comparison to global efforts, identifying areas for improvement. It highlights successes and challenges relative to countries with similar or higher socioeconomic levels, showing differences in pandemic responses and impacts on timely public health interventions.

Chapter Five synthesises key findings from Chapters two to four, drawing conclusions, situating them within the context of existing literature, discussing limitations and suggesting future research directions.

Chapter One: Background and Literature Review

1.1 Background

The novel Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) was initially reported in Wuhan, China in late 2019 (Wu *et al.*, 2020). This marked the beginning of a global health crisis, later termed the Coronavirus Disease 2019 (COVID-19) pandemic. While the earliest cases were mainly detected in Wuhan, subsequent investigations have revealed that the virus may have originated simultaneously in multiple countries (Liu *et al.*, 2021). This hypothesis was supported by retrospective analyses of stored biological samples, which have detected SARS-CoV-2 RNA in specimens collected prior to the recognised outbreak in Wuhan (Liu *et al.*, 2021).

Over the next few months, the World Health Organisation (WHO) and the Centres for Disease Control and Prevention (CDC) undertook extensive efforts to understand and mitigate the spread of the virus (Figure 1.1). The WHO declared the outbreak a Public Health Emergency of International Concern (PHEIC) on 30 January 2020, subsequently naming the disease COVID-19 and identifying the virus as SARS-CoV-2 (Gorbalenya *et al.*, 2020; Hu *et al.*, 2021). On 11 March 2020, the WHO elevated its assessment, declaring the COVID-19 outbreak a global pandemic.

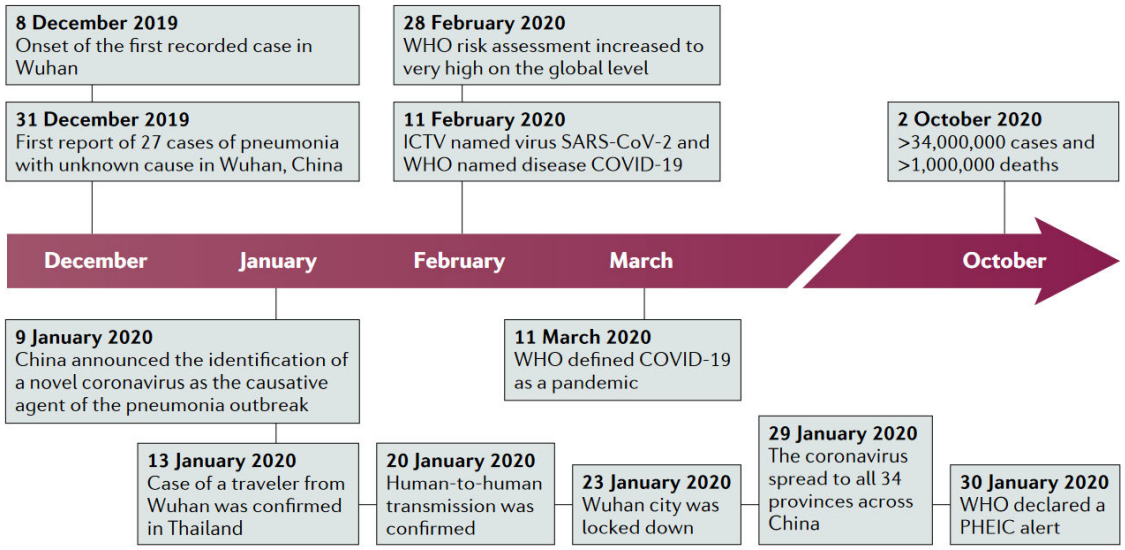


Figure 1.1: Chronology of Major Events in the COVID-19 Pandemic (Hu *et al.*, 2021)

These declarations underscored the urgent need for coordinated global action to address the rapidly escalating health crisis (Sohrabi *et al.*, 2020). In the subsequent weeks, the rate of SARS-CoV-2 transmission surged exponentially across the globe, resulting in >70 million cases within the first year of the pandemic (*Worldometers*, 2025). Globally, as of 5 January 2025, there have been over 777 million confirmed cases of COVID-19, including 7,083,246 deaths, reported to WHO (World Health Organisation, 2025). In South Africa, as of 5 January 2025, over 4,1 million confirmed cases of COVID-

19 were reported and over 103,000 associated deaths (Mathieu *et al.*, 2024; World Health Organisation, 2025).

The emergence of SARS-CoV-2 has had profound and lasting impacts on global health, economies and societies (Nicola *et al.*, 2020; Mofijur *et al.*, 2021). It has highlighted the importance of international collaboration, robust public health infrastructure and ongoing research for future pandemics. The lessons learned from this unprecedented crisis continue to shape the global health landscape, driving innovations in vaccine technology, infectious disease surveillance and public health strategies. The emergence of SARS-CoV-2 prompted global public health responses, including the establishment of genomic surveillance systems to monitor and track its evolution. In this context, genomic surveillance emerged as a key tool in the fight against SARS-CoV-2 (Robishaw *et al.*, 2021). By analysing the genetic structure of the virus, researchers gained valuable insights into its evolution, transmission patterns and potential impact on public health. This knowledge allowed for the development of effective strategies to control the spread of the virus, adapt public health measures and optimise vaccine development (Akande *et al.*, 2023).

The collection of ongoing genomic surveillance data is important in improving our understanding of SARS-CoV-2 and its epidemiology. By integrating genomic data with individual level metadata, researchers are able to identify demographic variables influencing transmission patterns (Hill *et al.*, 2021, 2023). Such information is valuable in understanding the transmission trends within communities and populations (Schriml *et al.*, 2020). Ongoing surveillance data has also lead to the improvement of existing systems, procedures and technologies, thereby increasing the effectiveness of genomic epidemiology of SARS-CoV-2 (Hill *et al.*, 2021). Publicly accessible databases, such as the Global Initiative on Sharing All influenza Data (GISAID), facilitated the sharing of over 16.7 million SARS-CoV-2 sequences from more than 215 countries worldwide (Khare *et al.*, 2021). This repository provided an excellent opportunity for evaluating trends and conducting comparative analyses on existing data from small communities and regions with those of national and international levels (Potdar *et al.*, 2021). By leveraging existing genomic surveillance data, we can improve our knowledge and approach for future epidemic and pandemic preparedness. Such data aids in understanding current viral dynamics and informs the development of surveillance systems capable of early detection and rapid response to emerging infectious diseases (Ling-Hu *et al.*, 2022). The consensus on genomic surveillance is that its ongoing evolution and refinement are essential for addressing future public health threats, thereby strengthening global health resilience (Akande *et al.*, 2023; Hill *et al.*, 2023).

1.1.1 Study significance and rationale

This study utilises existing genomic surveillance data from South Africa to identify gaps and challenges in current initiatives, aiming to strengthen strategies for future disease outbreaks. By analysing data at local, national, and global levels, it provides insights to strengthen public health

practices and outcomes. The primary objective is to use existing surveillance data effectively for comparative analysis, enabling meaningful conclusions that could inform public health policies and possibly improve outbreak response strategies. By identifying critical areas for improvement, this study underscores the role of genomic surveillance in strengthening public health preparedness and informing data-driven decision-making for future outbreaks.

1.1.2 Study goal and objectives

The overarching goal of this thesis is to utilise genomic data to uncover valuable insights into local trends, to understand South Africa's pandemic response and derive lessons to inform preparedness for future outbreaks. The study comprises three key analyses: first, a comparison of genomic data from two sequencing platforms, representing publicly available data and metadata used by researchers to detect trends and address surveillance limitations; second, an evaluation of sampling strategies in a well-annotated province to assess data representation and trends; and third, a global comparison of local genomic data with international datasets to evaluate data quality, quantity, and timeliness across various socioeconomic contexts.

These objectives directly address the gaps identified in the literature review, particularly the limited comparative evaluation of genomic data quality, the unevenness of surveillance across regions, and the need for contextualising South Africa’s genomic response within the global surveillance framework.

1.1.3 Study Considerations

Considering COVID-19's transition to endemic status and the cessation of pandemic containment measures, the study employs past tense throughout chapter one to clarify the initial stages of the pandemic and accurately depict these activities as historical events. This tense usage also reflects the current status of COVID-19, locally and globally. Traditional and Generative AI tools were employed in this thesis. For transparency, the specific prompts used to generate responses or visual content have been documented in Appendix D. These prompts guided the AI tools to refine language, resolve coding errors, and create visual materials while adhering to the principles of academic integrity.

1.2 Literature Review

The following literature review provides an overview of SARS-CoV-2, encompassing its virology, classification, epidemiology, and current prevention strategies, as well as highlighting key achievements and challenges during the pandemic. There have been many lessons that we learned to enable better preparation for future challenges. This review examines SARS-CoV-2 within the study's objectives, highlighting the role of genomic data in understanding processes, local trends, assessing South Africa's experience, and guiding future epidemic preparedness.

1.2.1 SARS-CoV-2 History and Classification

SARS-CoV-2, belonging to the family of coronaviruses (CoVs), is the largest group of the order *Nidovirales* (Sofi, Hamid and Bhat, 2020). The order *Nidovirales* is classified within the *Coronaviridae* family, which is a group of enveloped, positive-sense, single-stranded RNA (ssRNA) viruses known to infect mammals and birds (Dhama et al., 2020). This subfamily, *Coronavirinae*, is classified into four genera: *Alphacoronavirus*, *Betacoronavirus*, *Gammacoronavirus* and *Deltacoronavirus* (Figure 1.2) (Shahzad *et al*., 2020; Sofi, Hamid and Bhat, 2020; Wang *et al.*, 2020).

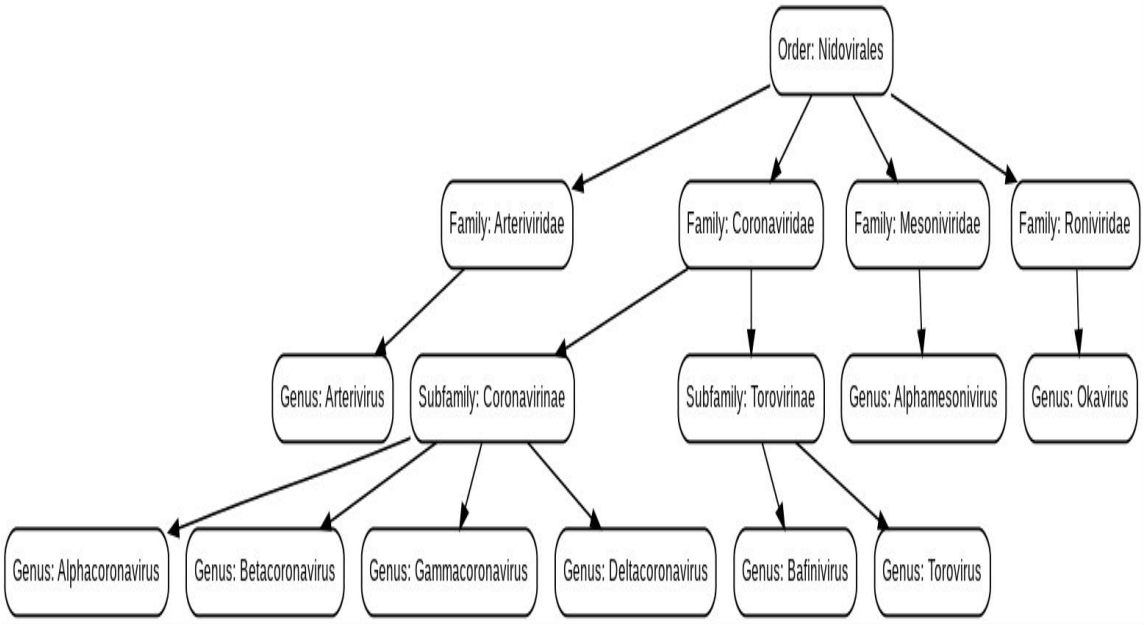


Figure 1.2: The taxonomy of the order *Nidovirales* (Sofi, Hamid and Bhat, 2020)

SARS-CoV-2 falls under the *Betacoronavirus* genus, which also includes the other notable human coronaviruses, MERS and SARS-CoV (Sofi, Hamid and Bhat, 2020). Within the *Betacoronavirus* genus, SARS-CoV-2 was classified as a lineage B *betacoronavirus* (Li *et al.*, 2021). The lineage B *betacoronaviruses* is further divided into four distinct subgenera: *Sarbecovirus, Merbecovirus, Nobecovirus* and *Hibecovirus* (Dhama *et al.*, 2020). SARS-CoV-2, belonging to the *Sarbecovirus* subgenus, also includes the closely related bat coronaviruses, such as RaTG13 and

RmYN02, identified in horseshoe bats (*Rhinolophus affinis* and *Rhinolophus malayanus*, respectively) (Andersen *et al.*, 2020; Hu *et al.*, 2021). The bat coronaviruses share a high genetic similarity with SARS-CoV-2, emphasising potential zoonotic origin as previously mentioned (Ye *et al.*, 2020). Similar to Middle East Respiratory Syndrome Coronavirus (MERS-CoV) and Severe Acute Respiratory Syndrome Coronavirus (SARS-CoV), such zoonotic viruses were also known for their ability to cause respiratory illnesses in both humans and animals (Figure 1.3) (Fadaka *et al.*, 2020; Hu *et al.*, 2021).

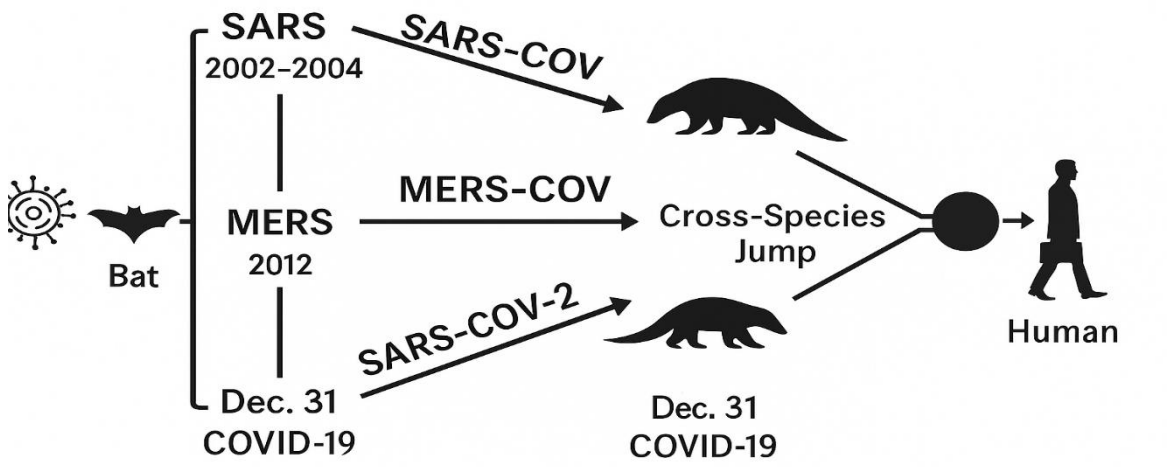


Figure 1.3: Possible transmission routes of coronaviruses from carrier to human (Fadaka *et al.*, 2020)

SARS-CoV-2 shares genetic similarities, origins, transmission mechanisms and clinical manifestations with MERS and SARS-CoV (Lu *et al.*, 2020). SARS-CoV was the virus responsible for the 2002-2003 Severe Acute Respiratory Syndrome outbreak and MERS-CoV was the virus responsible for the Middle East Respiratory Syndrome outbreak in 2012 (Yang and Leibowitz, 2015; Yin and Wunderink, 2018). Considering the effects of MERS and SARS, the emergence of SARS-CoV-2 elicited considerable apprehension within the scientific community due to its swift progression (Fadaka *et al.*, 2020).

The classification of SARS-CoV-2 at the variant level is based on the genetic variations observed in its genome (Koyama, Platt and Parida, 2020). Like most viruses, SARS-CoV-2 continuously evolves to adapt to the changes in the environment via random genomic mutations that are mostly focused on natural selection (Huang and Wang, 2021). The majority of mutations are neutral or damaging to the virus itself (Harvey *et al.*, 2021). However, some mutations were found to provide advantageous effects, such as viral escape to the host immune system, resistance to antivirals and increased viral fitness (Perez-Gomez, 2021; Abbasian *et al.*, 2023; Carabelli *et al.*, 2023). Due to the latter, an increase in the transmission of SARS-COV-2 was observed and this allowed for easy detection at a population level (Pritchard *et al.*, 2022).

The WHO and CDC have played crucial roles in the classification and monitoring of SARS-CoV-2 variants, which were important in understanding the evolution and spread of the virus (Chakraborty *et al.*, 2021). These organisations have implemented rigorous and systematic approaches to identify and categorise the different variants based on their genetic characteristics and potential impact on public health (Chakraborty *et al.*, 2021). The WHO established the SARS-CoV-2 Virus Evolution Working Group in June 2020, in collaboration with various global networks to monitor and evaluate the emergence and significance of different variants (Subissi *et al.*, 2022; World Health Organisation, 2023c). The group analysed data from genomic sequencing and other sources to identify genetic changes in the virus. The WHO developed a comprehensive classification system that categorises variants into different lineages and provides standardised nomenclature. They assigned Greek letters to different Variants of Concern (VOC) and some Variants of Interest (VOI) lineages only, to facilitate global communication and prevent stigmatisation of specific countries or regions (Konings *et al.*, 2021; Centres for Disease Control and Prevention, 2023). The CDC in the United States was also actively involved in monitoring and classifying SARS-CoV-2 variants.

The agency uses a variant classification system that includes four categories: VOIs, VOC, Variants of High Consequence (VOHCs) and Variants Under Monitoring (VUM) (Salehi-Vaziri *et al.*, 2022; Centres for Disease Control and Prevention, 2023). This classification was based on several factors, including evidence of increased transmissibility, potential reduction in neutralisation by antibodies and increased severity of disease. Notable VOCs (Figure 1.4) were the Alpha variant (B.1.1.7) (Wise, 2020), the first VOC identified in the United Kingdom; the Beta variant (B.1.351) (Tegally *et al.*, 2021), first identified in South Africa; the Gamma variant (P.1) (Faria *et al.*, 2021), first identified in Brazil; the Delta variant (B.1.617.2) (Cherian *et al.*, 2021), first identified in India; and the Omicron variant (B.1.1.529), first reported to WHO from South Africa (Salehi-Vaziri *et al.*, 2022; Viana *et al.*, 2022). It was essential to recognise that identifying variants in a particular country did not automatically suggest that these variants originated or evolved in that region (Choi *et al.*, 2023). The WHO and the CDC consistently revise their variant classifications in the light of emerging data (Ramesh *et al.*, 2021). It is important to note that while differences in transmissibility and immune escape between VOCs are well established, evidence for intrinsic differences in severity is more variable, with Delta representing the clearest example of reliably increased severity across multiple adjusted analyses.

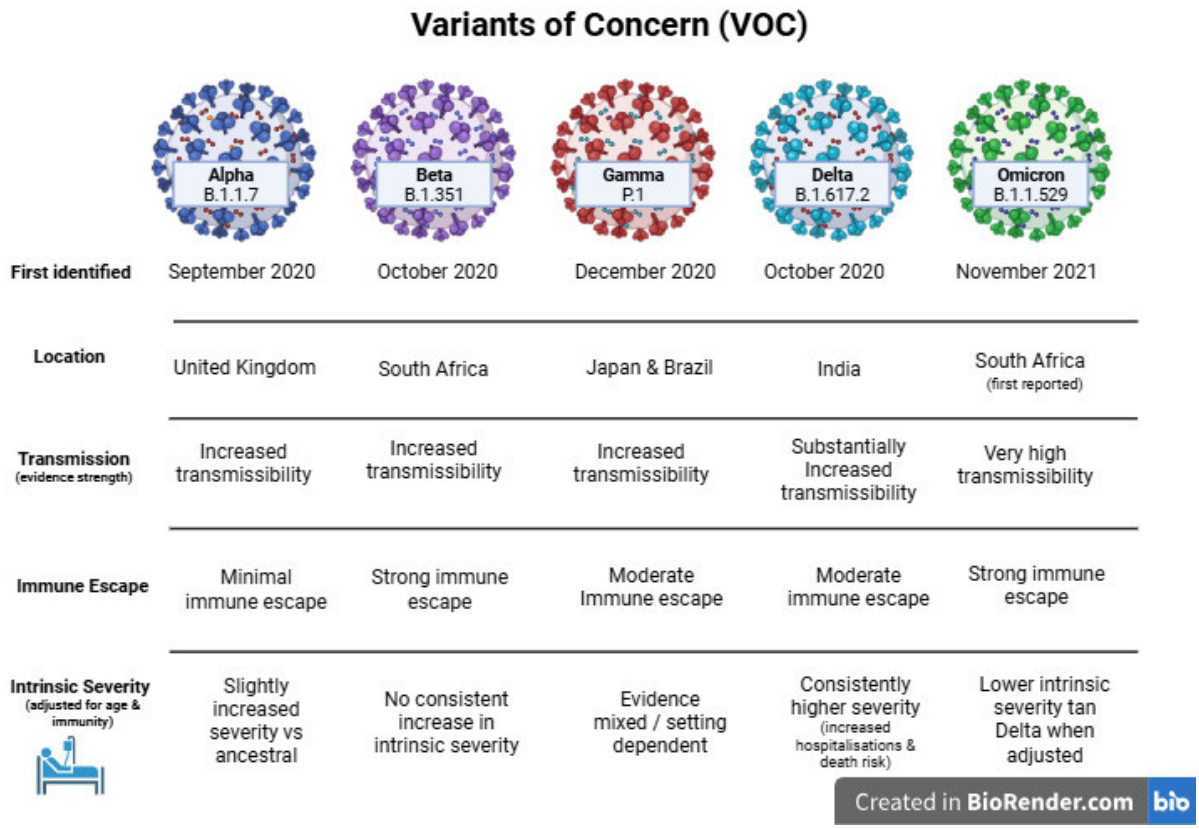


Figure 1.4: WHO-designated SARS-CoV-2 Variants of Concern (VOCs) with comparative evidence on transmissibility, immune escape and intrinsic severity. Evidence consistently supports increased transmissibility across all VOCs and strong immune escape for Beta and Omicron. The most robust signal of increased intrinsic severity is associated with Delta, whereas Omicron demonstrates lower intrinsic severity than Delta when controlling for prior immunity, age, and vaccination status. Severity estimates remain context-dependent due to differences in population immunity and healthcare capacity. Created on Biorender.com (Cele *et al.*, 2021; Davies *et al.*, 2021; Wolter *et al.*, 2021; Jassat *et al.*, 2022; Pulliam *et al.*, 2022; Salehi-Vaziri *et al.*, 2022; Viana *et al.*, 2022)

1.2.2 Scientific and Lineage-based Nomenclature Initiatives

The nomenclature initiatives for SARS-CoV-2 have been crucial in facilitating clear communication and coordination among the global scientific and public health communities (Rambaut *et al.*, 2020). These initiatives aimed to standardise the naming of the virus and its variants, ensuring that discussions regarding the virus and its mutations were precise and universally understood, thereby avoiding confusion and stigmatisation. The *Coronaviridae* Study Group's classification efforts and recommendations have been foundational, helping streamline communication in identifying and discussing new isolates (Gorbalenya *et al.*, 2020). Various lineages and sublineages are identified within SARS-CoV-2 variants, each characterised by the specific mutations and genetic signatures (Alkhatib *et al.*, 2021; Franceschi *et al.*, 2021; Chrysostomou *et al.*, 2022). These variants were designated using different scientific and lineage-based nomenclature initiatives. The Phylogenetic Assignment of Named

Global Outbreak Lineages (Pangolin), GISAID and Nextstrain were all lineage-based nomenclature initiatives (O'Toole *et al.*, 2021; Potdar *et al.*, 2021). Each provides a detailed understanding of the virus's evolution and spread and are essential for genomic surveillance, research, and public health response. However, each initiative has a distinct role and characteristic.

a) Pangolin is a bioinformatics framework and pipeline that assigns phylogenetic clades and lineages to viral sequences (Rambaut *et al.*, 2020). It is primarily used for SARS-CoV-2 and other viral genomes. Pangolin assigned alphanumeric designations using a systematic approach to identifying and categorizing new lineages, critical for real-time epidemiological tracking of SARS-CoV-2 (O'Toole *et al.*, 2021; McBroome *et al.*, 2023).

b) GISAID is an international initiative that promotes the sharing of influenza and other virus sequence data (Khare *et al.*, 2021). During the COVID-19 pandemic, GISAID played a crucial role in collecting and sharing genetic data of SARS-CoV-2 from different countries and researchers. It provides an open-access platform for scientists to deposit and access viral genome sequences, facilitating global collaboration in monitoring the virus's genetic evolution and understanding its spread (Elbe and Buckland-Merrett, 2017; Shu and McCauley, 2017; Khare *et al.*, 2021).

c) Nextstrain is an open-source bioinformatics tool that visualises and tracks the genomic evolution of pathogens, including SARS-CoV-2 (Aksamentov *et al.*, 2021). It uses phylogenetic analysis to create real-time visualisations of the virus's genetic diversity and spread. Nextstrain integrates genetic data from various sources, including GISAID and provides interactive phylogenetic trees and geographic visualisations. It assists researchers and public health officials in understanding the viral transmission patterns and how different lineages spread in specific regions (Hadfield *et al.*, 2018; Aksamentov *et al.*, 2021).

In summary, Pangolin is a pipeline for assigning standardised clade designations to virus genomes, GISAID is a platform for sharing viral sequence data and Nextstrain is a bioinformatics tool for visualising the genetic evolution and transmission of pathogens (Potdar *et al.*, 2021). Together, these initiatives contribute to the global efforts in monitoring and understanding the dynamics of SARS-CoV-2, as well as managing evolving public health challenges posed by SARS-CoV-2 variants.

1.2.3 SARS-CoV-2 Replication Cycle

SARS-CoV-2 infection begins when the viral spike (S) protein binds to the host angiotensin-converting enzyme 2 (ACE2) receptor on epithelial cells of the respiratory tract. This interaction is facilitated by host proteases such as TMPRSS2, allowing membrane fusion and entry of the viral RNA genome into the cytoplasm (Hoffmann *et al.*, 2020; Kakavandi *et al.*, 2023). Once released, the positive-sense RNA genome is directly translated to produce the non-structural proteins that assemble into the replication-transcription complex (RTC) (Malone *et al.*, 2022; Mishchenko and Ivanisenko, 2022). The

RTC drives synthesis of new genomic RNA and a series of sub-genomic RNAs that encode the structural proteins (S, E, M, and N). These components are assembled in the endoplasmic reticulum-Golgi intermediate compartment (ERGIC), after which complete virions are transported to the cell membrane and released via exocytosis to initiate further rounds of infection (Hartenian *et al.*, 2020; Klein *et al.*, 2020). This replication mechanism underpins viral proliferation and provides the substrate for genetic variation during transmission.

1.2.4 Emergence and Transmission Dynamics

The initial cases of SARS-CoV-2 were linked to a seafood market where live animals were sold (Li *et al.*, 2020). Investigations revealed that the virus likely originated in bats and may have been transmitted to humans through an intermediate animal host (Zhou *et al.*, 2020; Worobey *et al.*, 2022). SARS-CoV-2 belongs to the family of coronaviruses, which are enveloped viruses with single-stranded RNA genomes. In contrast to MERS and SARS-CoV, SARS-CoV-2 exhibits distinct genetic characteristics that set it apart from its predecessors (Lu et al., 2020). The transmission dynamics of SARS-CoV-2 primarily involve respiratory droplets expelled when an infected individual coughs, sneezes, talks, or breathes heavily (Figure 1.5) (Harrison, Lin and Wang, 2020).

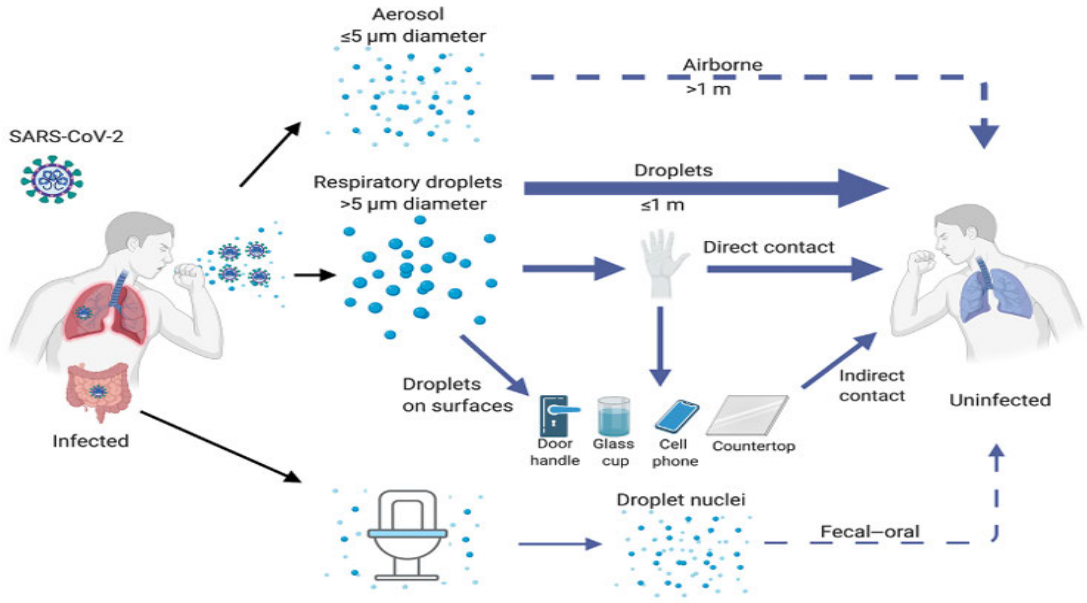


Figure 1.5: SARS-CoV-2 transmission routes potentially impacting the transmission rates and increase in cases worldwide (Harrison, Lin and Wang, 2020)

The reproductive number, often denoted as R_0, for an infectious disease, quantifies the average number of secondary cases generated by a single infected individual in a susceptible population. Early evidence has indicated that the virus has a high reproductive number (R_0 greater than one), indicating its ability to spread rapidly within communities (Billah, Miah and Khan, 2020; Dhungel *et al.*, 2022). Asymptomatic and pre-symptomatic individuals, who may have not exhibited obvious signs of illness,

also contributed to the transmission dynamic. This stealth transmission further complicated efforts to control the spread of the virus (Gao *et al.*, 2021). The global transmission of SARS-CoV-2 was facilitated by international travel across borders (Arora *et al.*, 2021). Studies also show that cities with high population density facilitate increased contact between individuals, amplifying the potential for transmission of airborne diseases (Mei *et al.*, 2015). Hamidi et al. (2020) found that metropolitan areas with tightly linked economic and commuting networks experience higher infection rates, underscoring the role of connectivity in the spread of COVID-19 (Hamidi, Sabouri and Ewing, 2020). Similarly, research by Boterman (2022) demonstrated that urban density influenced infection rates during specific phases of the pandemic in the Netherlands, with urban centers acting as epicenters for viral transmission during periods of relaxed mitigation measures (Boterman, 2023). Close contact environments, including households, healthcare facilities and enclosed spaces with poor ventilation, emerged as hotspots for transmission (Azuma *et al.*, 2020). The understanding of transmission dynamics increased over time as more data became available. It became evident that SARS-CoV-2 was not solely transmitted through large respiratory droplets but could also spread via smaller aerosolised particles that can remain suspended in the air for extended periods (Evans *et al.*, 2021; Wang *et al.*, 2021). This realisation necessitated the adoption of enhanced ventilation, mask-wearing and physical distancing measures to reduce the risk of airborne transmission (Chauhan *et al.*, 2023).

1.2.5 Genetic Variability and Mutation Rate

SARS-CoV-2 exhibits genetic variability and a mutation rate that contribute to its evolutionary dynamics (Hossain *et al.*, 2021). As mentioned earlier, the genome of SARS-CoV-2 is composed of a single-stranded RNA molecule (Figure 1.6) and as with other RNA viruses, it is prone to mutations during replication (Rahimi, Mirzazadeh and Tavakolpour, 2021). Genomic studies have revealed that SARS-CoV-2 exhibits a relatively low mutation rate compared to other RNA viruses (Mercatelli and Giorgi, 2020). In a previous study, the virus was found to have a relatively low mutation rate, with a cumulative count of 12,706 mutations within the SARS-CoV-2 genome (Akkiz, 2021). The study reported that the majority of the mutations were identified as single nucleotide polymorphisms (SNP's). Their sequencing data also revealed that on average, the virus accumulated approximately one to two mutations per month within its genome, which resulted in a considerable number of mutations due to the large number of infections worldwide. Moreover, most alterations in the SARS-CoV-2 genetic material were deemed to be either neutral or slightly harmful, indicating that they do not markedly impact the virus's properties or actions (Akkız, 2022). However, some mutations did lead to changes in the viral proteins, thus potentially influencing factors such as transmissibility, virulence, or immune evasion.

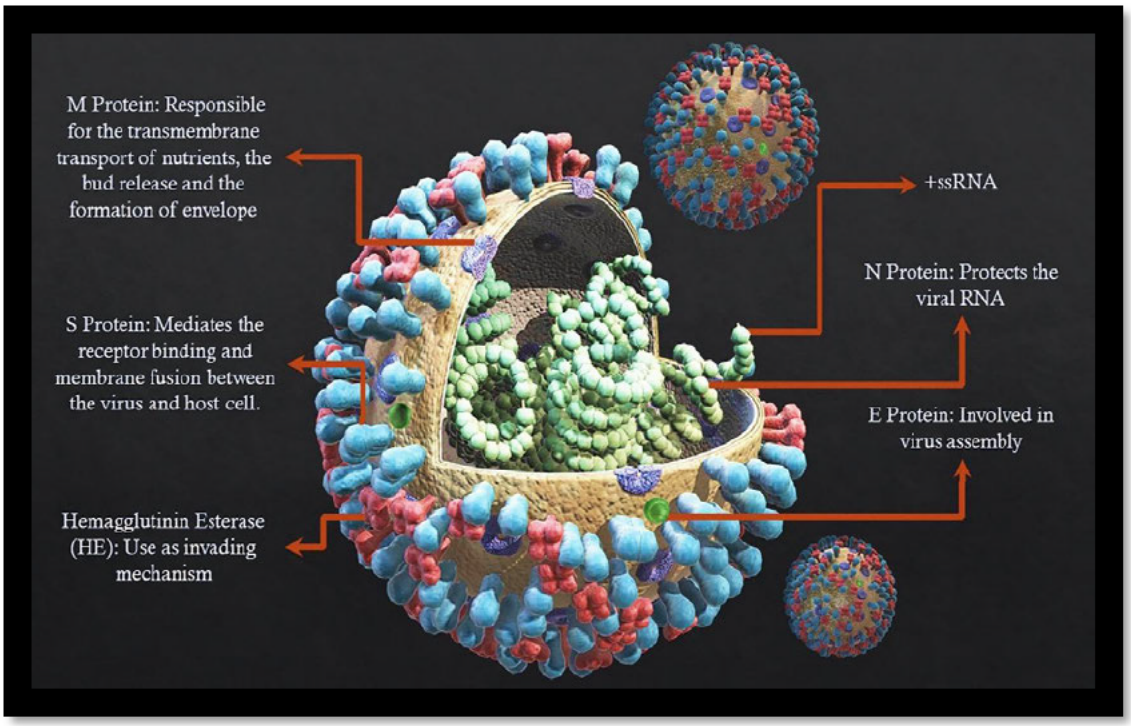


Figure 1.6: SARS-CoV-2 structure and its accessory proteins (Babadi et al., 2022)

In a review paper, the mutation rate of SARS-CoV-2 was estimated to be $1x10^{-6}$ to $2x10^{-6}$ mutations per nucleotide in each replication cycle (Markov *et al.*, 2023). Considering that the virus has a genome size approximating 30,000 nucleotides, this mutation rate translates into a 1 in 50 chance that a replication event will introduce a new mutation (Amicone *et al.*, 2022; Carabelli *et al.*, 2023). This mutation rate falls within the expected range for RNA viruses, which typically have mutation rates in the range of 10^{-3} to 10^{-5} mutations per nucleotide per replication cycle. The relatively lower mutation rate observed in SARS-CoV-2 and other beta coronaviruses may be attributed to their proofreading mechanism (Ortiz-Pineda and Sierra-Torres, 2022). This mechanism helps limit the accumulation of harmful mutations and preserves the stability of the viral genome over time. Furthermore, it is important to note that not all mutations are propagated. Most mutations are deleterious and do not persist beyond their initial appearance (Harvey *et al.*, 2021). Empirical evidence supporting these estimates comes from genomic sequencing of SARS-CoV-2 isolates from different parts of the world (Abbasian *et al.*, 2023). By comparing sequences collected over time, the estimated rate at which mutations accumulate in the viral genome can be determined (Kaushal *et al.*, 2020; Amicone *et al.*, 2022). Such analyses have shown that the mutation rate of SARS-CoV-2 is consistent with the lower end of the spectrum for RNA viruses, corroborating the estimates cited by Amicone et al. (2022) (Markov *et al.*, 2023). Furthermore, these estimates align with observations of other beta coronaviruses and RNA viruses, providing a comparative basis that underscores the relative stability of SARS-CoV-2's genome (Atzrodt *et al.*, 2020).

Stability was critical for the development of effective vaccines and therapeutics, as it suggests that while the virus does evolve, its mutation rate does not compromise the long-term effectiveness of these interventions. The spike protein, which is responsible for facilitating viral entry into host cells, is of particular interest due to its critical role in viral infectivity and its susceptibility to antibody-mediated immune responses (Babadi et al., 2022). Although mutations in the spike protein, such as those found in VOC, exhibits genetic variability, the virus itself demonstrates a high degree of genetic conservation (Martin et al., 2021). This means that the majority of the viral genome remains relatively stable, with only a small proportion undergoing mutations. Conserved regions of the genome are often targeted for diagnostic testing, as they ensure accurate detection of the virus across different variants. This knowledge supports evidence-based decision-making and contributes to our collective efforts to control the spread of such pathogens (Martin *et al.*, 2021).

1.2.6 Recombination as an Evolutionary Driver

In addition to point mutations, SARS-CoV-2 evolution is strongly influenced by recombination, a process in which the viral RNA-dependent RNA polymerase switches templates during replication when two genetically distinct viruses co-infect the same cell (Jackson *et al.*, 2021; Niemi *et al.*, 2021). This can generate chimeric genomes that contain genetic segments from different lineages, potentially introducing multiple changes in functional regions simultaneously. Several documented SARS-CoV-2 lineages, including recombinant forms such as XBB and XD, have emerged through this mechanism (Niemi *et al.*, 2021; Tao *et al.*, 2021). Recombination enables rapid evolutionary jumps that may alter viral properties, including transmission efficiency and immune evasion capacity, making it a key driver of variant diversity, particularly during periods of intense co-circulation of multiple lineages.

1.2.7 Punctuated Evolution and Variant Emergence

The evolutionary trajectory of SARS-CoV-2 has not been linear, but rather characterised by periods of relative genetic stability punctuated by sudden emergence of highly divergent variants (Harvey *et al.*, 2021). This pattern is evident in the transitions from early pandemic lineages to Alpha, Delta, and later Omicron, each of which appeared with an unusually large number of mutations in key genomic regions such as the spike protein (Tegally *et al.*, 2022). These punctuated evolutionary events are thought to be driven by strong selective pressures, such as population-level immunity or chronic infection in immunocompromised hosts, where prolonged viral replication allows accelerated mutational accumulation (Corey *et al.*, 2021; Kemp *et al.*, 2021). The resulting variants often show enhanced transmissibility or immune escape potential, enabling them to rapidly replace circulating strains and reshape epidemic dynamics.

1.2.8 The Impact of VOCs

The evolution of SARS-CoV-2, as previously outlined, has led to the emergence of several VOCs that have drawn considerable attention due to their implications for disease transmission, severity, the efficacy of preventive strategies, the likelihood of reinfection and the overall impact on public health measures. VOCs have exhibited specific genetic mutations that distinguish them from the original strain and, in some cases, from other circulating variants. Such variants have several potential impacts and concerns (Carabelli *et al.*, 2023).

As previously mentioned, one of the primary concerns associated with VOCs is their increased transmissibility. VOCs such as Alpha (B.1.1.7), Beta (B.1.351), Gamma (P.1), Delta (B.1.617.2) and Omicron (B.1.1.529) have demonstrated higher transmission rates compared to earlier variants of SARS-CoV-2 (Ahmad, Fawaz and Aisha, 2022). This heightened transmissibility resulted in an accelerated spread within communities, leading to increased case numbers, increased pressure on healthcare systems and a greater difficulty controlling the pandemic. VOCs have also shown the potential to influence disease severity by being associated with increased hospitalisation rates, greater risk of severe illness and potentially higher mortality rates (Salehi-Vaziri *et al.*, 2022). Such variants can have genetic mutations that affect their susceptibility to existing vaccines. Depending on the extent of the mutations, the effectiveness of current vaccines against the variant may be reduced. However, it is worth noting that vaccines often still provide protection against severe disease, even if they are less effective at preventing infection (Rabaan *et al.*, 2023). The emergence of new VOCs raises the possibility of reinfection in individuals who have previously recovered from COVID-19 or have been vaccinated (Ren *et al.*, 2022). Variants with genetic changes may evade previous immunity, leading to reinfection in some cases. Public health measures enforced by authorities helped control the viral spread. Initial health measures implemented included increased testing, contact tracing, travel restrictions, mask mandates and social distancing guidelines. However, such measures were relaxed and became less restrictive over time (Ayouni *et al.*, 2021; Foo *et al.*, 2022).

Another important consideration is the potential impact of VOCs on diagnostic testing and the efficacy of vaccines and therapeutics. Some variants have exhibited genetic changes in regions of the viral genome targeted by diagnostic tests. This can affect the accuracy of detection, as was observed with the s-gene target failures (SGTF) during the early detection phases of the Beta and Delta variants (Clark *et al.*, 2022; Viana *et al.*, 2022). Additionally, mutations in the spike protein, such as those seen in the Beta and Gamma variants, have raised concerns about potential reduced efficacy of certain monoclonal antibody treatments. However, vaccines developed against the original strain and subsequent variants have generally shown effectiveness in reducing severe illness, hospitalisation and death, albeit with varying degrees of efficacy against the different VOCs (Paul *et al.*, 2023).

From November 2021 to mid-2023, Omicron (B.1.1.529) and its sub-lineages dominated and caused significant disruptions worldwide (Moir and Amoako, 2022). Moir and Amoako (2022) also noted that the ongoing identification of genetically diverse Omicron lineages suggests that persistent human infections and/or animal hosts may be facilitating the virus's continuous evolution and spread. Omicron's circulation led to immune escape, vaccine escape and increased transmissibility and reinfection risk, resulting in its rapid global spread and unprecedented alarm (Chekol Abebe *et al.*, 2022; Dhawan *et al.*, 2022; Viana *et al.*, 2022). Omicron sublineages BA.1, BA.2and BA.3 drove the fourth wave in South Africa, later displaced by BA.4 and BA.5, which drove the fifth wave in early March 2022. As of mid-2023, Omicron sublineages were still being detected, indicating ongoing circulation, although their prevalence and impact varied due to factors such as vaccination coverage, natural immunity and public health measures. The situation with SARS-CoV-2 variants remains dynamic, with new variants potentially emerging and influencing circulation patterns.

1.2.9 Genomic Surveillance: A Key Tool in Understanding the Virus

Genomic surveillance is pivotal in managing infectious diseases, including SARS-CoV-2. Despite advancements in medicine and hygiene, infectious diseases remain a global challenge (Hurd and Nelson, 2009; Buermans and den Dunnen, 2014). Genomic research enhances our understanding, prevention and treatment of various diseases. Recent advances in sequencing technologies have proven essential during outbreaks of Dengue, Chikungunya, Ebola, Zika and COVID-19 (Faria *et al.*, 2016; Stewart-Ibarra *et al.*, 2018; Di Paola *et al.*, 2020; Tegally, Wilkinson, Giovanetti, *et al.*, 2021). In the fight against COVID-19, genomic surveillance involves analysing the virus's genetic structure to understand its evolution, transmission and public health impact (Munung, Mayosi and de Vries, 2018; Wilkinson *et al.*, 2021; Tegally, San, *et al.*, 2022). This approach has provided critical insights into SARS-CoV-2, informing effective public health strategies. Genomic surveillance has elucidated SARS-CoV-2 transmission dynamics and facilitated planning for future waves, particularly with the emergence of novel variants and sublineages.

Next-generation sequencing (NGS) enable rapid, cost-effective sequencing of the viral genome (Charre *et al.*, 2020). Platforms such as Illumina, Ion Torrent and Oxford Nanopore Technology provide comprehensive data on the virus's genetic makeup, revealing mutations, deletions and insertions. These genetic variations help trace transmission chains and identify new strains. Genomic surveillance identified numerous SARS-CoV-2 variants, which have exhibited varying degrees of increased transmissibility or immune evasion (Biancolella *et al.*, 2023). Tracking these variants was essential for adapting public health measures and developing targeted interventions. Additionally, genomic surveillance has illuminated the global distribution and movement of SARS-CoV-2 variants, informing travel restrictions and quarantine measures (Aggarwal *et al.*, 2022). Genomic surveillance also aids vaccine development by assessing the impact of genetic variations on vaccine efficacy. This allows for

timely adjustments of vaccine formulations and the development of booster shots. Collaborative networks and platforms for sharing genomic data enable real-time monitoring and analysis, supporting evidence-based public health decisions (Msomi *et al.*, 2021; Cheng *et al.*, 2023).

1.2.10 Overview of South Africa and SARS-CoV-2 Genomic Surveillance Initiatives

South Africa, a middle-income country located in the southern region of Africa, has the second largest economy in Africa, hosts a diverse population size estimated at over 61 million and is divided into nine provinces. Each province has its own distinct characteristics and contributes to the overall socio-economic landscape of the country. Figure 1.7 lists the provinces of South Africa, which are Eastern Cape, Free State, Gauteng, KwaZulu-Natal, Limpopo, Mpumalanga, Northern Cape, North West and Western Cape (Omotayo *et al.*, 2021; Alexander, 2023).

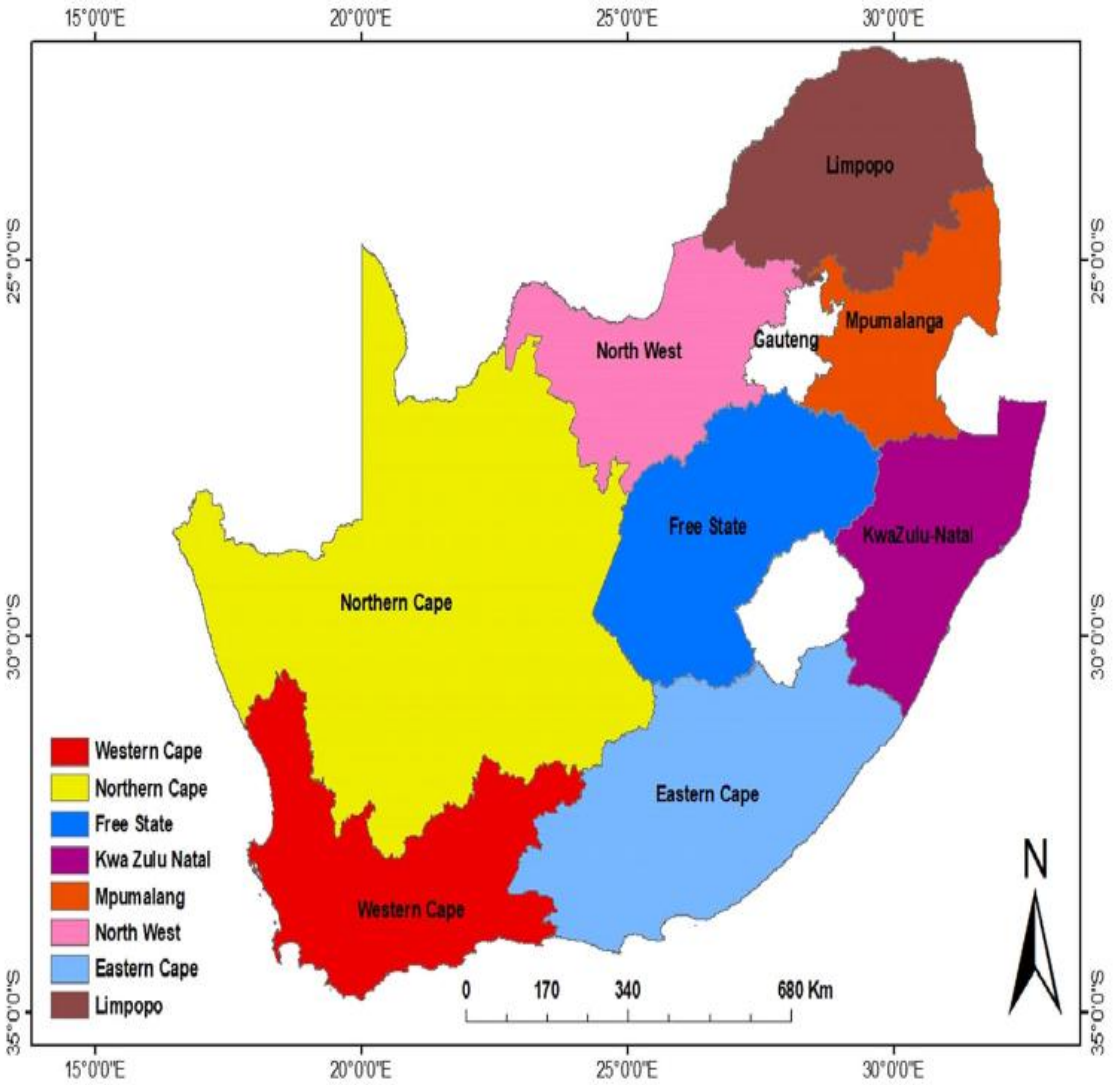


Figure 1.7: Map of South Africa's nine provinces (Omotayo *et al.*, 2021)

1.2.10.1 Provincial Overview

Table 1.1 provides a brief breakdown of the South African provinces in terms of area, estimated population size, the number of metropolitans and districts and COVID-19 statistics per given province.

Table 1.1: Provincial demographic statistics for South Africa, including GISAID submissions for SARS-CoV-2 genomic surveillance

Province	Area (km^2)[#1]	Population Estimates (2021)[#2]	No. of districts (incl. metropolitans)	Covid-19 Cases[#3] (01Feb2023)	Covid-19 Deaths[#3] (19Jan2023)	Total no. of Genomes[*] (24Apr2023)
Eastern Cape	169 966	6,5 million	8	366 907	16 940	3 380
Free State	129 825	2,9 million	5	217 666	7 964	1 983
Gauteng	18 178	15,9 million	5	1 346 712	21 148	12 459
KwaZulu-Natal	94 361	11,7 million	11	728 670	16 297	7 560
Limpopo	125 755	6,1 million	5	160 853	4 678	2 693
Mpumalanga	79 495	4,8 million	3	204 367	4 774	3 112
Northern Cape	372 889	1,3 million	5	115 882	3 249	2 107
North-West	104 882	4,1 million	4	203 844	5 067	2 716
Western Cape	129 462	7,1 million	6	710 754	22 478	11 402
Other	-	-	-	-	-	42
Total	1,22 million	>60 million	52	4 055 655	102 595	47 454*

[#1] **(Alexander, 2023)**
[#2] **(*South Africa: Urbanization from 2012-2022*, 2023)**
[#3] **(*COVID-19 South African Online Portal*, 2023)**
[*] Duplicates removed during data clean up from GISAID (Date accessed 24Apr2023).

In South Africa, comprehensive genomic surveillance efforts were established to monitor the spread and evolution of SARS-CoV-2. Recognising the critical role of genomic sequencing in understanding pandemic dynamics and guiding public health responses, the country implemented a robust system for SARS-CoV-2 genomic monitoring (Giandhari *et al.*, 2021; Msomi *et al.*, 2021). This system proved instrumental in identifying and characterising new variants of concern (VOCs). Researchers

analysed remnant samples from COVID-19 patients to sequence viral genomes, enabling them to pinpoint specific mutations and track their prevalence over time. These efforts were pivotal in detecting novel variants emerging in South Africa (Tegally *et al.*, 2021; Viana *et al.*, 2022). South Africa's genomic surveillance provided vital insights into the genetic diversity and evolution of circulating SARS-CoV-2 strains (Tegally, San, *et al.*, 2022). These findings informed public health strategies, including the design and implementation of targeted interventions to curb the spread of particular variants. Furthermore, by maintaining ongoing surveillance and sharing genomic data with international partners, South Africa advanced global understanding of SARS-CoV-2 and its variants. This collaborative approach facilitated coordinated efforts to mitigate the pandemic's impact worldwide (Engelbrecht *et al.*, 2021; Subramoney *et al.*, 2022).

Over 55,000 genomes from South Africa have been successfully sequenced and made publicly available on databases such GISAID (https://www.gisaid.org/) (Khare *et al.*, 2021)and Sequence Read Archive (SRA, https://www.ncbi.nlm.nih.gov/sra) (*NCBI*, 2021; *GISAID - Initiative*, 2024). These databases continue to accumulate sequencing data from around the world, hosting over 15 million SARS-CoV-2 sequences worldwide. Such sequences enabled researchers to develop therapeutics such as vaccines that target the spike protein and decrease mortality rates. As of August 2024, more than 70,7% of the world's population have received at least one dose of a COVID-19 vaccine, with over 13,7 billion doses administered globally. In South Africa, 38,8% of the national population have received at least one dose of the COVID-19 vaccine, leading to more than 41,8 million doses being administered in the country (Mathieu *et al.*, 2024).

1.2.10.2 The Initial Detection of SARS-CoV-2 in South Africa

The Network for Genomic Surveillance in South Africa (NGS-SA) was established in response to the impact of the COVID-19 pandemic on South Africa (Msomi *et al.*, 2020). This network was a collaborative effort that monitored and characterised the evolution and transmission dynamics of SARS-CoV-2 (Msomi *et al.*, 2020, 2021). This collaborative effort involved various institutions, including the National Institute for Communicable Diseases, KRISP at the University of KwaZulu-Natal and several other universities across the nation (Figure 1.8) ("SARS-CoV-2 Genomic Surveillance Update - September 2024," 2024).

The South African government implemented a range of strategies to control the spread of COVID-19, including extensive testing, contact tracing, public health education, and vaccination campaigns (Majam *et al.*, 2021; Modisenyane *et al.*, 2023). Key findings from several studies on these efforts reveal insights into both successes and challenges faced (Arndt *et al.*, 2020; Lewis *et al.*, 2022; Purasram *et al.*, 2022). These studies collectively highlight that while South Africa's approach involved comprehensive public health interventions, ongoing surveillance, vaccination, and education remain essential in mitigating COVID-19 and preparing for future public health threats.

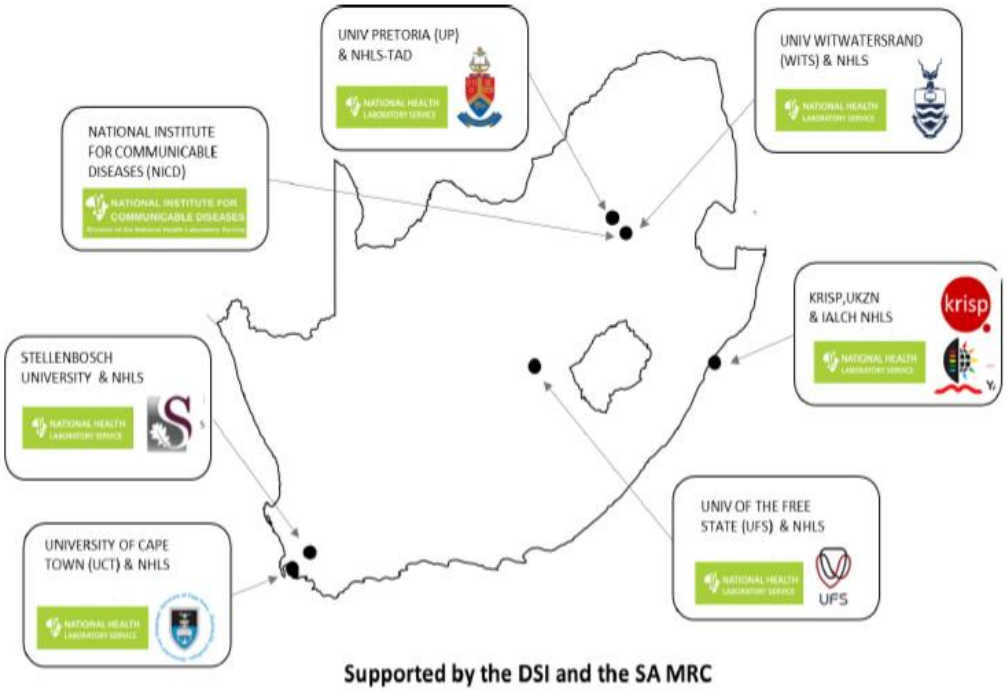


Figure 1.8: NGS-SA consortium of laboratories in South Africa (Msomi *et al.*, 2020; "SARS-CoV-2 Genomic Surveillance Update - September 2024," 2024)

South Africa recorded its first COVID-19 case in early March 2020 and has since experienced multiple infection waves (Figure 1.9). The first wave, which occurred in mid-2020, saw a rapid increase in cases and put a strain on the healthcare system. The government implemented various measures, including lockdowns, social distancing and widespread testing, to slow the spread of the virus (Haider *et al.*, 2020; Durizzo *et al.*, 2021; Schröder *et al.*, 2021). South Africa also faced the challenge of the emergence of new variants of the virus. The Beta variant (B.1.351), first identified in South Africa, became dominant during the second wave in late 2020 and early 2021. This variant was of concern due to its ability to evade some immune responses and its increased transmissibility. However, subsequent research has shown that existing COVID-19 vaccines provided notable protection against severe disease caused by the Beta variant (Chirico *et al.*, 2022; Marta *et al.*, 2022). In late 2021, South Africa reported the identification of the Omicron variant (B.1.1.529). This variant, as previously described, was characterised by a large number of mutations and raised concerns globally due to its potential impact on transmissibility and vaccine effectiveness. Omicron led to international travel restrictions and reinforced efforts to understand its properties (Viana *et al.*, 2022).

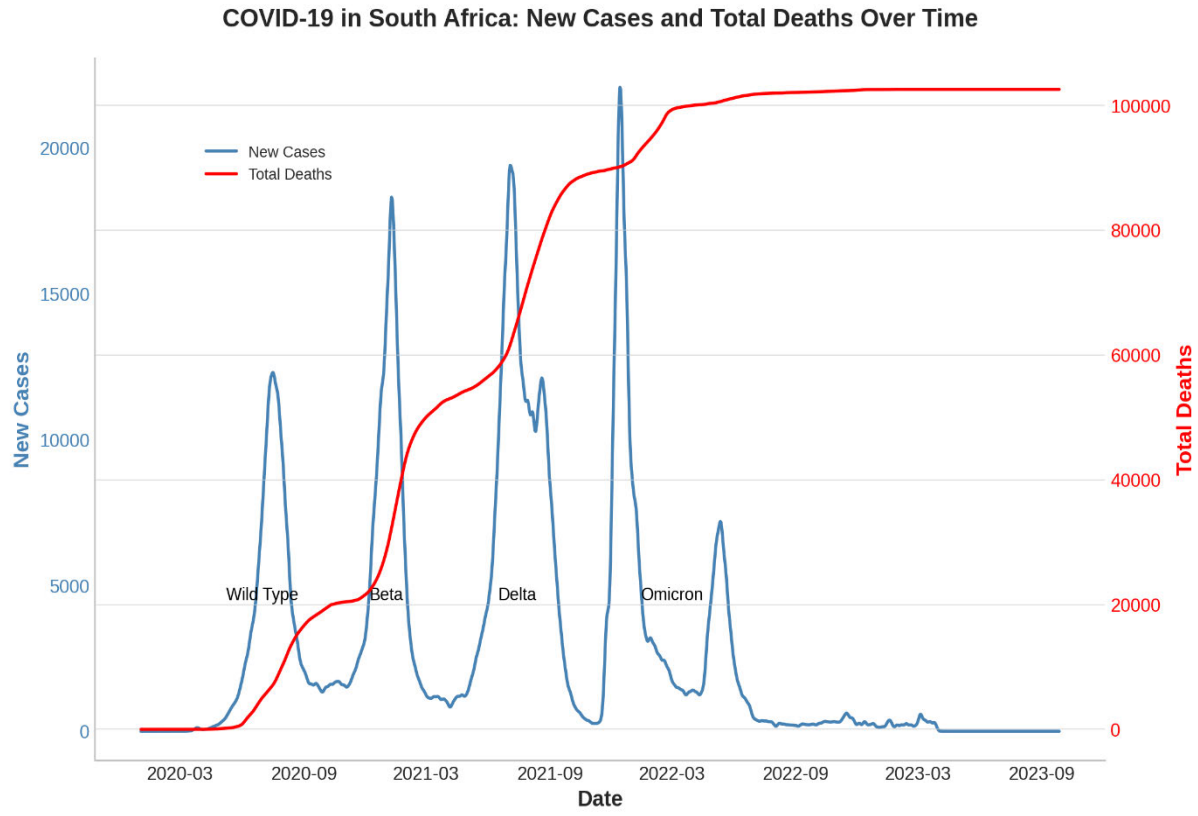


Figure 1.9: Daily confirmed SARS-CoV-2 cases and total deaths since the beginning of the pandemic in South Africa. Adapted from Our World in Data, we highlight the five epidemic waves experienced in South Africa, represented by the recorded daily cases and daily deaths over a 7-day rolling average (Mathieu *et al.*, 2021, 2024).

1.2.10.3Achievements in South African Genomic Surveillance

South Africa made advancements in genomic surveillance, particularly during the COVID-19 pandemic, enhancing both national and global understanding of the virus. The development of genomic sequencing capacity positioned the country as a leader in scientific research and pandemic preparedness (*South Africa National Department of Health*, 2024). The establishment of the NGS-SA provided robust infrastructure for decoding the genetic composition of SARS-CoV-2, improving the ability to track the virus's evolution and tailor public health strategies. South Africa rapidly identified and reported SARS-CoV-2 VOCs, such as Beta and Omicron (World Health Organisation, 2021). Collaborations with key institutions like the Africa Health Research Institute (AHRI) and participation in international genomic databases like GISAID facilitated these efforts (Karim *et al.*, 2020; Cele *et al.*, 2021; Khan *et al.*, 2022). This capability provided crucial insights into the genetic changes of the virus and supported global surveillance and response strategies. South Africa's genomic data has been pivotal in understanding viral transmission dynamics, influencing global pandemic control strategies and advancing vaccine research. By hosting clinical trials for COVID-19 vaccines, South Africa demonstrated its integral role

in evaluating vaccine efficacy and promoting equitable vaccine distribution through partnerships with WHO, Africa CDC and COVAX (*Play Your Part - Brand South Africa*, 2021; World Health Organisation, 2023a, 2025). South Africa's adaptability during the COVID-19 pandemic, including scaling up testing capabilities, training initiatives and utilising digital technologies for contact tracing, showcased the country's resilience and commitment to managing public health crises (Lessells, Moosa and de Oliveira, 2020; Giandhari *et al.*, 2021; de Oliveira, 2022). The COVID Alert South Africa app and extensive testing by the National Health Laboratory Service (NHLS) demonstrate successful strategies in containing the virus (*COVID-19 South African Online Portal*, 2023). These achievements underscored South Africa's proactive approach to genomic surveillance and public health management, establishing a foundation for future pandemic preparedness.

1.2.10.4Challenges, Limitations and Future Directions

Genomic surveillance in South Africa has faced significant challenges and limitations, despite its pioneering role in Africa. The limited infrastructure for genomic sequencing, especially in rural and underserved regions, creates disparities in laboratory capacity and affects the timeliness and geographic coverage of surveillance data (Mfuh, Abanda and Titanji, 2023). This inequity hampers effective pathogen tracking and public health responses (Kluge *et al.*, 2018). Insufficient funding further complicates the situation, as genomic surveillance requires advanced technology and skilled personnel, making it resource-intensive (Gardy and Loman, 2018; Ekusai-Sebatta *et al.*, 2023).

The integration of genomic data with epidemiological data remains suboptimal (Hill *et al.*, 2021). Effective genomic surveillance necessitates the integration of genetic sequences with clinical and epidemiological information to inform public health decisions, but the lack of streamlined data systems and interdepartmental coordination hinders this process (Gim, 2022; Gomes, Rebelo and Alves De Sousa, 2022; Ling-Hu *et al.*, 2022). The availability of skilled personnel is another major challenge. While South Africa has skilled experts, there is a shortage of trained professionals, exacerbated by global demand and brain drain (Govender, 2024). Logistical challenges in sample collection, storage, and transport from remote areas to sequencing labs have long caused delays, undermining the timeliness of data for real-time decisions (Patrick, Chomič and Armstrong, 2016). Despite these challenges, South Africa's genomic surveillance efforts have provided crucial insights into pathogen epidemiology and enhanced global understanding of infectious disease dynamics (Giandhari *et al.*, 2021; Inzaule *et al.*, 2021). Future directions should focus on improving infrastructure by expanding and upgrading sequencing facilities, securing stable funding streams, developing integrated data systems and investing in education and training programmes to build local expertise (Dandara *et al.*, 2014; Inzaule *et al.*, 2021; World Health Organisation, 2024a). Strengthening partnerships with global health organisations and other countries can enhance knowledge exchange, capacity building and access to advanced technologies (World Health Organisation, 2024b). Addressing these challenges through collaborative

efforts will strengthen South Africa's capacity to monitor and respond to infectious diseases effectively, leveraging genomic surveillance as a foundation of public health strategy (de Oliveira and Baxter, 2024).

Despite significant progress in SARS-CoV-2 genomic research, several research gaps remain. Firstly, while numerous studies have characterised the virus's transmission and mutation patterns, fewer have critically assessed the operational, infrastructural, and data-sharing challenges that influence genomic surveillance in resource-limited settings. Secondly, there is limited integration between genomic data and epidemiological metrics, which constrains the ability to evaluate the real-time effectiveness of surveillance frameworks. Thirdly, few comparative analyses have been undertaken to position South Africa's genomic surveillance progress within a global context. This study directly addresses these gaps by combining technical assessments of sequencing performance, regional surveillance trends, and comparative global data analysis to identify areas for improvement and enhance preparedness for future outbreaks.

1.3 Conclusion

This literature review underscores the pivotal role of genomic surveillance in global health, a necessity brought into clear focus by the COVID-19 pandemic. International health agencies, such as the WHO and CDC, have highlighted the importance of robust systems to track and control infectious diseases (Luo et al., 2023). South Africa's genomic surveillance efforts have been instrumental in understanding the dynamics of SARS-CoV-2, particularly through the identification of key variants such as Beta and Omicron (Cele *et al.*, 2021; Tegally *et al.*, 2021; Subramoney *et al.*, 2022). For example, genomic surveillance facilitated the rapid detection of the Omicron variant, enabling effective global tracking of its spread and informing timely public health measures (Viana *et al.*, 2022). Despite challenges, including funding limitations and disparities in laboratory capacity, particularly in rural areas, South Africa has built a surveillance infrastructure that continues to evolve (Tegally, San, *et al.*, 2022). By integrating genomic data with epidemiological and clinical insights, South Africa has effectively guided critical public health responses, such as identifying the immune escape potential of the Omicron variant (Khan *et al.*, 2022; Pulliam *et al.*, 2022). To ensure sustained progress, ongoing investment in workforce training, stable funding, and international collaboration is essential. Strengthening these areas will enhance South Africa's genomic surveillance infrastructure, improve readiness for future outbreaks, and reinforce global health security (Madhi *et al.*, 2022).

Linking Chapters one and two

Interlinking the narratives of Chapters One and Two, we build a framework for understanding SARS-CoV-2 genomic surveillance. Chapter One provides a global and South African overview of the pandemic, the role of genomic surveillance against a backdrop of varying resources, showcasing both achievements and research gaps. Chapter Two advances this discussion by focusing on data generation and sequencing infrastructure used in genomic surveillance. It explores the impact of NGS technologies, particularly their role in enhancing early-stage surveillance in South Africa. This analysis not only details the use of two major platforms but also evaluates their contributions to improving data quality and consistency, crucial for tracking variants like Delta in regions such as the Eastern Cape. This comprehensive synthesis serves as a foundation for further discussions on enhancing genomic surveillance capabilities.

Chapter Two: Reflective Evaluation of Next-Generation Sequencing data during early phase detection of the Delta variant

Abstract

During the SARS-CoV-2 pandemic, next-generation sequencing (NGS) technologies like the Ion Torrent S5 and Illumina MiSeq, alongside advanced software, improved genomic surveillance in South Africa. This study analysed anonymised samples from the Eastern Cape using Genome Detective and NextClade, showing Ion Torrent S5 and Illumina MiSeq success rates of 96% and 94%, respectively. The study focused on genomic coverage (above 80%) and mutation detection (below 100), with the Ion Torrent S5 achieving 99% coverage compared to Illumina MiSeq's 80%, likely due to different primers used in amplification. The Ion Torrent S5 was more effective in sequencing varied viral loads, whereas Illumina MiSeq had difficulties with lower loads. Both platforms were adept at identifying clades, successfully differentiating between Beta (<45%) and Delta variants (<30%), despite minor discrepancies in assignments due to Illumina MiSeq's lower coverage, leading to a failure rate of up to 6%. Manual library preparation showed similar capabilities in sample processing and clade identification for both platforms. However, respective differences between the Ion Torrent and Illumina MiSeq were noted in sequencing duration (3.5 vs. 36 hours), automation level, genomic coverage (80% vs. 99%) and viral load compatibility, showcasing each platform's unique advantages and challenges in SARS-CoV-2 genomic surveillance. In conclusion, the Illumina MiSeq and Ion Torrent S5 platforms are both efficacious in executing whole-genome sequencing (WGS) via amplicons, facilitating precise, accurate and high-throughput examinations of SARS-CoV-2 viral genomes. However, it is important to note the existence of disparities in the quality of data produced by each platform. Each system offers unique benefits and limitations, rendering them viable choices for the genomic surveillance of SARS-CoV-2.

Keywords: Next-generation-Sequencing (NGS), SARS-CoV-2, Illumina MiSeq, Ion Torrent S5, Genomic surveillance, viral load, data analysis.

2.1 Background

At the end of 2019, incidence rates of SARS-CoV-2 increased substantially within a short period, followed by a rapid global diffusion and evolution of the virus resulting in five novel variants of concern (VOC) - each driving new infection waves (Konings *et al.*, 2021; Tay *et al.*, 2022). A VOC, as listed by WHO and Centers of Disease Control and Prevention, is defined as variants that include attributes of increasing transmissibility, causing severe disease and increased hospitalisations, showing a reduction in therapeutics such as vaccine efficacy and diagnostics detection failures (Centers for Disease Control and Prevention, 2023; World Health Organisation, 2023b). Globally, more than 700 million confirmed cases of COVID-19 have been reported as of 05 January 2025, resulting in over 7 million deaths (*Worldometers*, 2025).

Pathogen genome sequencing is a fundamental surveillance tool used to support the understanding of molecular epidemiology of disease outbreaks (Babb de Villiers *et al.*, 2021). Recent advances in sequencing technologies have shown their applicability for research use in outbreak situations, such as those observed with Ebola, Zika virus, SARS, MERS and most recently, the novel SARS-CoV-2 (Faria *et al.*, 2016; Quick *et al.*, 2016; Yin and Wunderink, 2018; Tegally *et al.*, 2020; Butera *et al.*, 2021; Tegally, Wilkinson, Lessells, *et al.*, 2021). Tracking signature genetic mutations of the virus allows researchers to estimate the influence of early outbreaks and ensures more accurate detection and characterisation of variants, possible drug resistance mutations, vaccine escape variants, virulence and pathogenicity factors (Engelbrecht *et al.*, 2021; Giandhari *et al.*, 2021; Tegally, Wilkinson, Giovanetti, *et al.*, 2021; Wilkinson *et al.*, 2021; Khan *et al.*, 2022). Therefore, genomic surveillance, complemented by real-time monitoring and data sharing networks, is a valuable tool to improve the understanding of SARS-CoV-2 transmission and epidemic dynamics. Sequencing centers across the country initiated genomic surveillance programmes for the WGS of SARS-CoV-2 as part of the Network for Genomic Surveillance in South Africa (NGS-SA) (Msomi *et al.*, 2020, 2021). Sequencing technologies used included Illumina, Ion Torrent and Oxford Nanopore technologies. Together, the network kept abreast of the latest SARS-CoV-2 variants circulating within South Africa's infection waves and produced over 46,000 genomes, which were made publicly available in May 2020 (Charre *et al.*, 2020; Pillay *et al.*, 2020; Tshiabuila *et al.*, 2022; *GISAID - Initiative*, 2024).

Although there are distinct differences between these two systems, the Ion Torrent S5 and Illumina MiSeq platforms are well-known for their mid- to high-throughput sequencing, variant calling and overall good quality short read sequences (Quail *et al.*, 2012; Salipante *et al.*, 2014; Lahens *et al.*, 2017; Marine *et al.*, 2020). Illumina uses a fluorescence-based paradigm for determining nucleotide sequence in which all the enzymatic processes and imaging steps take place in a flow cell. Ion Torrent, an alternative sequencing technology, reads nucleotide sequences based on a measure of pH by proton release and makes use of a semiconductor sequencing chip and ion spheres bound to DNA (Merriman, Torrent

and Rothberg, 2012; Lahens *et al.*, 2017; *Ion GeneStudio™ S5 System*, no date). Additionally, there are differences in the type of data generated by each platform. The sequence reads generated from Illumina data in a single run have the same length and are generally paired-end reads, whereas Ion Torrent reads vary in length and are single-ended (Quail *et al.*, 2012; Liu *et al.*, 2016). A comparative study involving SARS-CoV-2 further highlights the ease of use and operation with the automated Ion Torrent S5 workflow when coupled with the Ion Chef (Plitnick *et al.*, 2021). The Ion Chef is used to automate library preparation for small sample numbers and is an essential component in the templating of prepared libraries onto the Ion Torrent sequencing chip. Cost comparisons of these platforms are comparable, assuming that sample multiplexing is increased on the Ion Torrent S5 (Marine *et al.*, 2020). The operation of the Illumina MiSeq at maximum capacity ensures an overall low cost due to the high efficiency of the platform (Marine *et al.*, 2020; Plitnick *et al.*, 2021).

Since the onset of extensive research into SARS-CoV-2 in 2020, concerns have persisted regarding the emergence of new variants and mutations. Consequently, ongoing surveillance of SARS-CoV-2 variants is imperative for the timely identification of emerging variants. For this reason, it is necessary to ensure that sequencing data generated by various WGS platforms is comparable in terms of performance and sequencing output. A recent study performed a benchmarking comparison on several different SARS-CoV-2 genome-sequencing protocols and reported performance variation across WGS technologies (Liu *et al.*, 2021). Another study assessed the WGS of SARS-CoV-2 using both the Ion Torrent and Illumina technologies with their respective protocols in considerable detail (Plitnick *et al.*, 2021). Although the study's findings demonstrate that genomic coverage was high with faster turnaround times on the Ion Torrent platform, it is challenging to compare the data generated without using the same analysis pipeline to prevent discrepancies in assembly and quality control processes.

Therefore, in this comparative study, we used one set of remnant SARS-CoV-2 positive samples collected from the Eastern Cape for a direct comparison of data generated by the Ion Torrent S5 and Illumina MiSeq platforms. The same analysis pipeline was used by both platforms to assess if genomic coverage, quantification of mutations (deletions, insertion sand substitutions) and clade assignment were comparable for data generated across the two platforms.

2.2 Methods

2.2.1 Sample population, collection, processing and ethics

As part of the NGS-SA initiative, we used remnant routine genomic surveillance samples that were collected from the Eastern Cape between March and June 2021 to assess the sequencing data generated from two NGS platforms (Msomi *et al.*, 2020). The National Health Laboratory Service

(NHLS) in Port Elizabeth, Eastern Cape collected nasopharyngeal and oropharyngeal swabs from in-patients and outpatients in clinics and hospitals. The NHLS team also determined SARS-CoV-2 positivity using qualitative polymerase chain reaction (qPCR) assays on the Cobas® SARS-CoV-2 (Roche Molecular, Pleasanton, CA, USA), Xpert Xpress SARS-CoV-2 (Cepheid, CA, USA) or Seegene Allplex™ 2019-nCoV Assay and the CFX96 DX™, Bio-Rad (Seegene, Inqaba Biotec, SA). Sample Ct values were provided as part of the metadata files accompanying all samples. The remnant nasopharyngeal and oropharyngeal swabs from 183 patients were used for this study (Table S1) irrespective of their Ct values. The RNA was extracted at the KwaZulu-Natal Research Innovation and Sequencing Platform (KRISP) based in Durban, KwaZulu-Natal, South Africa. The extracted RNA was used for independent library preparation at the respective sequencing sites followed by sequencing on the Illumina MiSeq based at KRISP and on the Ion Torrent S5 based at the Central Analytical Facilities (CAF) in Stellenbosch, Western Cape. Samples were sequenced over two runs on each platform for a direct comparison of the sequence data generated. This study falls under the approval of the Biomedical Research Ethics Committee of the University of KwaZulu-Natal, South Africa, with protocol reference number BREC/00004745/2022. As no human subjects were involved, no informed consent was required.

2.2.2 Nucleic acid extraction

All 183 samples were extracted as per the manufacturer's instructions using the CMG-1049 kit on the Chemagic 360 instrument (Perkin Elmer, Hamburg, Germany). Total nucleic acid (TNA) extraction was performed using a total volume of 200 µL per sample added to 450 µL lysis buffer and 14 µL Poly A RNA / proteinase-K reaction mixture. The TNA was eluted in 100 µL elution buffer in which two aliquots were made and stored at -20 C until further use.

2.2.3 Complementary DNA (cDNA) synthesis

cDNA synthesis of samples sequenced on the Illumina MiSeq was performed with the SuperScript IV reverse transcriptase using random hexamers (Life Technologies) while cDNA synthesis on samples sequenced on the Ion Torrent S5 was performed with the SuperScript Vilo cDNA synthesis kit (Life Technologies).

2.2.4 Library preparation and next-generation sequencing strategies

Sequence libraries for Illumina MiSeq and Ion Torrent S5 sequencing were manually prepared using the Nextera DNA Flex Library Prep kit and the Ion AmpliSeq Library Kit Plus, respectively. Templating of the prepared libraries onto the sequencing chip for the Ion Torrent S5 was automated using the Ion Chef and the Ion AmpliSeq SARS-CoV-2 Research Panel and Ion AmpliSeq kit for Chef DL8.

2.2.4.1 Multiplex Tiling PCR, Illumina MiSeq Library Preparation and Sequencing

Samples sequenced using Illumina MiSeq were amplified using a multiplex PCR as previously published (Pillay *et al.*, 2020). ARTIC primers were designed on Primal Scheme (http://primal.zibraproject.org/) to generate 400 base pair (bp) amplicons with 70 bp overlaps (Quick *et al.*, 2017). The primers (v3 as of June 2021) were used to amplify the SARS-CoV-2 whole genome (30 kb). Amplicons were purified using Ampure XP purification beads (Beckman Coulter, High Wycombe, UK), using a 1:1 ratio. All purified amplicons were quantified on the Qubit 4.0 instrument using the Qubit double-stranded DNA (dsDNA) High Sensitivity assay kit (Life Technologies). Purified amplicons were stored at 4 °C prior to further use. Indexed paired-end libraries were prepared using the Nextera DNA Flex Library Prep kit and the Nextera DNA CD indexes (Illumina, San Diego, USA) in compliance with the manufacturer's instructions. Libraries were purified and normalised to 4nM prior to the pooling. The pooled library was denatured using 0.2N sodium hydroxide followed by dilution to obtain a final concentration of 8 pM. A minimum of two controls (negative and positive) were included in each sequencing run with 96 samples processed in total. The library was spiked with 1% PhiX Control v3 (adapter-ligated library used as a control) and was sequenced using a 500-cycle v2 MiSeq Reagent Kit on the Illumina MiSeq instrument (Illumina, San Diego, CA, USA) (Pillay *et al.*, 2020).

2.2.4.2 Ion Torrent S5 Library preparation and sequencing

Manual library preparation workflow was performed using the overlapping amplicon strategy and a 2-pool primer panel on the Ion AmpliSeq Library Kit Plus. Primer pool 1 consisted of 125 primer pairs and primer pool 2 consisted of 122 primer pairs generating amplicons within a range of 125 bp to 275 bp in length. It is imperative to note that the panel design allows for the tiling of ~237 amplicons across the SARS-CoV-2 genome (~30kb) resulting in a sequencing coverage of 99,0%, covering positions 43 to position 29,842 (positions relative to the SARS-CoV-2 reference, GenBank accession number NC_045512). An additional five primer pairs targeting human expression were used as controls within the panel. SARS-COV-2 targets were set up using a 16-cycle target amplification for samples with a broad range of viral load. Amplified targets for the two primer pools were combined and ligated using the Ioncode Barcode Adapters 1 – 96 kit. Automated templating of 70 pM libraries was loaded onto two high sequencing data output Ion 540 Chips per sequencing run using the Ion Chef followed by sequencing on the Ion Torrent S5 as per manufacturer's protocol (Ion 540™ Kit – Chef User Guide Pub. No. MAN0010851). A minimum of two controls were included in each run with 96 samples processed per sequence run using two 540 chips. All runs were pre-planned and set up using the Ion Torrent suite software (v5.16.0). All information on the Ion AmpliSeq SARS-CoV-2 Research Panel is available at https://ampliseq.com.

2.2.5 Sequence Data Analysis

The data analysis was processed by one bioinformatician at KRISP and the sequences generated by the Illumina MiSeq were originally analysed prior to those generated on the Ion Torrent S5 due to routine sequencing schedules. The raw paired-end reads generated from Illumina MiSeq and the raw single-end reads from Ion Torrent sequencing (FASTQ files) were assembled using the web-based application Genome Detective, version 1.126 (https://www.genomeDetective.com/) (*Genome Detective Virus Tool*, 2019). Genome Detective is a web-based assembly tool that incorporates the use of de novo and reference-based mapping algorithms to assemble whole viral genomes. The initial assemblies obtained from Genome Detective were refined by aligning mapped reads to the reference and generating consensus sequences for each of the comparisons runs on both sequencing platforms. Consensus sequences for both platforms were assessed using NextClade (https://clades.nextstrain.org/, version 1.7.4) for sequence clade assignment, identification, quantification of mutations and sequence quality analyses. NextClade is a classification tool that utilises Nextstrain nomenclature to distinguish differences between a given sequence and a reference sequence in order to identify various clades and VOCs (Aksamentov *et al.*, 2021). Additional data regarding S-gene coverage was obtained by uploading consensus sequences to Genome Detective. Sequences (FASTA files) that passed quality control with greater than 80% genomic coverage and less than 100 mutations were deposited onto GISAID (https://www.gisaid.org/) (*GISAID - Initiative*, 2024). Internal quality control of 100 mutations was obtained from previous sequencing data using the mutational rate of the virus and the time lapse into the SARS-CoV-2 pandemic. All the raw sequence data generated for this research is available publicly in the NCBI Sequence Read Archive (project no. PRJNA636748). Epidemic data is freely available on the open-source GISAID database and accessed using the GISAID Identifier: EPI_SET_230203ym; doi: https://doi.org/10.55876/gis8.230203ym. The GISAID accession identifiers are also included as part of Table S2.2. For uploading purposes only, the majority of the sequences uploaded onto GISAID were obtained from data generated on the Illumina MiSeq platform, as these were initially sequenced and analysed. Outstanding sequences that passed quality control were obtained from sequencing data generated on the Ion Torrent S5 and uploaded onto GISAID thereafter.

2.2.6 Statistical Evaluation and Considerations

Data visualisation and statistical analysis were performed using ggplot2 v3.3.6 package and R v.4.2. A spearman's ranked sum correlation test was performed to determine the relationship between viral load and coverage (genomic and S-gene) obtained from sequences generated on each platform. The Wilcoxon test was used to establish the difference in the range of genomic coverage obtained between platforms and to assess the difference in the quantification of mutations (total mutations, insertions, deletions and substitutions) detected between the Ion Torrent S5 and the Illumina MiSeq platforms.

2.3 Results

2.3.1 Process Flow

The Illumina MiSeq and Ion Torrent S5 sequencing platforms followed specific workflows for WGS of SARS-CoV-2 samples. Figure 2.1 provides a detailed overview on the processes involved from sample receipt to data analysis. Post nucleic acid extraction, the Illumina MiSeq and Ion Torrent S5 workflows take on a separate direction in sample processing. The Illumina MiSeq and Ion Torrent S5 workflows accommodate manual libraries of 96 samples in total. Illumina MiSeq accommodates 94 samples and two controls per sequence run; Ion Torrent S5 accommodates a minimum of 91 samples and a maximum of five controls per sequence run using two Ion 540 chips.

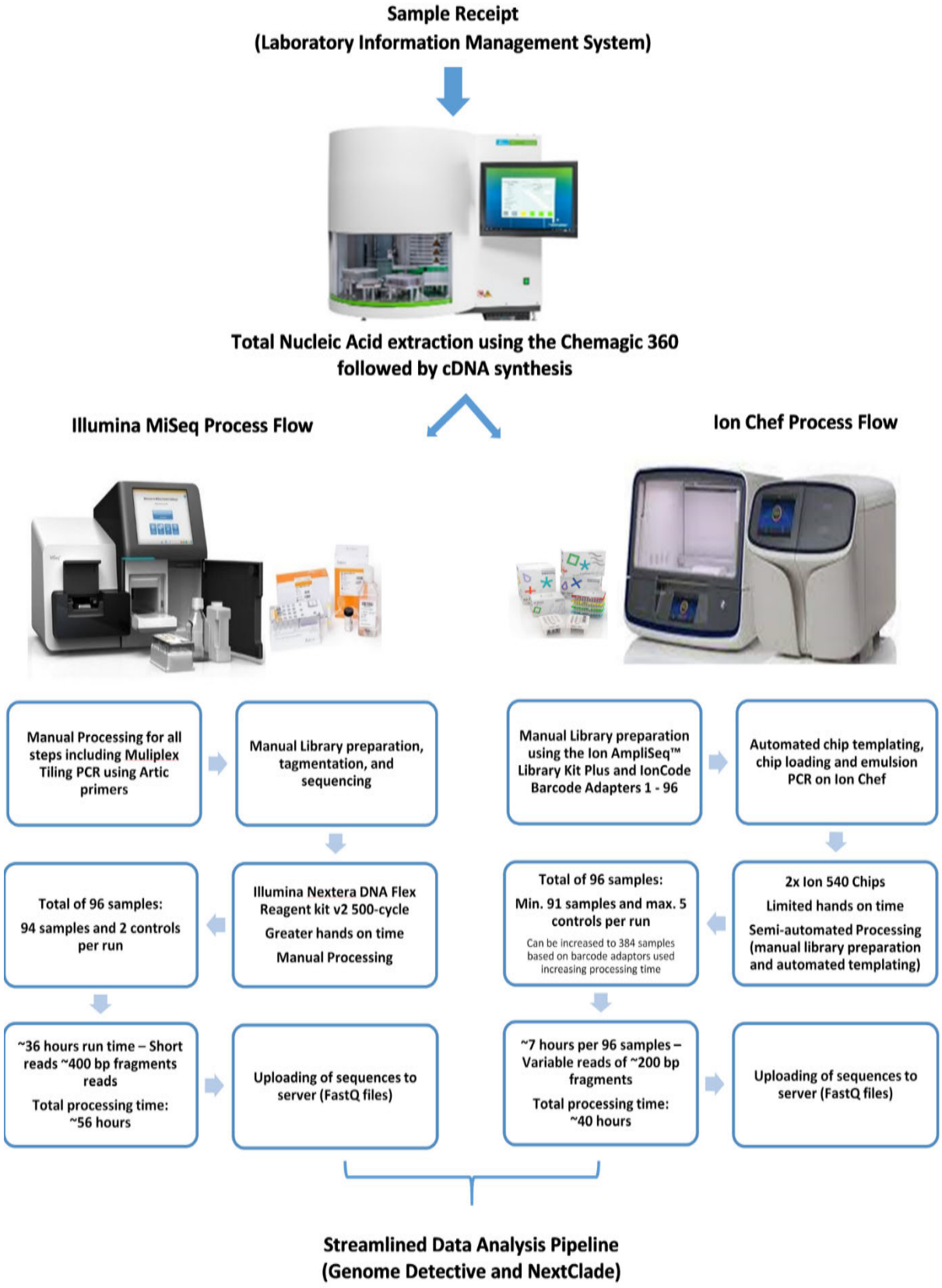


Figure 2.1: Overview of the Next-Generation Sequencing process flow for the whole genome sequencing of SARS-CoV-2 using the Illumina MiSeq and Ion Torrent S5 sequencing platforms

2.3.2 Sequencer-specific attributes

Sequencing data obtained for the independent runs and listed in Tables S3 and S4, were subjected to WGS on the Illumina MiSeq and Ion Torrent S5, respectively. The sequencing data listed in Table 2.1 was obtained for each respective platform upon run completion and was tabulated separately due to non-comparable parameters listed.

Table 2.1: Comparison of (a) Illumina MiSeq and (b) Ion Torrent S5 sequencing platforms used for WGS of 96 SARS-CoV-2 samples.

Attributes	Illumina MiSeq	Ion Torrent S5 and Chef
Primer details	Artic Primers (version 3) – 2 pools	AmpliSeq Primers – 2 pools
Sequencer performance and kit details	Nextera Flex V2 500 cycle kit 24–30 million paired-ended reads 12–15 million single reads Throughput: 8,5 Gb Sequencing Run time: ~ 36 hours	AmpliSeq Research Panel using the Ion 540 Chip 60 – 80 million reads Throughput: 10 - 15 Gb Sequencing Run time: ~3,5 hours per chip
Sample quantity	Processing low to high sample numbers	Processing of low to high sample numbers
Sample processing (automated vs manual)	Can be automated using the addition of an automated liquid handler to reduce hands-on time and error rates Automation allows for handling of large sample numbers High throughput with manual library preparation and indexing of 96 libraries at a time Increased hands-on times and increased error rates	Automation of library preparation on Ion Chef allows for small sample number processing (8 samples per 7,5 hours) Automation limits hands-on time and limits error rates Higher throughput is achievable by preparing libraries manually with the use greater IonCode Barcode Adaptors Manual library preparation increases hands-on time and increases error rates
Read lengths	Assembled Paired-end reads 400 bp fragments	Single-end reads 125 – 275 bp fragments

Each platform was independently assessed based on kit and platform specific outputs as listed in Table 2.1. Each platform displays similarities in the use of 2-pool primer sets, processing capacity of a broad range of sample quantities, semi-automation of process, high sequence success rates and genomic coverage. The platforms differ across the platform-specific kit requirements and primer sets used the performance output in read length, fragment size and varied sequencing duration observed per sequence run.

2.3.3 Sequencing Platform Performance Comparison

Run A and Run B, constituting 183 libraries in total, were sequenced on both platforms, generating 86 consensus sequences for each of the Illumina MiSeq runs, followed by 92 and 83 consensus sequences for the Ion Torrent S5 sequencing runs (Table 2.2).

Table 2.2: Summary of comparable consensus sequence data generated from same sample set on Illumina MiSeq and Ion Torrent S5 platforms

Properties	**Sequencing Platform**	
	Illumina MiSeq	**Ion Torrent S5 and Chef**
Total number of samples processed per platform	183	
Sequencing Run Time per run (hrs)	36	7 (3,5 hrs per 540 chip)
Sequence Success Rate (n/N (%))	172/183 (93,9%)	175/183 (95,6%)
Sequence Failure Rates (n/N (%))	11/183 (6,0%)	8/183 (4,4%)
Greater than 80% genomic coverage (n/N (%))	138/172 (80,2%)	174/175 (99,4%)
Sequences with less than 100 mutations (%)	100%	100%
Mean Genomic Coverage (%)	83,4%	98,9%
Mean S-gene coverage (%)	78,7%	99,0%
Total mutations per platform	7036	8116
Total Substitutions per platform	4671	5039
Total Insertions per platform	17	46
Total Deletions per platform	2348	3031
Paired consensus genomes for both platforms (n/N (%))	164/183 (89,6%)	
Matched clade assignments between platforms (n/N (%))	147/183 (80,3%)	
Mismatched clade assignments between platforms (n/N (%))	17/183 (9,3%)	

The sequence success rate was 93,9% on the Illumina MiSeq and 95,6% on the Ion Torrent S5. The sequencing runtime per run on the Illumina MiSeq was 36 hours, whereas the Ion Torrent S5 platform had a total sequencing runtime of seven hours per 96 samples. Of the 183 samples sequenced, 164 (89,6%) had paired consensus sequences available from both platforms. Table 2.2 combines the samples sequenced in Run A and Run B on both the Illumina MiSeq and the Ion Torrent S5 for a minimum of 91 samples and a maximum of five controls per run. A total of 183 libraries were sequenced on each platform and produced 172 and 175 successful consensus sequences from the Illumina MiSeq and Ion Torrent S5, respectively. Total hands-on time for processing and sequencing is illustrated in Figure 2.1. Consensus genomes from the Illumina MiSeq had a mean coverage of 83,4% with 80,2% having a coverage of 80% and above. Consensus genomes from the Ion Torrent S5 had a mean coverage of 98,9% with 99,4% of sequences having coverage of 80% and above.

2.3.4 Genome Sequence Quality Metrics

The consensus sequences generated on Genome detective were evaluated to determine genome coverage and the quantity of assigned mutations. In total, 99,4% of the genomes that were generated on the Ion Torrent S5 passed the genomic coverage quality metrics for GISAID submissions based on KRISP's internal QC specification compared to 80,2% on the Illumina MiSeq (Table 2.2). All consensus sequences generated from both platforms had less than 100 mutations quantified for each sequence, which are expected for Beta and Delta VOCs.

2.3.5 Genomic and S-gene coverage of SARS-CoV-2

Whole-genome assemblies generated on the Ion Torrent S5 showed a generally higher mean genomic coverage compared to those generated on the Illumina MiSeq (Figure 2.2). Mean genomic coverages of 99,0% and 83,7% was observed on the Ion Torrent S5 and Illumina MiSeq, respectively. A highly statistically significant difference (Wilcoxon, $p<0.0001$) was observed in genomic coverage obtained between the two platforms.

Genomes generated on the Ion Torrent S5 ranged from 63,0% to 100,0% coverage, whereas genomes generated on the Illumina MiSeq ranged from 1,9% to 99,5% coverage. Furthermore, the S-gene coverage ranged from 64,2% to 100,0% on the Ion Torrent S5 and 0,9% to 100,0% on the Illumina MiSeq, with an average S-gene coverages of 99,0% and 78,7%, respectively. A highly significant difference (Wilcoxon, $p<0.0001$) in the S-gene coverage was observed across both platforms (Figure 3). Overall, the genomic and the S-gene coverages were consistently higher on the Ion Torrent S5 platform than on the Illumina MiSeq (Figures 2.2 and 2.3).

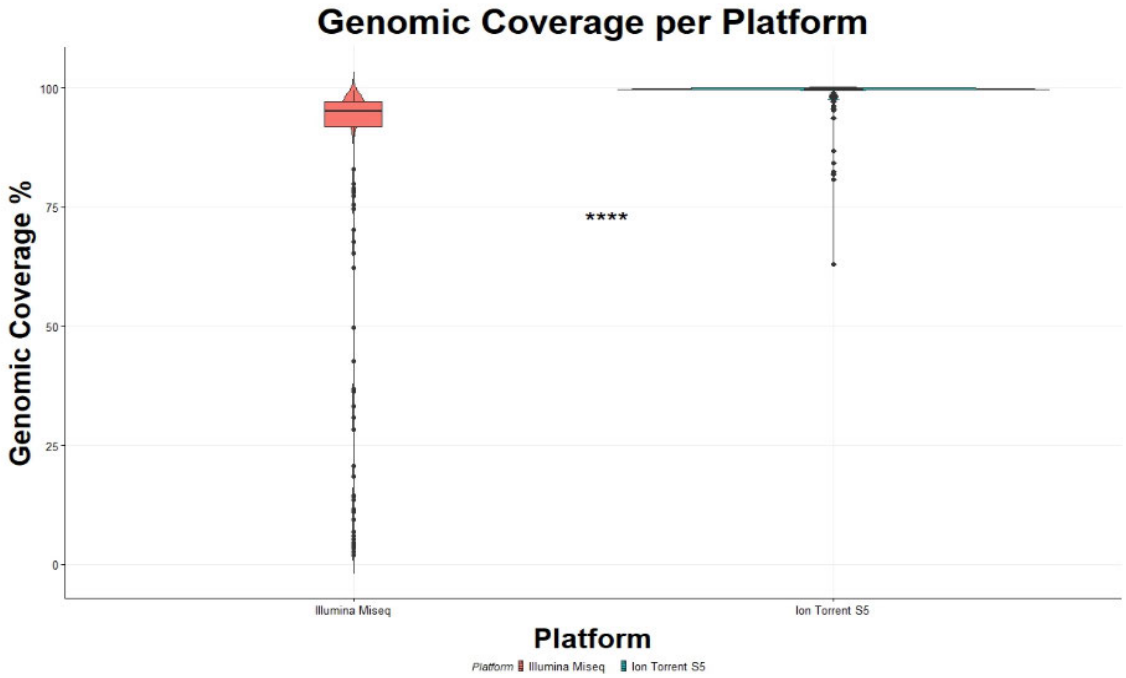


Figure 2.2: Comparison of the genomic sequence coverage for sequences generated on the Illumina MiSeq and Ion Torrent S5 platforms. Statistical comparison was performed using a Wilcoxon rank sum test. Statistical significance (Wilcoxon rank sum tests) was represented by "*" (****: $p < 0.0001$)

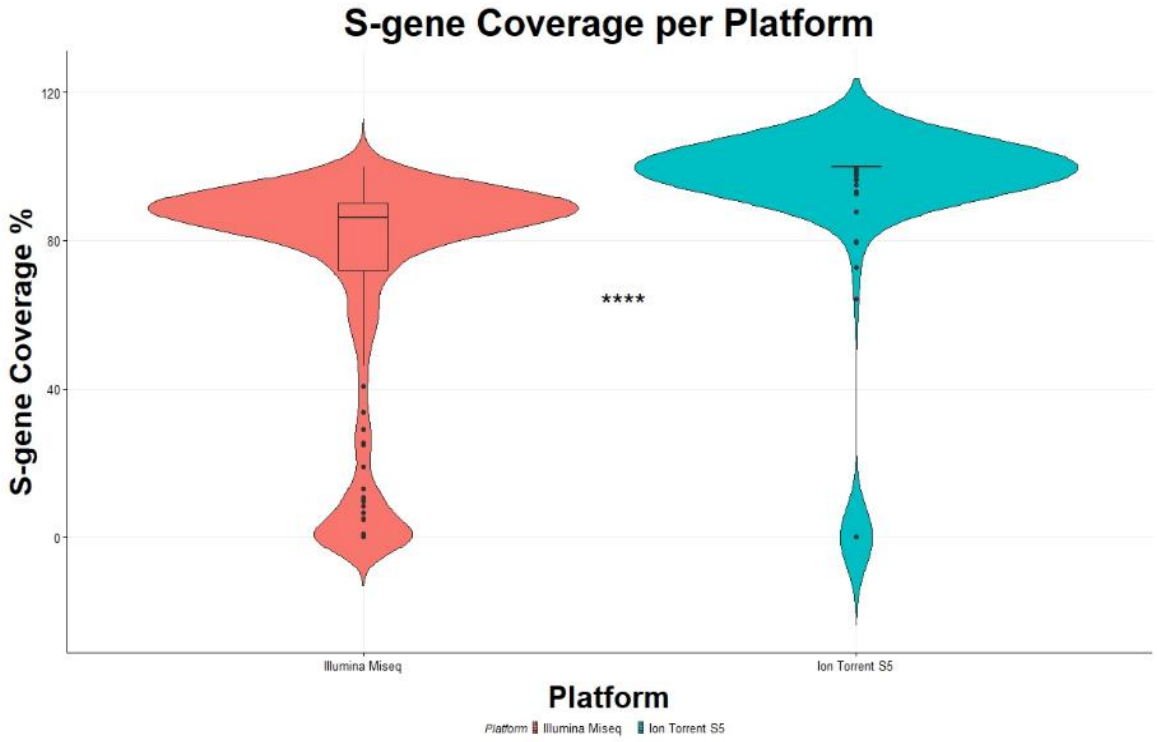


Figure 2.3: Comparison of the S-gene coverage for sequences generated on the Illumina MiSeq and Ion Torrent S5 platforms. Statistical comparison was performed using a Wilcoxon rank sum test. Statistical significance (Wilcoxon rank sum tests) was represented by "*" (****: $p < 0.0001$).

2.3.6 Effect of Viral load on genomic and S-gene coverage of SARS-CoV-2

A Spearman's ranked sum correlation test was performed to determine the effect of increasing viral load on genomic and S-gene coverage on sequences generated from each sequencing platform (Figures 2.4 and 2.5).

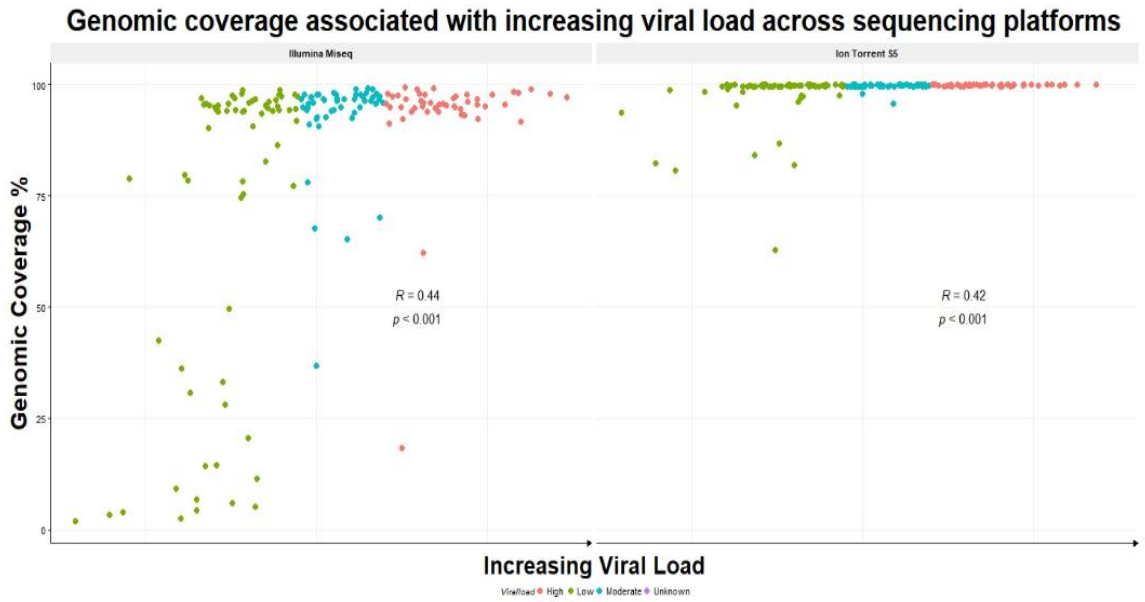


Figure 2.4: Comparison of the influence on viral load and resulting genome coverage produced on the Illumina MiSeq and Ion Torrent S5 platforms.

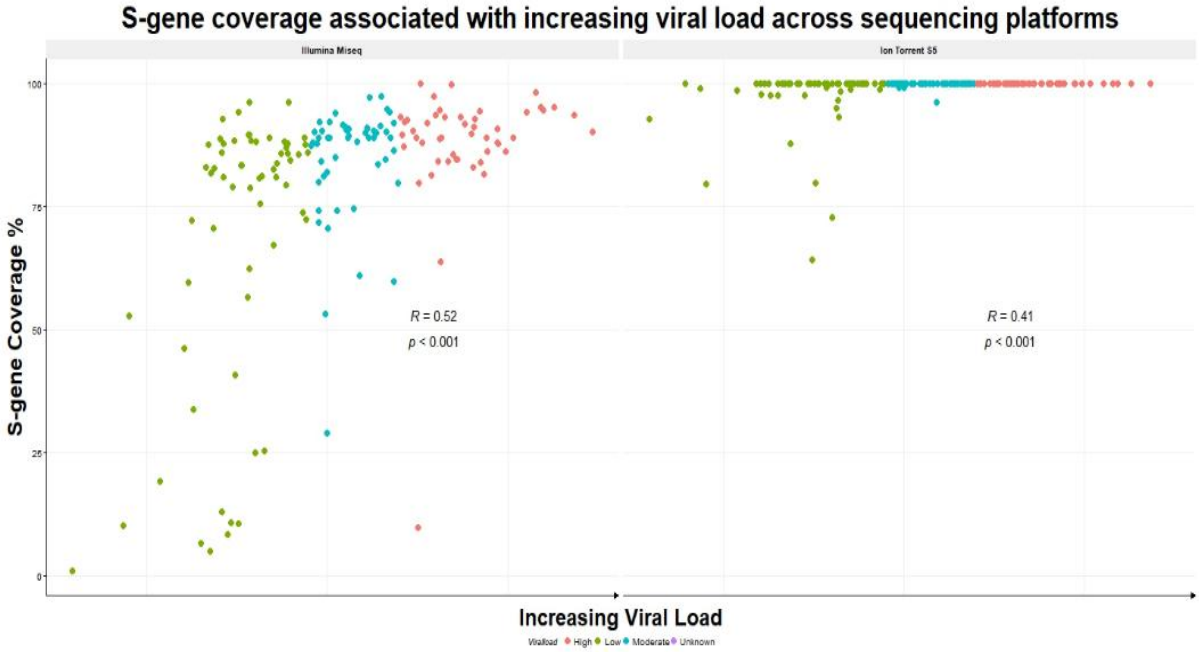


Figure 2.5: Comparison of the influence on viral load and resulting spike gene coverage produced on the Illumina MiSeq and Ion Torrent S5 platforms.

Of the 183 samples sequenced on each platform, 180 sequences had corresponding Ct scores available. Estimated viral loads were qualitatively based on mean Ct values provided with available metadata and grouped as per Table 2.3. Sequences generated from samples with mean Ct values ≤ 25 (high viral loads) accounted for 28,4% of the total consensus genomes, followed by 29,5% with mean Ct > 25 and ≤ 30 (moderate viral loads) and 40,4% with mean Ct values > 30 (low viral load). High statistical significance obtained for sequences generated on the Illumina MiSeq ($p<0.001$) and Ion Torrent S5 ($p<0.001$) showed that estimated viral load directly influences genomic coverage. A similar trend was also observed for S-gene coverage in association with viral load. In addition, several samples sequenced on the Illumina MiSeq resulted in an absence of coverage in the S-gene region (Table S2.5). These sequences occurred in samples having very low viral load and low template material and therefore obtained low genomic coverages.

Table 2.3: Mean Ct score range for viral load estimation

Ct score (mean)	Sample No. (n (%))	Estimated Viral Load (Qualitative)	Mean Genomic Coverage (Ion Torrent S5 / Illumina MiSeq)
Ct ≤ 25	52 (28,4%)	High	99,9 / 93,7
25 < Ct ≤ 30	54 (29,5%)	Moderate	99,7 / 93,0
Ct > 30	74 (40,4%)	Low	97,7 / 71,4
No Ct score available	3 (1,6%)	Unknown	99,3 / 39,0

Mean genomic coverages were also tabulated in Table 2.3 for each of the estimated viral loads assigned. Although consensus sequences generated on the Ion Torrent S5 had higher overall mean genomic coverages compared to the Illumina MiSeq, genomic coverage is known to gradually decline with decreasing viral load.

The range of Ct values with estimated the viral loads used is summarised in Table 2.3. Mean Ct values less than and equal to 25 were assigned as having high viral loads and accounted for 28,4% of the total consensus genomes, followed by 29,5% with mean Ct values greater than 25 and less than and equal to 30 having moderate viral loads and 40,4% with mean Ct values greater than 30 having a low viral load. No Ct values were provided for three samples so the viral load was not estimated and was listed as unknown. The mean genomic coverage obtained for each viral load estimate is also provided to illustrate the decrease in coverage obtained with decreasing viral load (template material).

2.3.7 SARS-CoV-2 Clade assignments

The consensus sequences were uploaded onto NextClade and the clade assignments were determined and compared between sequences generated on the two platforms. Clade assignments obtained from sequences run on the Ion Torrent S5 and Illumina MiSeq were comparable across most of the clades identified. As illustrated in Table 2.4, Beta and Delta variants were identified in 41,5% and 30,0% of samples sequenced on the Ion Torrent S5 followed by 44,8% and 29,5% on the Illumina MiSeq, respectively.

Table 2.4: Clade assignment summary between sequencing platforms

Clade	**Ion Torrent S5 (*n* (%))**	**Illumina MiSeq (*n* (%))**
19A	14 (7,7%)	7 (3,8%)
19B	-	2 (1,1%)
20A	8 (4,4%)	4 (2,2%)
20B	13 (7,1%)	13 (7,1%)
20C	7 (3,8%)	8 (4,4%)
20H (Beta,V2)	76 (41,5%)	82 (44,8%)
20I (Alpha,V1)	2 (1,1%)	2 (1,1%)
21A (Delta)	1 (0,5%)	1 (0,5%)
21J (Delta)	54 (29,5%)	53 (29,0%)
Blanks	8 (4,4%)	11 (6,0%)

The number of samples that failed sequencing (eight and 11 on the Ion Torrent S5 and Illumina MiSeq, respectively) were not assigned a clade and are detailed in Table 2.5.

Table 2.5: Samples that failed sequencing on the NGS platforms

		Ion Torrent S5		Illumina MiSeq	
Accession Identifiers	**Mean Ct (VL)**	**Genomic Coverage (%)**	**Clade Assignment**	**Genomic Coverage (%)**	**Clade Assignment**
EPI_ISL_3275349	26,8 (Moderate)	99,8	20H (Beta, V2)	-	-
EPI_ISL_3275351	22,4 (High)	99,8	20B	-	-
EPI_ISL_3275352	24,4 (High)	99,8	21J (Delta)	-	-
EPI_ISL_3275353	33,0 (Low)	99,8	21J (Delta)	-	-
EPI_ISL_3275354	28,2 (Moderate)	99,7	21J (Delta)	-	-
-	33,1 (Low)	99,7	21A (Delta)	-	-
EPI_ISL_3275374	35,7 (Low)	100,0	20H (Beta, V2)	-	-
EPI_ISL_3275376	34,1 (Low)	99,9	20C	-	-
EPI_ISL_3275378	19,2 (High)	100,0	20C	-	-
EPI_ISL_2727188	18,1 (High)	-	-	91,7	20A
EPI_ISL_3275379	33,1 (Low)	100,0	20H (Beta, V2)	-	-
EPI_ISL_2727191	33,7 (Low)	97,1	19A	-	-
EPI_ISL_2727197	22,6 (High)	-	-	95,8	20H (Beta, V2)
EPI_ISL_2727205	35,8 (Low)	-	-	95,4	20H (Beta, V2)
EPI_ISL_2727214	32,3 (Low)	-	-	86,4	20H (Beta, V2)
EPI_ISL_2727215	25,7 (High)	-	-	94,9	20H (Beta, V2)
EPI_ISL_2727222	30,8 (Low)	-	-	95,0	20H (Beta, V2)
-	36,5 (Low)	-	-	14,2	20H (Beta, V2)
EPI_ISL_2727233	15,3 (High)	-	-	97,3	21J (Delta)
Sampled unsuccessfully sequenced		8		11	

Of the 183 consensus genomes compared, 17 sequences were classified as different clades between the NGS platforms (Table 2.6). We highlighted the mismatched clades assigned to each sequence generated per NGS platform as well as the genomic and S-gene coverages obtained for each, followed by the respective amino acid mutations identified. It is evident from the data obtained that genomic coverage influenced the clade classification of the variants sequenced by each platform. Table 2.6 is for illustration purposes only and draws attention to the high coverage obtained for sequences on the Ion Torrent S5 that had contributed to a reliable clade assignment of variants in relation to Illumina MiSeq data. Specific key mutations for each of the mismatched clade assignments were also examined with a focus on specific key mutations in the spike region. It was observed that not all specific key mutations were identified for each respective clade assigned. An inconsistency in mutations identified were observed in sequences between the platforms, that is, each sequence identified had a different set of mutations called. This discrepancy may be attributed to several sequencing factors such as primers used, number of reads obtained as well as platform-specific processing. Table 2.6 highlights the 17 samples, which were sequenced on both the Illumina MiSeq and Ion Torrent S5 but were classified as different clades by Nexclade online tool. Clades identified by the Illumina MiSeq include 20H Beta V2 (n=6), 20A (n=1), 20B (n=1), 20C (n=3), 19A (n=2), 19B (n=2) and 21J Delta (n=2). Clades identified by the Ion Torrent S5 include 20H Beta V2 (n=3); 20A (n=6); and 19A (n=8)). An overall higher genomic coverage was observed for sequences generated on the Ion Torrent S5 than on the Illumina MiSeq. Amino acid mutations identified by both platforms for a given sample were highlighted in green and S-gene specific key mutations for the respective clade assignment were listed in bold for each sample sequenced per platform. Furthermore, it was observed that a few sequences generated on the Illumina MiSeq obtained no coverage for the S-gene region of the SARS-CoV-2 genome.

Table 2.6: Clade calling discrepancy between the Illumina MiSeq and Ion Torrent S5 platforms with their corresponding mutations identified (mutations in bold signify key S-gene mutations specific to the VOC identified)

Sample ID	Run:	Coverage (%) per Platform				Clade Calling per Platform		Key Mutations	
		Illumina MiSeq		Ion Torrent S5		Illumina MiSeq	Ion Torrent S5	Illumina MiSeq	Ion Torrent
		Genomic Coverage %	S-gene Coverage %	Genomic Coverage %	S-gene Coverage %				
K016691	A	18,5	9,7	99,8	100,0	19A	20H (Beta, V2)	ORF7b:T40I	E:P71L, N:P13S, N:T205I, ORF1a:T265I, ORF1a:K1655N, RF1a:K3353R, ORF1b:P314L, ORF1b:A1057S, ORF3a:Q57H, ORF3a:S171L, S:L18F, S:D80A, S:D138Y, S:D215G, S:R246G, S:**K417N**, S:**E484K**, S:**N501Y,** S:D614G, S:A701V
K016709	A	78,1	71,7	99,7	100,0	20H (Beta, V2)	20A	E:P71L,N:T205I, ORF1a:T265I, ORF1a:G507R, ORF1a:P2046L, ORF1a:N2596S, ORF1a:M2796C, ORF1a:K3353R, ORF1b:P314L, ORF1b:I1074V, ORF3a:S171L, S:L18F, S:D80A, S:D215G, S:**K417N**, S:**E484K**, S:**N501Y**, S:D614G, S:A701V, S:S940F, S:A1087S	ORF1b:P314L, S:D80A, S:**N501Y**, S:D614G
K016715	A	11,0	4,8	99,7	100,0	20C	20A	ORF1a:T265I, ORF1a:K3353R, ORF1a:D4335Y, ORF1b:P314L	S:N501Y, S:D614G
K016754	A	36,3	46,2	99,0	97,7	21J (Delta)	19A	M:I82T, N:D63G, ORF1b:E513*, ORF7b:T40I, ORF9b:T60A, S:T19R, S:**L452R**, S:**T478K**, S:D614G, S:D950N	M:I82T, N:D63G, N:D377Y, ORF1a:T3255I, ORF1b:G662S, ORF1b:A1918V, ORF7a:V82A, ORF7b:T40I, ORF9b:T60A, S:**L452R**, S:D614G, **S:P681R**
K016760	A	28,2	10,8	99,7	100,0	19A	20A	ORF1a:K3353R	M:I82T, N:D63G, N:G215C, ORF1a:A1306S, ORF1a:P2046L, ORF1a:P2287S, ORF1a:T3255I, ORF1b:P314L, ORF1b:G662S, ORF1b:P1000L, ORF1b:A1918V, ORF3a:S26L, ORF7a:V82A, ORF9b:T60A, S:T19R, S:T478K, S:N501Y, S:D614G
K016870	B	11,5	25,4	97,5	98,3	20H (Beta, V2)	19A	ORF1a:K1197N, ORF1a:T1638I, ORF1a:P1640S, ORF1a:K1655N, S:**K417N**	ORF3a:S26L, ORF3a:Q57H, ORF3a:S171L, S:D614G, S:A701V
K016899	B	13,5	***0,0***	98,1	98,6	20C	19A	ORF1b:R1315C	E:P71L, S:A701V
K016902	B	42,7	19,1	98,4	98.6	20H (Beta, V2)	19A	N:G60F, N:K61P, N:E62S, N:D63S, N:K65P, N:R68L, N:G69L, N:I74L, N:D81Y, N:D82Y, ORF1a:T265I, ORF1a:T2154I, ORF1a:N2767H, ORF1a:K3353R, ORF1b:V22I, ORF1b:C44S, ORF1b:L271I, ORF9b:E65S, ORF9b:D66Y, ORF9b:Q70H, ORF9b:Q77H, ORF9b:M78I, S:A701V	ORF1a:T265I; S:D614G

K016908	B	30,8	33,7	95,4	97,6	20B	20A	N:A152S,N:R203K, N:G204R, N:N213I, ORF1a:T395I, ORF9b:R32L, S:D614G	ORF1b:S1779I, S:N450K, S:D614G, S:P681R
K016917	B	77,3	73,7	97,7	98,8	20H (Beta, V2)	19A	E:P71L, N:D128Y, N:T205I, N:Y268N, ORF1a:K1655N, ORF1a:K3353R, ORF1b:S1182L, ORF3a:W131L, ORF3a:S171L, ORF8:D63N, S:L18F, S:D215G, S:**K417N**, **S:E484K**, **S:N501Y**, S:D614G	N:D128Y, ORF1a:K3353R, ORF3a:Q57H, S:L18F, S:**K417N**, S:D614G, S:A701V
K016920	B	33,1	8,3	99,7	97,6	20C	19A	ORF1a:T265I, ORF1a:T2087S, ORF1a:K3353R, S:D614G	N:T205I S:D614G
K016923	B	5,9	10,6	86,8	79,7	20A	20H (Beta, V2)	S:**K417N**	E:P71L, M:I82T, N:T135I, N:T205I, ORF1a:T265I, ORF1a:I547F, ORF1a:K1655N, ORF1b:P314L, ORF1b:L1698F, ORF3a:Q57H, ORF3a:G100S, ORF3a:S171L, S:T19A, S:L24F, S:P25T, S:K182N, S:D215G
K016926	B	3,9	10,2	98,9	99,0	19B	20A	S:D614G	M:I82T, ORF1a:P2046L, ORF1a:S2048F, ORF1a:T3255I, ORF1b:G662S, ORF1b:P1000L, ORF1b:A1918V, S:D614G
K016931	B	9,3	***0,0***	99,7	100,0	21A (Delta)	20H (Beta, V2)	M:I82T, N:D63G, ORF9b:T60A	E:P71L; N:T205I, ORF1a:T265I,ORF1a:K1655N, ORF1a:K3353R, ORF3a:Q57H, ORF3a:S171L, S:**K417N, S:E484K, S:N501Y**, S:D614G, S:A701V
K016940	B	4,4	6,6	98,5	97,6	20H (Beta, V2)	19A		ORF3a:Q57H, RF3a:S171L, S:D614G, S:A701V
K016943	B	5,2	***0,0***	97,7	93,1	19B	20A	ORF1a:A540T, ORF1a:K1655N	S:D614G
K016945	B	79,9	59,6	99,9	100,0	20H (Beta, V2)	19A	E:P71L, N:T205I, N:T271I, ORF1a:K1655N, ORF1a:K3353R, ORF1b:P314L, ORF1b:A941S, ORF1b:G1129V, ORF3a:G100S, ORF3a:S171L, S:D215G, **S:K417N**, S:D614G, S:A701V	S:D614G
					Clade Summary per Platform (total = 17)			**20H Beta V2 (n=6), 20A (n=1), 20B (n=1), 20C (n=3), 19A (n=2), 19B (n=2) 21J Delta (n=2)**	**20H Beta V2 (n=3); 20A (n=6); 19A (n=8)**

2.3.8 Quantification of mutations (insertions, deletions, and substitutions) for sequences generated on the Illumina MiSeq and Ion Torrent S5

The number of mutations detected for each sample was individually compared to establish if there was a significant difference in the assignment of mutations. These analyses included the total

substitutions, insertions and deletions, as detected by the Ion Torrent S5 and Illumina MiSeq platforms (Figure 2.6).

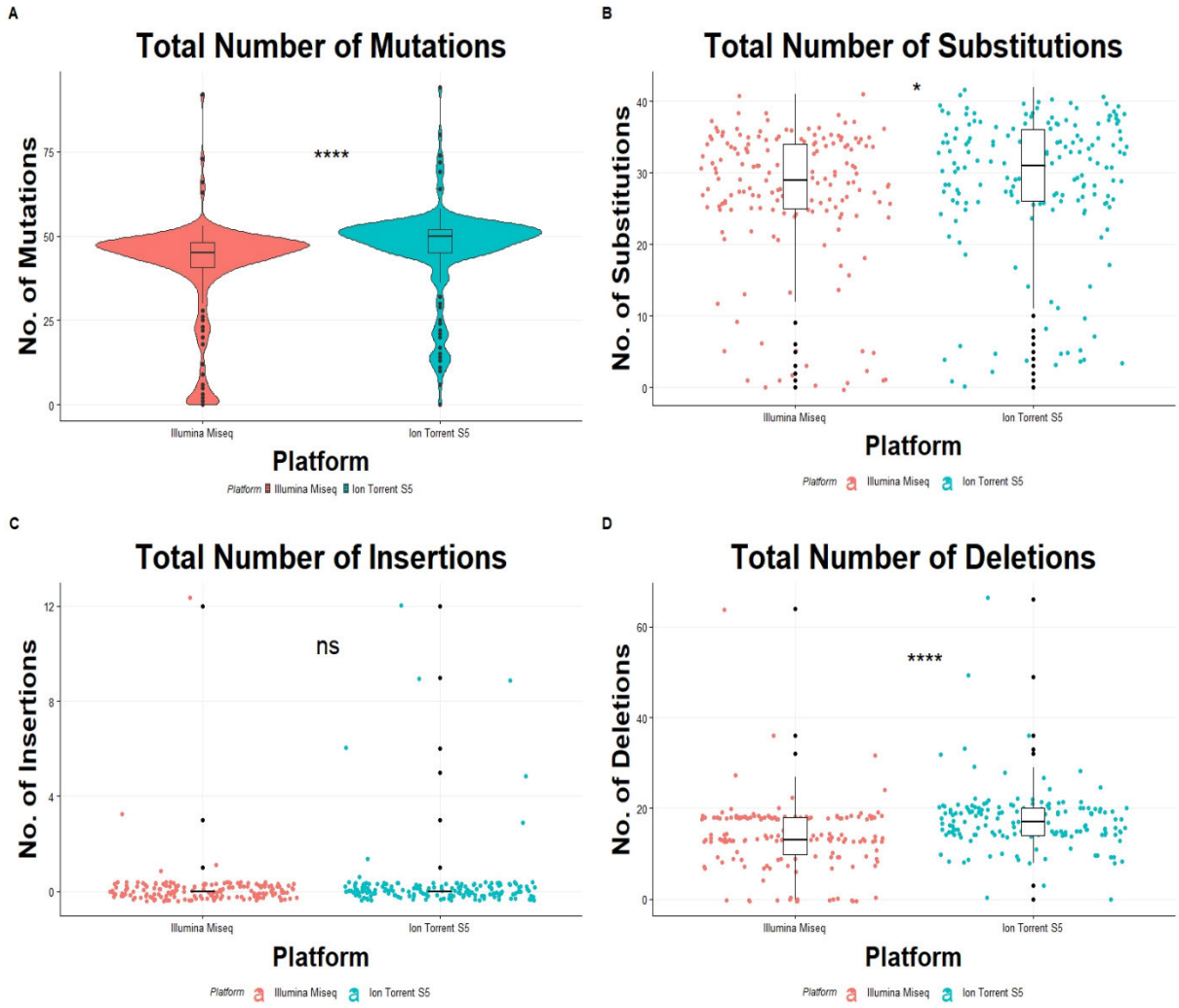


Figure 2.6: Mutational analysis of all sequences from runs A and B generated on the Illumina MiSeq and Ion Torrent S5 platforms

Consensus sequences were produced using Genome Detective and uploaded to NextClade for quantification of mutations. A comparison in the number and type of mutations detected by each platform was performed using a Wilcoxon rank sum test and statistical significance was represented by "*" (ns: non-significant, *: $p < 0.05$, ****: $p < 0.0001$). The total number of mutations (A), substitutions (B), insertions (C)and deletions (D) on the Illumina MiSeq and Ion Torrent S5 platforms for the combined two runs are illustrated.

In total, we analysed 347 consensus sequences. There was a highly significant difference in total mutations observed across the Ion Torrent S5 and Illumina MiSeq (Wilcoxon, $p<0.0001$) with a greater number of mutations detected by the Ion Torrent S5 (6 – 94 mutations, total: 8116) than the Illumina MiSeq (1 - 92 mutations, total: 7036). A marked significant difference was also noted for the

number of substitutions (Wilcoxon, $p<0.05$) and deletions (Wilcoxon, $p<0.0001$) identified by both platforms; however, no significant difference was observed for insertions (Wilcoxon, p=0,25 ns) across both platforms. The variation in the number of mutations detected across the sequencing platforms were listed in Table 2.2.

2.4 Discussion

The Ion Torrent S5 and Illumina MiSeq provide alternative methods for researchers to study SARS-CoV-2 at a genomic level (Plitnick *et al.*, 2021). This study compared the performance and data generated by the two WGS platforms. We hypothesised the sequencing platforms that were used for the genomic surveillance of SARS-CoV-2 in a high throughput laboratory setting, generated sequencing data that differed when analysed with the same bioinformatic pipeline. In a brief overview of the data generated for genomic coverage, clade assignments and quantification of mutations, we concluded that the platforms were similar in sequencing capabilities but differed in sequencing data outcomes. Our findings indicate that the Ion Torrent S5 produced sequences with higher genomic coverage over a broader range of viral loads and in a shorter time in comparison to the Illumina MiSeq. These findings were in agreement with previous comparison studies (Lahens *et al.*, 2017; Marine *et al.*, 2020; Plitnick *et al.*, 2021; Rachiglio *et al.*, 2021).

In terms of the sequencing process, the Ion Torrent S5 and Illumina MiSeq followed a streamlined process flow, which allowed the platforms to display their adaptability in the WGS of SARS-CoV-2. The sequencing runtime for a sample set of 96 on the Illumina MiSeq was 36 hours, whereas on the Ion Torrent S5 it was seven hours using two Ion 540 sequencing chips. The difference in sequencing time allows a greater number of samples to be sequenced on the Ion Torrent S5 in a 36-hour period than on the Illumina MiSeq. Although the sequencing duration is much shorter on the Ion Torrent S5, it is important to note that the remaining processes, such as amplification and library preparation, consume a shorter duration on the Illumina MiSeq. The automated process for templating the manually prepared libraries onto the Ion Torrent sequencing chip is approximately 15,5 hours, whereas the amplification and tagmentation step utilise less than 10 hours. These findings confirmed previous observations that found similarities in processing times with each respective workflow (Plitnick *et al.*, 2021). With limited hands-on time, full automation with the Ion Chef allows for faster turnaround times but limits sample numbers that can be processed at once (eight libraries per seven and a half hours on the Ion Chef) on the Ion Torrent S5. This makes the Ion Torrents selling point of full automation its major downfall in a high-throughput laboratory setting. However, taking into consideration the manual route for library preparation, the Ion Torrent S5 is similar to the Illumina MiSeq in handling larger sample numbers, provided a manual library preparation is followed. The Illumina MiSeq workflow can also be automated by including an external liquid handler to further reduce hands-on time and overall duration of processing. In addition, reagents used in the upstream preparation processes on the Illumina

MiSeq can be further optimised and validated to accommodate greater sample numbers with reduced reagent volumes by miniaturisation of process workflow (Pillay *et al.*, 2023). In contrast, the use of full automation on the Ion Torrent S5 coupled with the Ion Chef allows for limited handling of small sample numbers. It increases the overall turnaround time in a day-shift laboratory site. It is therefore feasible to incorporate a manual library preparation for such platforms to minimise turnaround times and increase sample numbers processed in a high throughput laboratory setting as illustrated in this study.

The same remnant sample set of 183 was used to limit variability between samples and ensured comparison of the data generated during analysis from the two platforms. The sequence process had a direct impact on the sequencing outcomes in relation to genomic coverage and sequence quality metrics. Sequence quality was based on in-house quality control specifications established at KRISP for GISAID submissions. These included sequences which had greater than 80% genomic coverage and less than 100 mutations. Although the Ion Torrent S5 and Illumina MiSeq are both capable of producing complete SARS-CoV-2 genomes, sequences generated on the Ion Torrent S5 maintained an overall higher mean genomic coverage in comparison to sequences generated on the Illumina MiSeq. Various factors contribute to the genomic coverage obtained from both platforms. Ct values are semi-quantitative numbers that generally categorise the concentration of viral RNA in a given sample following qPCR testing. An inverse correlation is observed between viral load and Ct values. Low Ct scores are associated with high viral loads and were found to influence sample quality and, therefore, overall sequence quality (Rabaan *et al.*, 2021; Zuckerman *et al.*, 2021). Echoing previous findings, we observed an association of viral load on genomic coverage for all sequences generated. Moderate to low viral load samples sequenced on the Ion Torrent S5 resulted in an overall good mean genomic coverage (>60%) resulting in higher success rates and increased test eligibility. These findings imply that the Ion Torrent S5 sequencing capabilities are less likely to be affected by sample Ct values and, therefore, can be employed in sequencing of samples during early stages of infection when viral load is lower. However, further investigation may be required to assess this using a larger sample cohort across various laboratories. In contrast, the Illumina MiSeq relied on samples with higher viral load for better coverage of genomes as observed in other studies (Charre *et al.*, 2020; Pillay *et al.*, 2020; Tshiabuila *et al.*, 2022).

Additionally, the increase in genomic coverage obtained from the Ion Torrent S5 may be attributed to the greater number of reads obtained per sample with the use of two Ion 540 chips in a sequencing run of 96 samples (Marine *et al.*, 2020). As highlighted in Table 1, the AmpliSeq Research Panel on Ion Torrent S5 produces over twice the number of reads of the Illumina MiSeq and half the size of fragments sequenced compared to the Illumina MiSeq (200bp versus 400bp) (Thermo Fisher Scientific, 2021). In essence, the number of reads achievable for the samples sequenced in this dataset would be at least double on the Ion Torrent S5 to obtain a coverage that is higher or equal to the Illumina MiSeq. It is possible that the greater number of reads obtained per sample could also have contributed

to the greater coverage and reliability in clade assignments observed in sequences generated on the Ion Torrent S5.

Furthermore, the Ion Torrent S5 detected a significantly greater number of total mutations (insertions, substitutions and deletions) than the Illumina MiSeq. Previous studies have reported that unlike the Illumina MiSeq, semiconductor sequencing platforms like the Ion Torrent S5 are known to produce a predominated homopolymer-associated base-call error by means of INDELS (Loman *et al.*, 2012; Laehnemann, Borkhardt and McHardy, 2016; Marine *et al.*, 2020). Interestingly, these were often deletions instead of insertions, similar to findings of this study, which may have contributed to the larger number of total mutations from Ion Torrent S5 sequences. According to Marine et al., 2020, while such INDELS may be adjusted and corrected for in well characterised viruses, this may not be the case when characterising novel viruses (Marine *et al.*, 2020). This further validates the need for in-depth quality control parameters during analysis of such sequences.

In using Genome Detective as the prime assembly method in generating consensus genomes for both platforms, we eliminate the variability between other assembly methods. Additionally, the advantage of using NextClade allows the user to determine the difference in quality of the consensus sequences generated, to classify clades accordingly and to establish similarity in the identification of evolutionary changes between sequences from each platform (Aksamentov *et al.*, 2021). In contrast to our findings above highlighting higher genomic coverage obtained for sequences generated on the Ion Torrent S5, we find the majority of these sequences to be grouped as lower quality on NextClade in comparison to those generated on the Illumina MiSeq (data not shown). This, however, can be attributed to sequencing errors or miscalled bases generated on the Ion Torrent systems as previously observed (Bragg *et al.*, 2013; Laehnemann, Borkhardt and McHardy, 2016; Quick *et al.*, 2016; Marine *et al.*, 2020). Furthermore, other findings indicate that the Ion Torrent S5 and Illumina MiSeq sequences can easily differentiate between the Beta and Delta VOCs based on mutation calling and respective clade assignment. NextClade assigned the same clades for 147/183 (80,3%) samples during the early delta-replacing-beta phase observed in the Eastern Cape, South Africa. A mismatch in clade assignment was observed in 17/183 (9,3%) samples successfully sequenced on both platforms followed by a low failure rate of 4,4% and 6,0% on the Ion Torrent S5 and Illumina MiSeq, respectively. It would be interesting to expand the clade classification and sequencing across a larger cohort of known VOCs using both platforms to assess the reliability and accuracy in clade assignment of NextClade. The Pangolin lineage assignment tool is an alternative software for lineage classification; however, it was not included in this study (Rambaut *et al.*, 2020).

A major limitation and consideration for genomic surveillance laboratories is the use of updated primer sets. It is important to note that unlike the Illumina MiSeq Artic V3 primers, which were found to be problematic with novel variants such as Delta, the Ion Torrent primers (AmpliSeq Primers) covered 99% of the SARS-CoV-2 genome, therefore attributing to higher genomic and S-gene coverage

(Alessandrini *et al.*, 2020; Davis *et al.*, 2021; Kuchinski *et al.*, 2022). Due to the consistent evolution of SARS-CoV-2, difficulty in mutational regions arose resulting in poor coverage of some Artic V3 primers that were located in regions having the key Delta mutations (Davis *et al.*, 2021). Since the initial primers were designed and based on the reference SARS-CoV-2 genome sequence, it was expected that there would be difficulty in identifying large structural variants. As a result, systematic limitations were observed in the presence of high levels of genomic variation. Subsequently, several S-gene target failures (SGTF) were observed during diagnostic qPCR testing for COVID-19 (Challen *et al.*, 2021; Vogels *et al.*, 2021; Wolter *et al.*, 2022). A decreased coverage of specific regions within the SARS-CoV-2 genome was also observed with the emergence of novel variants since the beginning of the pandemic (Vogels *et al.*, 2021). Low coverage sequences generated on the Illumina MiSeq may have contributed to the discrepancy in clade assignment during the initial surge of the Delta variant. It was previously reported that the G142D amino acid substitution was substantially underrepresented among early Delta variant genomes identified (Davis *et al.*, 2021). Furthermore, Kuchinski et al., 2022, reported a disruption in genomic sequencing of SARS-CoV-2 as a result of emerging mutations identified in novel variants (Kuchinski *et al.*, 2022). Since the ARTIC primer set is one of the most widely used SARS-CoV-2 sequencing primers, the V3 primers were updated to address the amplicon drop-off observed among the Delta variant of concern, resulting in a version 4 being released in June 2021. Unfortunately, V4 primers were not used during the execution of the study as they were not procured. Lambisia *et al.*, 2022, subsequently conducted a study to assess the impact of the updated V4 Artic primers on genome recovery using the ONT and concluded a great improvement in the recovery of the Delta variant amongst others (A. W. Lambisia *et al.*, 2022). The Ion Torrent sequencing panel was also updated to accommodate the amplicon drop-off in novel variants. The updated panel, Ion AmpliSeq SARS-CoV-2 Insight Research Assay, was designed to improve the coverage and uniformity of the previous Ion AmpliSeq SARS-CoV-2 Research Panel used in this study. Continuous improvement of current primers irrespective of kit specifications is, therefore, an essential requirement of an effective genomic surveillance regime.

Plitnick et al., 2021, directly compared the performance of the SARS-CoV-2 AmpliSeq Research Panel to the results obtained with the Illumina MiSeq-based ARTIC Nextflow analysis pipeline (Plitnick *et al.*, 2021). Post-bioinformatic analysis of data from such studies showed that both methods produced similar levels of coverage (>98%) across a broad range of viral loads (Ct values of 15.56 to 32.54 [median, 22.18]) and that both approaches sequenced SARS-CoV-2 effectively (Plitnick *et al.*, 2021; Rachiglio *et al.*, 2021). Although the bioinformatic analysis pipelines used in this study differ, the findings of our study are similar to those documented in the above study by Plitnick et al., 2021. Standardisation of analysis regimes accommodates the comparison of data from different NGS technologies without bias from independent assembly and analysis tools. Assembly software affects the overall genomic coverage of sequences obtained from various platforms. There is an additional need for quality control processes to improve the overall quality of such sequences made publicly available

as recommended by Jacot et al., 2021 for diagnostic purposes (Jacot *et al.*, 2021). Such achievements have included removal of frame shifts and unknown stop codons in some instances. In this study, we observed sequences generated on the Illumina MiSeq to be simpler to process, with quality control easily implemented across such sequences yielding sequences of better quality as per NextClade analysis. Although sequencing capabilities are similar on both platforms, an overall higher genomic coverage of sequences was generated on the Ion Torrent S5. However, the majority of these sequences were found to be of lower quality as per analysis on NextClade. It is, therefore, important to note that standardising assembly and analysis software allows for improved comparison of the data generated by the different platforms and data analysed by different software. Nonetheless, this study complements previous research investigating the efficacy of Ion Torrent and Illumina platforms for sequencing viral pathogens (Quail *et al.*, 2012; Salipante *et al.*, 2014; Szargut *et al.*, 2022).

In light of these discoveries, it is of the utmost significance to remember that although samples were collected during the period spanning mid-March to early June 2021, the commencement of sequencing occurred only in mid-June 2021, upon the receipt of these samples. This observation underscores a challenge in real-time genomic surveillance monitoring, given that the Delta wave in South Africa was officially reported from 17 May 2021 to 14 November 2021 (Tegally, Wilkinson, Althaus, *et al.*, 2021). Nonetheless, the surveillance samples collected in this study on 29 March 2021 revealed an earlier presence of the Delta variant in the Eastern Cape region. It is, therefore, necessary to acknowledge that the delayed initiation of sequencing for novel emerging variants can exert a direct influence on the transmission dynamics within communities and provinces. The implications arising from such challenges may hinder the timely reporting and monitoring of emergent variants, potentially leading to delays in implementing appropriate public health measures to mitigate their spread and impact (Msomi *et al.*, 2020; Boehm *et al.*, 2021).

This study offers valuable knowledge about the needs and difficulties involved in using various platforms for genomic surveillance in a high throughput research laboratory as well as the challenges encountered in such a programme. When using publicly available sequences, users should be cautious and consider the technologies, assembly and analysis pipelines and the quality of sequences from different samples. This is especially important when comparing data from different platforms. It is essential for users to consistently update and maintain the primers specific to each technology, particularly with the emergence of new variants. This practice plays a critical role in accurately monitoring existing circulating variants and promptly detecting emerging ones, facilitating real-time genomic surveillance with precision.

While more advanced statistical models (such as regression or mixed-effects analyses) could further interrogate associations between viral load and sequencing performance, these were not applied here due to the retrospective and exploratory nature of the dataset, as well as variability in sampling and run conditions. Instead, descriptive and non-parametric analyses were intentionally employed to

provide a balanced comparative assessment that accurately reflects real-world laboratory performance during the early phase of Delta variant detection.

2.5 Conclusions

Genomic surveillance played a crucial role in mitigating the spread of SARS-CoV-2, aiding in the rapid detection of new mutations within the Delta VOC (Tegally, Wilkinson, Giovanetti, *et al.*, 2021; Wilkinson *et al.*, 2021). Both the Ion Torrent S5 and Illumina MiSeq platforms were capable of accurately distinguishing between the Beta and Delta VOCs. Notably, the Ion Torrent S5 demonstrated superior performance in processing samples with lower viral concentrations (indicated by higher Ct values) compared to the Illumina MiSeq, though further analysis is warranted. Regarding the capabilities of these sequencers, including genomic coverage, the production of high-quality sequences and the overall data output, both the Illumina MiSeq and Ion Torrent S5 were deemed suitable for whole-genome sequencing (WGS) of SARS-CoV-2. Nonetheless, distinctions in data quality exist between the two. Consequently, this study's findings highlight that, despite their individual strengths and weaknesses, the Ion Torrent S5 and Illumina MiSeq, in conjunction with analytical tools like Genome Detective and NextClade, stood as dependable options for SARS-CoV-2 genomic surveillance.

Linking Chapters two and three

Chapter Two addresses the complexities of NGS technologies and data analysis in genomic surveillance, noting the challenges of data variability when different sequencing technologies are used. Chapter three provides an examination of genomic surveillance data at the provincial level in KZN. By concentrating on KZN's unique demographic and geographic diversity, this chapter aims to uncover gaps and enhance surveillance strategies, thereby improving future epidemic responses. The insights gained will not only benefit KZN but also provide a valuable model for other regions globally, advancing the effectiveness of genomic surveillance.

Chapter Three: Retrospective Evaluation of SARS-CoV-2 Genomic Surveillance Data from KwaZulu-Natal, 2020 to 2023

Abstract

KwaZulu-Natal (KZN), South Africa, experienced distinct waves of SARS-CoV-2 driven by diverse variants, underscoring the need for robust genomic surveillance to guide public health interventions. This study evaluates genomic surveillance strategies in KZN by analysing district-specific data trends to identify gaps, challenges, and limitations within the current framework. A retrospective analysis of 6,089 genomic sequences collected between March 2020 and March 2023 was conducted using sequences and metadata from the GISAID database. Trends in circulating lineages, operational parameters, and district-level disparities were evaluated. The analysis identified 134 SARS-CoV-2 lineages circulating across the 11 districts of KwaZulu-Natal between March 2020 and March 2023. These included the Alpha, Beta, and Delta variants of concern, with Omicron sub-variants becoming dominant in the later stages of the study period. Demographic data showed higher case incidence among females and individuals aged 19–30. The EThekwini district experienced higher case numbers and greater viral diversity due to urbanization and population growth. Disparities in demographic sequencing coverage were identified, particularly in rural areas, along with limited testing in certain districts. Operational trends highlighted variability in turnaround times and metadata completeness, affecting the overall efficiency of surveillance efforts. This study highlights the need for equitable genomic monitoring and localised public health strategies. Insights from KZN's genomic surveillance may guide the understanding of regional epidemiology, offering a framework to improve surveillance systems and prepare for future health crises. By combining genomic and epidemiological data, the research underscores the vital role of genomic epidemiology in guiding timely, informed public health responses.

Keywords: COVID-19, SARS-CoV-2, genomic surveillance, KwaZulu-Natal, demographic trends, healthcare disparities, urbanisation, urban vs. rural transmission, pandemic preparedness.

3.1 Background

The COVID-19 pandemic, caused by the novel coronavirus SARS-CoV-2, was first identified in December 2019 in Wuhan, China. The crisis rapidly escalated into a global one. South Africa reported its first confirmed case on 5 March, 2020 (National Institute for Communicable Diseases, 2020). Since then, the country has experienced multiple waves of infection, with new viral variants emerging. As of August 2023, there have been over 4 million confirmed cases and around 102,000 deaths attributed to the virus in South Africa (*Worldometers*, 2025). These waves were driven by genetic VOCs that showed changes in transmissibility and the ability to escape the immune system. Genomic surveillance has played a crucial role in tracking the evolution and spread of SARS-CoV-2. By sequencing viral genomes from diagnosed cases, researchers have been able to monitor the emergence of mutations and identify variants that contribute to changes in the epidemic. This genomic data has provided insights into the transmission pathways and epidemiological links of the virus and has guided public health interventions.

In South Africa, the national COVID-19 genomic surveillance network has identified several key variants during the pandemic. The Beta variant (B.1.351), first detected in the Eastern Cape in October 2020 (Tegally *et al.*, 2021), dominated the second wave due to its increased transmissibility. It was followed by the Delta variant (B.1.617.2), which caused the third wave (Tegally *et al.*, 2021)and then the Omicron variant (B.1.1.529), first reported in November 2021, which led to the fourth wave (Viana *et al.*, 2022). Each wave had specific genomic characteristics that correlated with patterns of transmission intensity and vaccine efficacy. Despite extensive national genomic surveillance, there are still gaps in understanding the full range of factors that influence the virus's dynamics. Factors like population density, regional connectivity and local public health responses play critical roles but require detailed regional analysis to fully understand their impact.

KwaZulu-Natal (KZN), situated in the southeast of South Africa, provides a unique setting for an in-depth analysis of genomic surveillance data. As the province with the second-highest population after Gauteng, KZN displays a diverse demographic with both urban and rural communities (Statistics South Africa, 2020). This province has been extensively sampled throughout the pandemic, yielding a rich dataset for retrospective analysis (*GISAID - Initiative*, 2024). KZN's distinctive epidemiological and sociodemographic profiles render it an exemplary model for examining the interactions between viral genomic trends and regional factors influencing the spread and control of infectious diseases. The province has been notably impacted by the HIV epidemic, with a prevalence rate of 27,0%, the highest in South Africa in 2019 (Kharsany *et al.*, 2019). The analysis focuses on KZN at the provincial level, a shift from previous studies that typically considered national data, allowing for a more nuanced exploration of the region's genomic surveillance capabilities and shortfalls. KZN's metadata, particularly at the district level, provides a baseline for evaluating the efficacy of the existing surveillance system. The comprehensive baseline helps identify specific local challenges, which is

essential for refining public health strategies. Notably, KZN was the location of the first detected case of SARS-CoV-2 in South Africa (Giandhari *et al.*, 2021). With a population of 11,7 million, KZN has recorded 18% of the national SARS-CoV-2 cases and 16% of the deaths reported to date, highlighting its epidemiological contribution (*Worldometers*, 2025). Despite extensive national epidemiological studies, a data gap remains. Detailed information on viral diversity in KZN's eThekwini Metropolitan Municipality is limited. The same applies to its ten districts, which range from partly rural to semi-urban. This gap has been evident, especially since the second wave (Tegally *et al.*, 2021). These districts, which vary in socio-economic conditions, may exhibit diverse patterns of disease transmission and may show differences in genomic surveillance coverage.

This chapter evaluates the genomic surveillance strategy in KZN by analysing district-level data. It identifies possible gaps, challenges, and limitations in the current framework. This retrospective analysis contributes to global SARS-CoV-2 research while emphasising the need for localised public health strategies.

3.2 Methods

3.2.1 Compilation of Genomic Sequence data

SARS-CoV-2 genomic surveillance sequences and corresponding metadata files (n = 47,426) were downloaded from GISAID (EPI_SET_230908yt, doi: 10.55876/gis8.230908yt) on 24 April 2023. Sequences and metadata were sorted by province. We obtained 7,560 sequences from KZN, 1,471 of which were excluded due to absent or incorrect district information (Figure S3.1). The KZN cohort of 6,089 SARS-CoV-2 sequences (EPI_SET_230914wa, doi: 10.55876/gis8.230914wa) had sample collection dates ranging from 23 March 2020 to 21 March 2023 and were obtained from both male and female individuals spanning a wide age range. Genomic sequences with the corresponding metadata and sequencing technology files were concurrently downloaded and used for operational trend analysis. The geographical locations of the originating and sequencing laboratories were manually curated. Acknowledgment is given to all sequencing laboratories as part of supplementary files S1 and S2.

3.2.2 Trend Analysis of genomic surveillance data across the KZN districts

The KZN metadata was collated and organised by district, specifically: eThekwini, Amajuba, Harry Gwala, King Cetshwayo, iLembe, Ugu, uMgungundlovu, uMkhanyakude, uMzinyathi, uThukelaand Zululand. Districts are categorised based on levels of urbanicity or rurality. To facilitate analysis of rural and urban distinctions, districts were grouped as outlined in Figure 3.1a. Demographic trends were examined for each district by gender and age for the period 2020 to 2023. The analysis extended to the distribution of SARS-CoV-2 cases and deaths, as well as genomes generated in relation

to the district population. Supplementary data was sourced from the NHLS, Our World In Data (OWID) and https://municipalities.co.za (*Municipalities of South Africa*, 2023). Correlation analyses are conducted to explore relationships between population size and total cases, total cases and deaths and genomes generated and total cases for each district, utilising regression plots produced with statistical and visualisation packages in R.

3.2.3 Trend analysis of circulating lineages and VOCs within KZN

Clades and lineage assignments were verified using NextStrain (https://clades.nextstrain.org). Variants and circulating lineages were trended based on their frequency per district, utilising Pangolin lineages derived from NextClade. Lineages were categorised based on variant identifiers as follows: Alpha (B.1.1.7), Beta (B.1.351.x), Delta and its sub-lineages (B.1.617.2 and AY.x), Omicron and its sub-lineages (B.1.1.529 and BA.x), along with others, which include additional B-lineages, C.1.x, recombinants and other low-frequency lineages. Plots were created to visualise individual lineages for genomes generated, as well as lineages that are predominant in KZN.

3.2.4 Trend analysis of Operational parameters

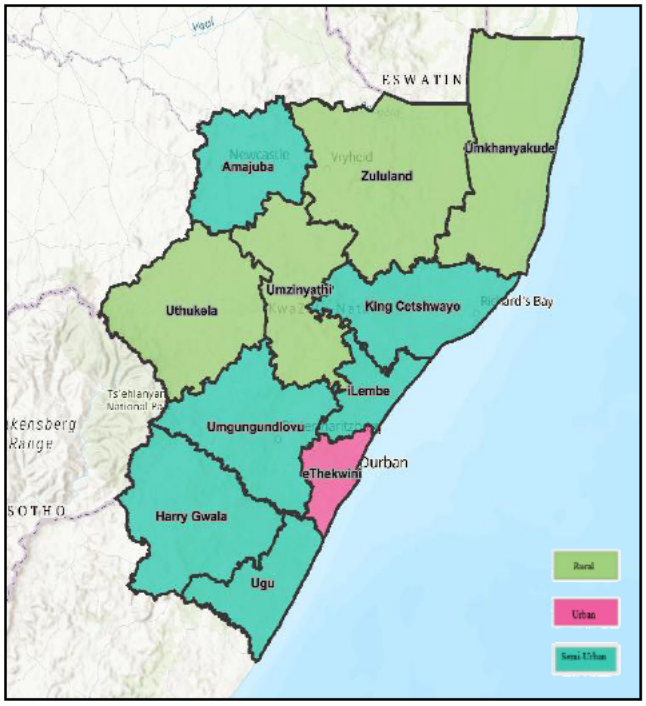
To elucidate the operational parameters essential for an effective genomic surveillance regime, trends based on comprehensive and informative metadata were examined. The following trends were evaluated: sequencing technologies employed in data generation, the assembly software utilised by various sequencing laboratories, the sectors engaged in sample collection as part of the genomic surveillance strategy and the overall turnaround time from sample collection to data submission to GISAID. These metadata files were also reviewed to determine the degree of completion of metadata and the quality of the data recorded.

3.2.5 Data Analysis and Visualisation

All data visualisation and statistical analysis were performed by using available packages such as ggplot2 v3.3.7 and R version 4.3.1. Statistical analysis was performed to establish standard deviation, mean, median, p-values, correlation and confidence intervals. Visualisation of urban and rural districts within KZN are illustrated using ArcGIS (www.demarcation.org.za) (*Municipal Demarcation Board*, 2021).

3.3 Results

3.3.1 KwaZulu-Natal cohort and analysis of district trends



District	Area (km^2)	Population estimates (2016)
Amajuba	7 102	531 327
eThekwini	2 556	3 702 231
Harry Gwala	10 386	510 865
Ilembe	3 270	657 612
King Cetshwayo	8 214	971 135
Ugu	4 791	753 336
uMgungundlovu	9 602	1 095 865
uMkhanyakude	13 855	689 090
uMzinyathi	8 652	554 882
uThukela	11 134	706 588
Zululand	14 800	892 310

Figure 3.1: The distribution of urban and rural districts within KwaZulu-Natal, including the population size and land area (km²) of each district. This visual representation provides insight into the geographic placement of rural versus urban districts, as well as population density patterns across the province. (*Municipalities of South Africa*, 2023).

The 6,089 genomic sequences used in this study were distributed across the 11 urban and rural districts within KZN (Figure 3.1 and Figure S3.1). The number of genomes sequenced per district with the circulating lineages were tabulated in Table S3.1, which also highlighted the number of sequences and circulating lineages for each district across 2020, 2021, 2022 and 2023, respectively. eThekwini, although the smallest district by land area in KwaZulu-Natal, has the highest population due to its urbanised nature. Consequently, it had the most comprehensive genomic annotation across all four years analysed. Trend analysis of cases, deaths, total genomes generated per district were examined to illustrate the disease severity and surveillance efforts over the four years (Figure 3.2). A statistical evaluation of the severity of disease across the districts was represented as the case fatality rate percentage (CFR) (Figure 3.2a). The genomes sequenced from each district in KZN are illustrated in Figure 3.2b, including the mean represented by a red dashed line. The shaded red area indicates the range within one standard deviation from the mean, providing a visual representation of the variability in genome sequencing across the districts (Figure 3.2b). The infection rates within each district were represented per 100,000 people (Figure 3.2c), showing the mean and median IR in red and green dashed lines, respectively.

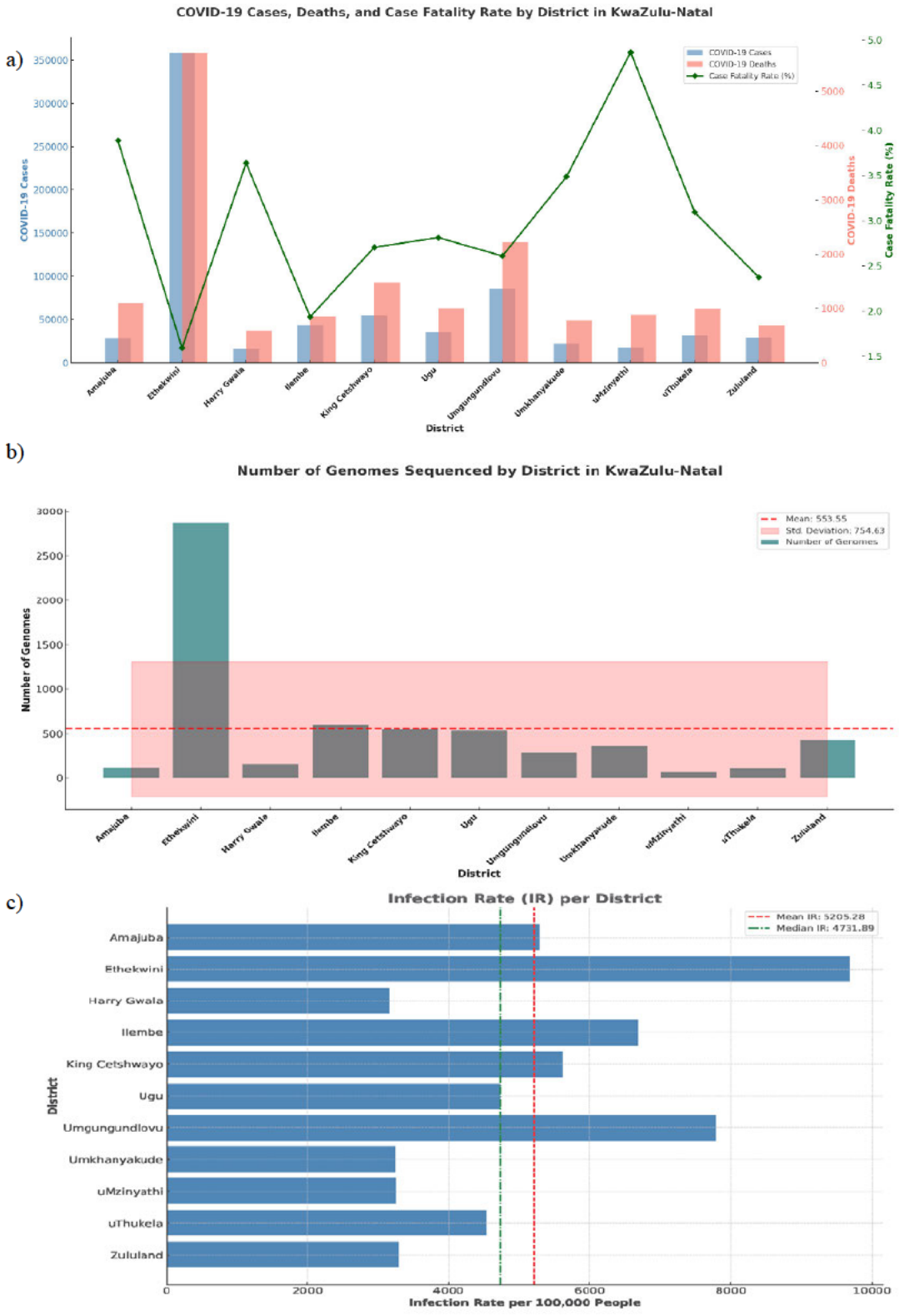


Figure 3.2: (a) SARS-CoV-2 cases and deaths per district and the Case Fatality Rate (CFR) (%), (b) Visualisation of the number of genomes sequenced per district with the mean represented by a red dashed line and, (c) Infection Rates (IR) across the districts in KwaZulu-Natal. Infection rates for each district were standardised and reported per 100,000 people (Figure 3.2c), showing the mean and median IR in red and green dashed lines, respectively.

3.3.2 VOCs and Circulating lineage Trends in KZN districts

The annotated districts were examined to assess the distribution of genomes and lineages across age groups and gender for each year. This analysis revealed variation in demographic patterns and circulating variants of concern between rural and urban districts in KwaZulu-Natal (Table 3.1).

Table 3.1: Aggregate genomic data and all circulating lineages in KwaZulu-Natal (2020 – 2023)

VARIABLE	CLASS	GENOMES (*N* (%))		GENOME DISTRIBUTION ACROSS KZN DISTRICTS				
				Alpha (B.1.1.7)	Beta (B.1.351.x)	Delta and sublineages (B.1.617.2 and AY.x)	Omicron and sublineages (B.1.1.529 and BA.x)	**Others including C.1.x, B.x., Xxx.
GENDER	Female	3603	(59,2)	3	708	549	986	1357
	Male	2236	(36,7)	5	425	403	667	737
	Unknown*	250	(4,1)	1	34	31	72	112
AGE RANGE (YEARS)	0-18	899	(14,8)	0	240	204	198	257
	19-30	1316	(21,6)	1	239	214	411	451
	31-40	1247	(20,5)	1	208	179	360	500
	41-50	960	(15,8)	1	171	156	288	344
	51-60	745	(12,2)	4	132	112	219	271
	>60	778	(12,8)	2	161	99	211	313
	Unknown*	144	(2,4)	0	16	20	38	70
DISTRICT	Amajuba	124	(2,0)	0	30	0	15	79
	eThekwini	2870	(47,1)	2	214	441	1257	956
	Harry Gwala	158	(2,6)	0	60	21	8	69
	iLembe	593	(9,7)	2	170	66	72	283
	King Cetshwayo	546	(9,0)	3	131	237	92	83
	uGu	540	(8,9)	0	143	87	112	198
	uMgungundlovu	283	(4,6)	0	26	52	88	117
	uMkanyakude	358	(5,9)	0	151	15	24	168
	uMzinyathi	64	(1,1)	0	3	3	6	52
	uThukela	117	(1,9)	1	9	7	21	79
	Zululand	436	(7,2)	1	230	53	30	122

**Unknown represents those individuals whose data was unavailable or not recorded at the time of specimen collection*

*** Others includes all remaining lineages present such as x sublineages, recombinant viruses, other A. and B. lineages*

A subset of data was categorised as 'Unknown,' indicating the absence or unavailability of particular demographic information at the time of specimen collection. For completed metadata, females represent over half of the genomes at 59,2%, while males account for 36,7% and a small proportion (4,1%) remain of unknown gender. The 19–40 year age group was notably well represented in the genomic data, averaging 21,1%. This may indicate a higher infection rate or a greater likelihood of testing among younger adults. Nevertheless, the data indicate robust representation across all age groups, with each exceeding 12%. The 6089 genomes for KZN were distributed across the 11 districts, with 47,1% collected from eThekwini. The remaining 10 districts accounted for 1,1-9,7% of the genomes. These genomes, encompassing several VOCs, and their sublineages were distributed accordingly throughout the province. To gain a deeper understanding of the distribution and diversity of lineages that circulated in KZN (Figure 3.3) over time, the identified lineages were plotted for the KZN region (Figure 3.4a) and within each district (Figures 3.4b and 3.5).

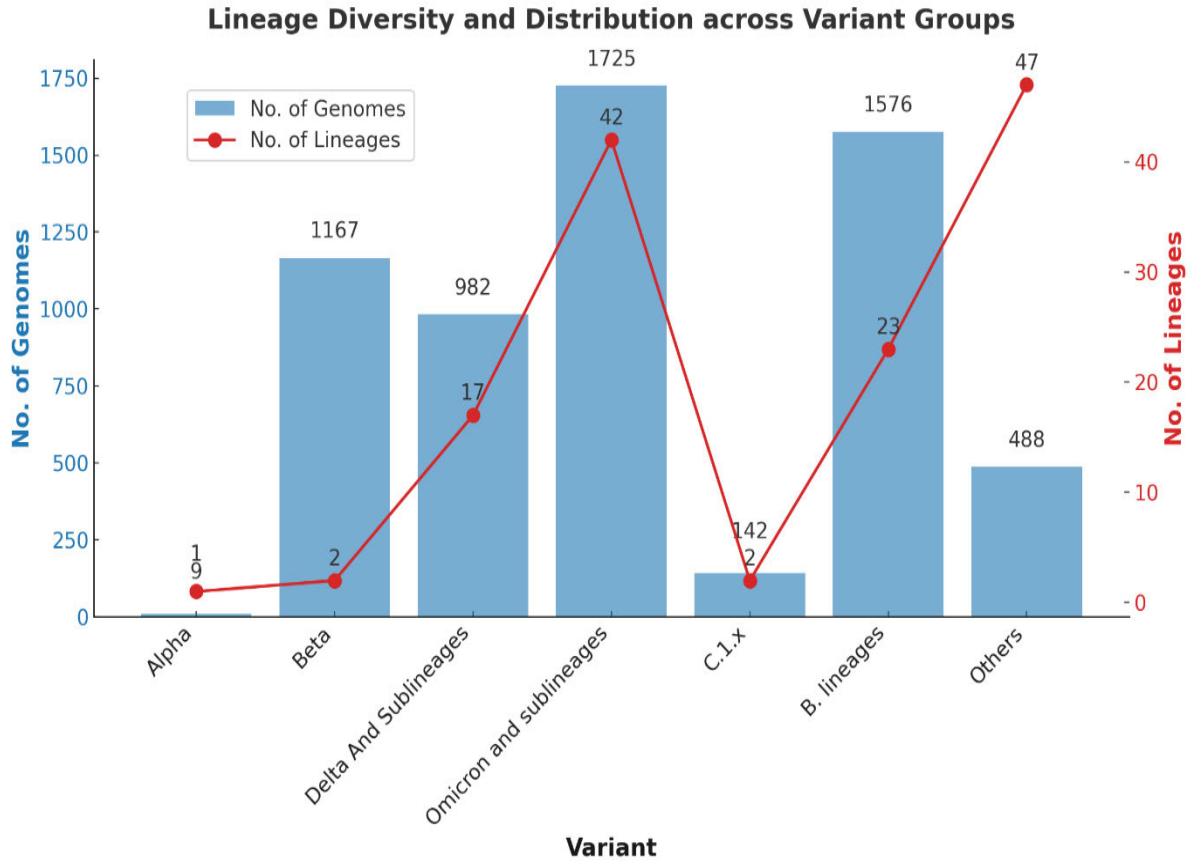


Figure 3.3: Lineage diversity and distribution across variants groups for the total genomic cohort

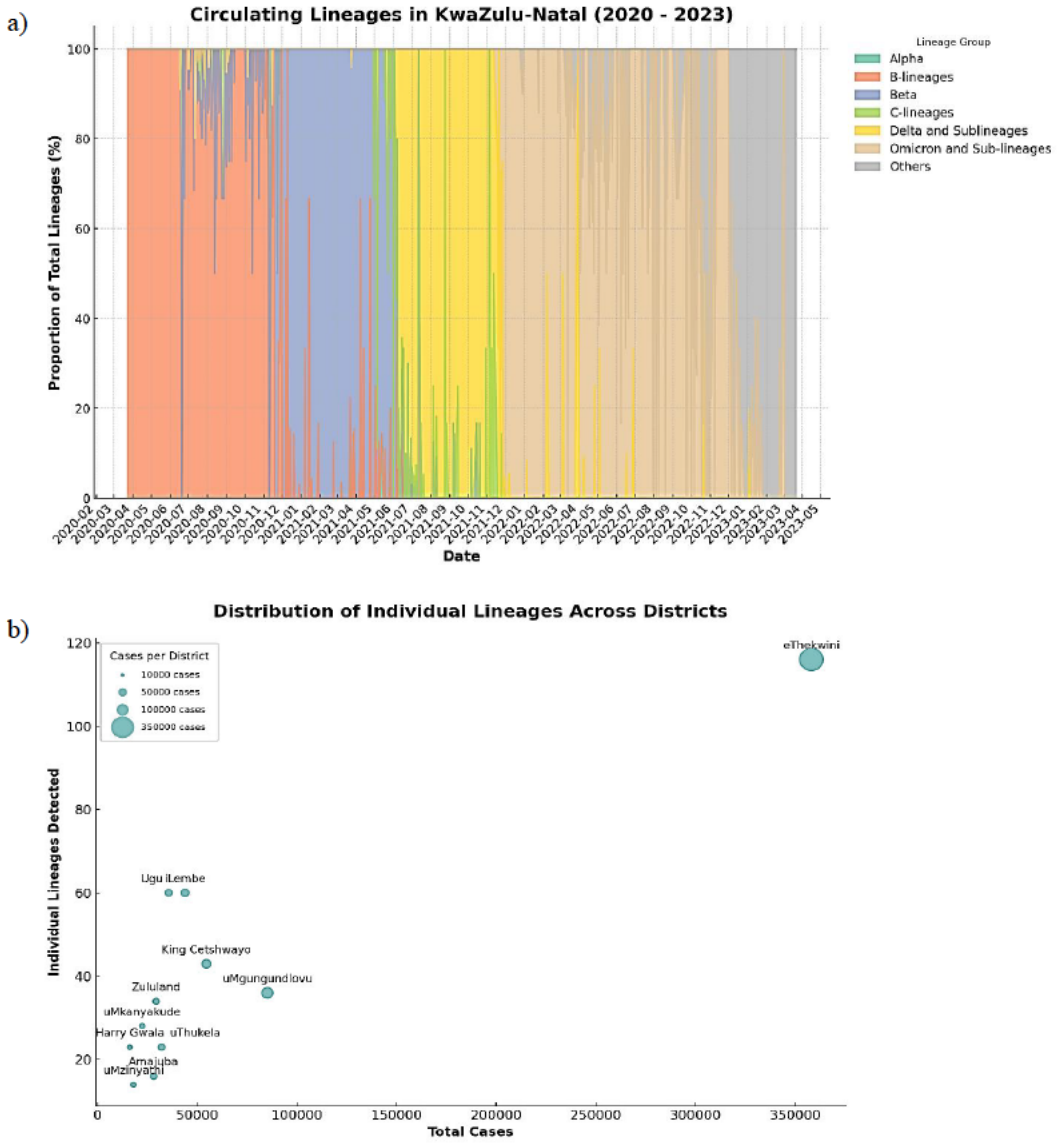


Figure 3.4: Lineage diversity a) within KwaZulu-Natal from March 2020 to 2023 and, b) number of lineages detected per district total cases

Lineages were further examined for the total cases reported for each district (Figure 3.5). The total cases are indicative of the level of infection for each respective district within KZN and are represented by coloured circles of differing size. We further examined the individual lineages circulating within each district over time. The raw data for genome and lineage distribution per district are found in supplementary figures (Figures S3.2 and S3.3, respectively).

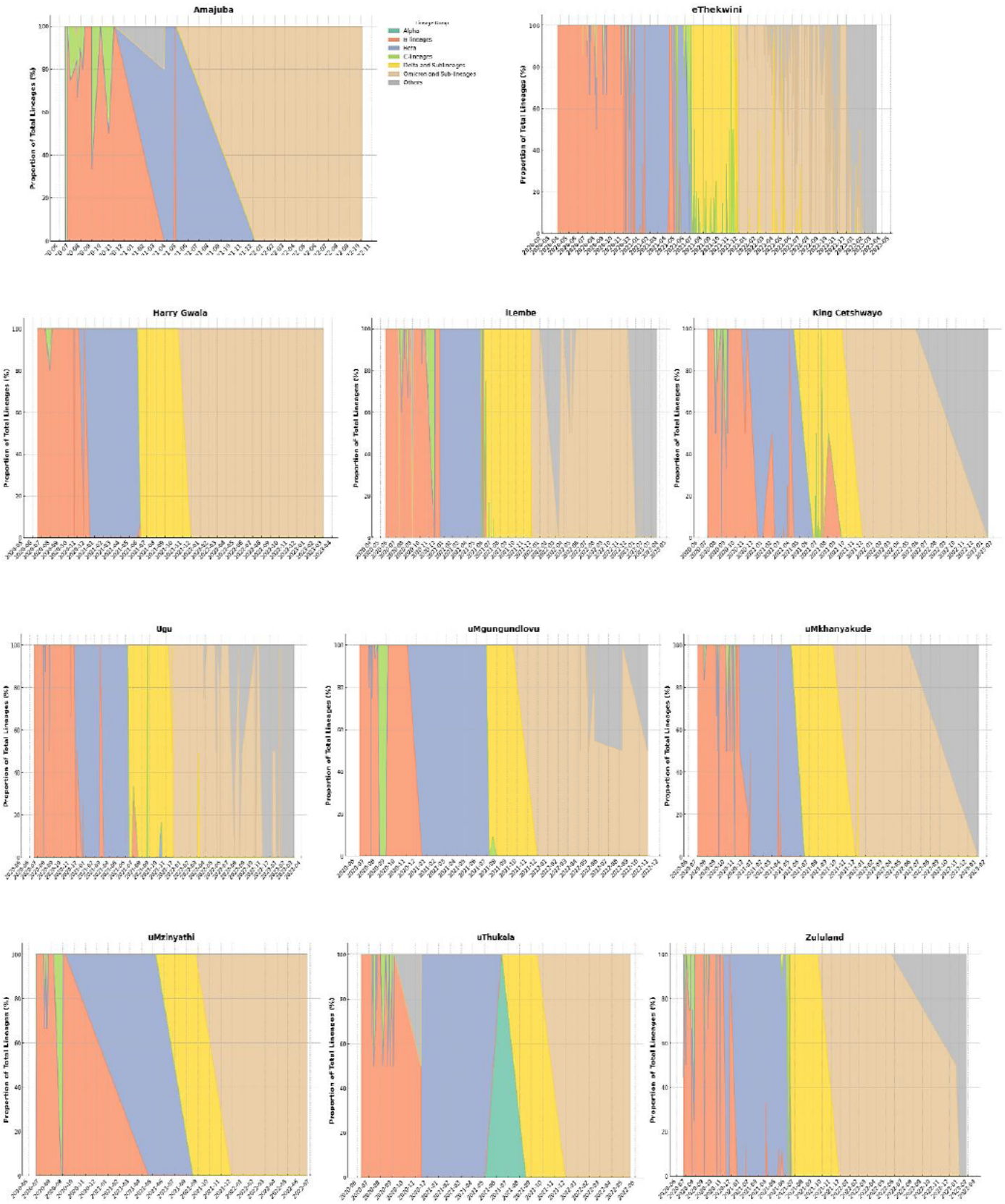


Figure 3.5: Lineage diversity across the 11 districts from March 2020 to 2023.

Lineages detected at least 50 times (>1%) from March 2020 to March 2023 in the KZN genome dataset are illustrated in Figure 3.6. All other lineages (including <1%) detected for the same period are illustrated in Figure S3.3.

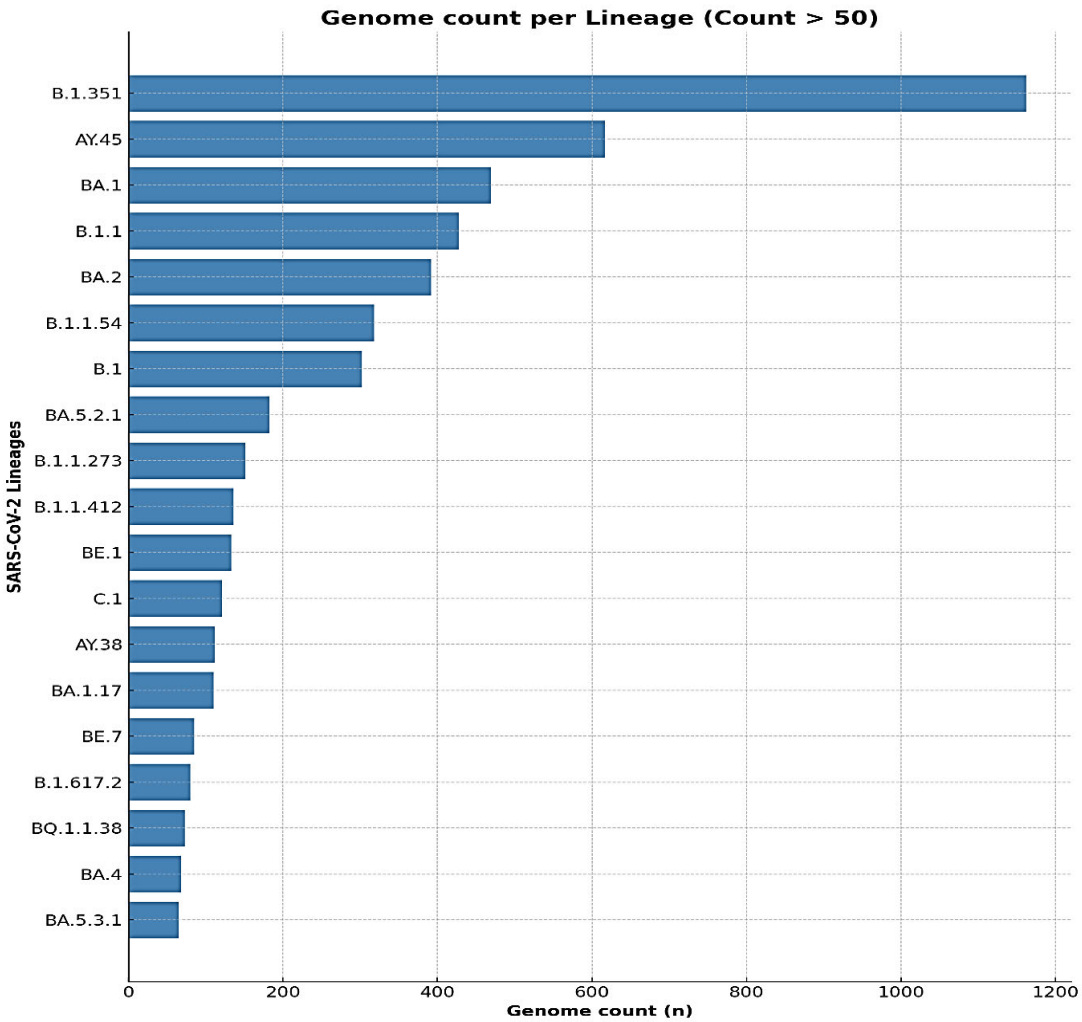


Figure 3.6: Most frequently sampled lineages between March 2020 – 2023.

3.3.3 Operational Trends for Genomes generated from KZN districts

A visual comparison of the COVID-19 genome sequencing rates relative to the total cases across various districts was examined (Figure 3.7). Figure 3.7 b) depicts urban and semi-urban districts to have statistically significant sequencing rates (t-test = 5.42, $p < 0{,}001$). Rural districts also show sequencing activity, but at a borderline level of significance (p value < 0,1). There is no strong statistical difference in sequencing rates between urban/semi-urban and rural areas, suggesting that sequencing efforts are relatively evenly distributed. These statistical measures are crucial for assessing the consistency and outreach of genomic surveillance across the region.

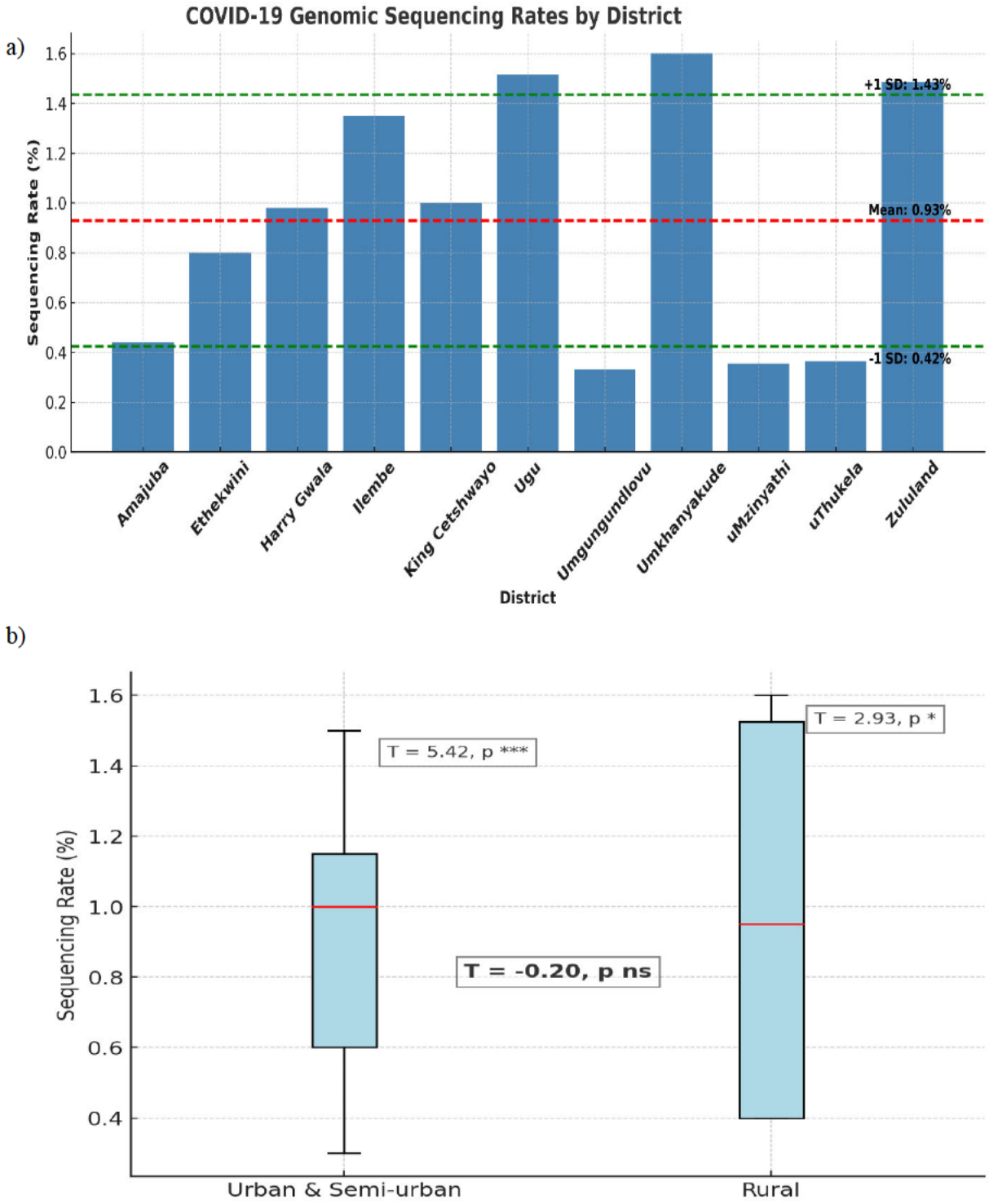


Figure 3.7: COVID-19 genomes sequencing rates as a percentage of the total cases for a) each district and b) within urban and semi-urban, and rural districts. Median (red line) and statistical annotations summarising the t-test was represented by "*" (***: $p < 0.001$, *: $p < 0.1$, *ns* : not significant). The mean sequencing rate across the districts is approximately 0,93%, displayed by a red dashed line. The standard deviation ±1 of the sequencing rates is approximately 0,50 and displayed in green. The lines indicating one standard deviation above and below the mean provide insight into the variability of the sequencing efforts across different districts.

The sequencing platforms used for data generated by the submitting laboratories were examined in Table 3.2, as well as the assembly software used during analysis in Table 3.3.

Table 3.2: Genomic sequencing technology distribution across a network of submitting laboratories

Sequencing Platform	Submitting laboratories			Total (*n* (%))	
	KRISP/CERI	NHLS	NICD		
Illumina MiSeq	4643	3	0	4646	(76,3%)
Illumina NextSeq	23	0	22	45	(0,7%)
Ion Torrent	0	0	21	21	(0,3%)
Oxford Nanopore Technology (GridIon, MinIon)	1377	0	0	1377	(22,6%)
Total (*n* (%))	6043 (99,2%)	3 (0,1%)	43 (0,7%)	6089	(100%)

A tabulated overview of samples received for sequencing from the respective originating laboratory sectors is also listed in Table 3.3. Information about the distribution of originating laboratories for a given dataset were categorised into the Private Sector, Government Sector and Research Institutes.

Table 3.3: Assembly software used during pre-analysis of genomes and originating laboratory sectors

Assembly method	(*n* (%))	
Genome Detective	5951	(97,7%)
Exatype	2	(0,0%)
Others (Artic, Basepace, Dragen)	136	(2,2%)
Originating laboratories	(*n* (%))	
Private Sector	246	(4,0%)
Government Sector	5261	(86,4%)
Research institutes	582	(9,6%)

The interval between the collection of the sample and its submission to the GISAID database was also assessed (Figure 3.8). The TAT was calculated based on the number of genomes generated from sample collection on a monthly from March 2020 to March 2023. TATs were categorised and denoted as ≤ 21 days, 21 < days ≤ 60, 60 < days ≤ 120and > 120 days. The dataset was analysed using a bar plot to visualise the number of submitted genomes for each category per month. The mean and ANOVA tests were calculated for each category over the past three years and the p-values were represented as "*" (***: $p < 0.05$).

Genomic sequencing across the submitting laboratory network was predominantly conducted by KRISP/CERI, which contributed over 99% of all genomes analysed. Illumina MiSeq was the primary sequencing platform used (76,3%), reflecting a preference for high-accuracy short-read sequencing, while Oxford Nanopore Technology accounted for 22,6% of data generation, supporting rapid sequencing during periods of increased demand. Assembly workflows were highly standardised, with 97,7% of genomes processed using Genome Detective, ensuring analytical consistency across the dataset. Most samples originated from the Government Sector (86,4%), highlighting the central role of public healthcare facilities in national SARS-CoV-2 genomic surveillance, while contributions from research institutes (9,6%) and the Private Sector (4,0%) were comparatively smaller. TAT from sample collection to submission to GISAID varied over the study period and was categorised for analysis. Statistically significant differences ($p < 0.05$) between TAT categories demonstrated fluctuations associated with changing epidemic dynamics and laboratory throughput pressures, indicating that reporting speed improved during periods of heightened public health response and slowed during intervals of reduced case load or resource strain.

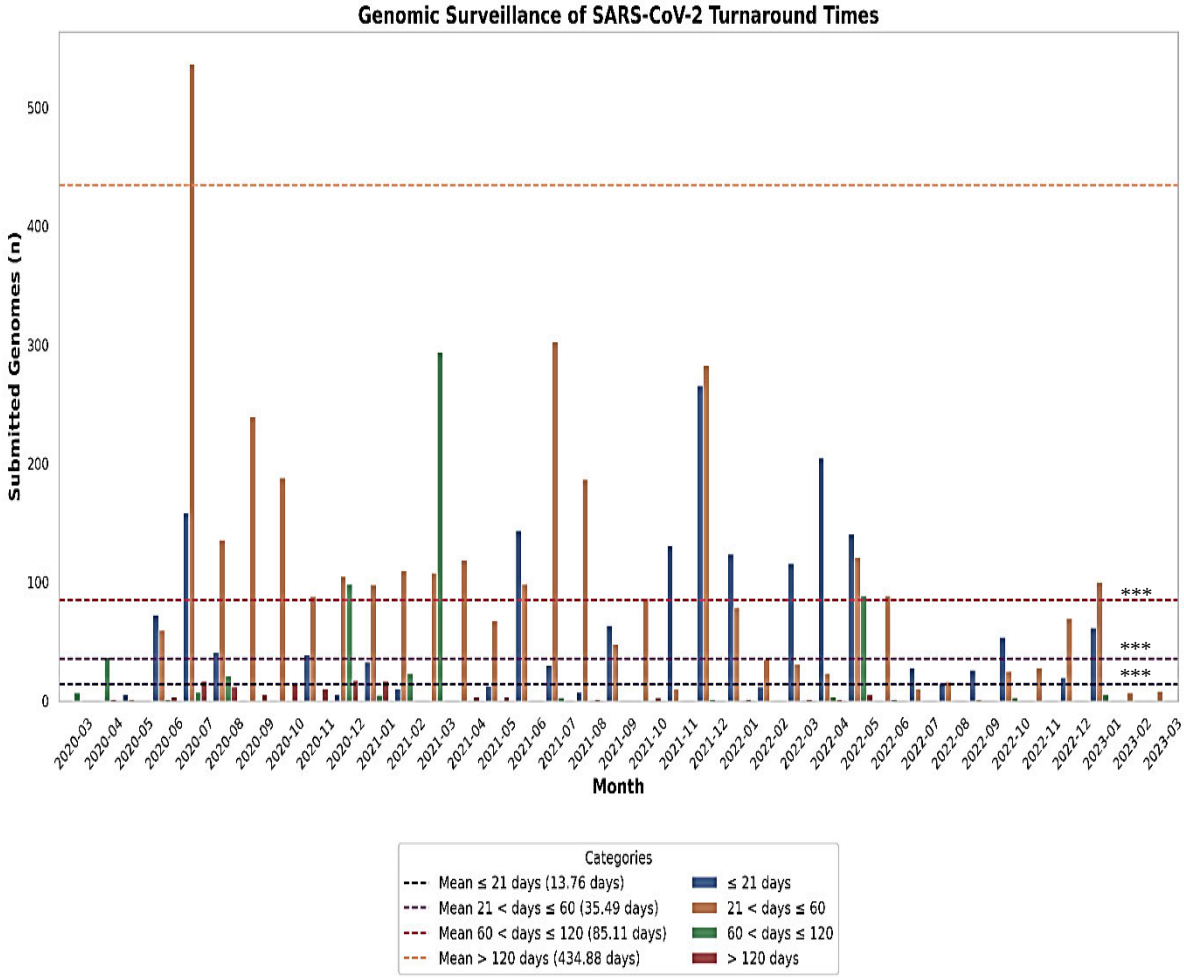


Figure 3.8: SARS-CoV-2 genomic surveillance turnaround time (TAT) (***: $p < 0.005$)

3.4 Discussion

The analysis of genomic surveillance data from KZN within the framework of the COVID-19 pandemic has yielded significant insights into the regional dynamics of SARS-CoV-2. The robust dataset of 6,089 genomic sequences facilitated a detailed investigation into the correlation of viral transmission and demographic as well as geographic factors. The demographic distribution shows a higher testing rate among younger adults aged 19-40, which aligns with global observations of higher social activity and mobility in this cohort, potentially increasing transmission risk as also observed with HIV in 16-30 year olds (Kharsany *et al.*, 2019; Harris, 2020; Pritchard *et al.*, 2022; Tegally, San, *et al.*, 2022). In terms of gender, genomes generated from females made up 59,2% of all submitted genomes from KZN, while males accounted for 36,7%. Individuals, whose gender was not captured, were listed as 'Unknown' and made up 4,1% of submitted genomes. This proportion of age and gender was in line with a previous study that observed a higher SARS-CoV-2 seropositivity in 20- 40 year olds and in women than in men (Moyo *et al.*, 2023).

The urban-rural divide in KZN presents distinct patterns of disease spread, with urban areas like eThekwini accounting for nearly half of the sequenced genomes in the province. This distribution reflects a higher population density and greater access to healthcare and testing facilities in urban districts. The CFR representing the percentage of deaths out of total cases varies considerably across districts, especially in the rural districts, pointing to disparities in healthcare access, demographic influences, or disease severity (Figure 3.2a). These disparities compromise the efficiency of the measures imposed in surveillance and control and become obstacles to the realisation of primary care for all (Maimela *et al.*, 2024). The disparity in genome sequencing across districts highlights the challenge of achieving equitable genomic surveillance, essential for effective pandemic monitoring (Hamzaoui *et al.*, 2024). Approximately 19,5% of genomes from KZN were excluded from the study due to the absence of metadata at the time of collection. This underscores the need for stricter data collection protocols and enhanced standardisation in genomic surveillance. Several inconsistencies and erroneous entries were identified during the analysis of metadata from GISAID, including: i) spelling errors and variations in spelling, ii) incorrect location, city and province information, iii) inaccurate submission and originating laboratory listings, iv) incorrect formats for age, dates and gender, v) duplicate entries and vi) insufficient information captured. KZN was selected for this study due to its status as the most thoroughly annotated province, featuring comprehensive sampling across all districts. Since a significant proportion of the genomes from KZN were found to lack supporting data, it is possible that the proportion of missing data will be higher in other provinces. This further emphasises the disparities in data standardisation. These findings underscore the urgent need for improvements in metadata quality and the establishment of standard operating procedures (SOPs) for data capturing in such databases (Gozashti and Corbett-Detig, 2021; Ma *et al.*, 2023). Enhanced metadata quality ensures that genomic surveillance efforts can more effectively inform policy decisions, guide resource

allocation and ultimately improve public health outcomes by enabling more precise and actionable insights (Ling-Hu *et al.*, 2022).

The correlation analyses between the number of genomes sequenced and the cases reported per district shed light on the operational strengths and gaps within the current genomic surveillance system. The visual and statistical analysis indicates variability in surveillance intensity, which is crucial for detecting emergent strains and implementing timely interventions. The operational trends suggest that while the sequencing efforts are commendable, there is variation in how comprehensively different districts are monitored. The global benchmark for sequencing rates of SARS-CoV-2 varies depending on public health goals, variant surveillance needs, and sequencing capacity in different regions. However, based on recommendations from organisations like the WHO, GISAID, and CDC, an ideal benchmark of 5-10% confirmed cases should be sequenced. Many high-income countries with strong sequencing infrastructure aim for this gold standard as observed in the United Kingdom (10-15% at peak pandemic), United States of America (3-7%) and Europe (~5%). Low-income countries often sequence below 1% due to resource constraints. Although a sequencing rate of 5-10% of confirmed cases is recommended for optimal genomic surveillance, a minimum of 1-2% sequencing rate is necessary for baseline variant detection, while strategic sampling approaches should be employed during periods of low case numbers to efficiently detect emerging variants (Liu *et al.*, 2021). The analysis of circulating lineages and VOCs within KZN highlights the dynamic nature of the viral population. The predominance of the Beta and Delta variants during the specific waves illustrates the impact of viral evolution on the pandemic's trajectory (Tegally *et al.*, 2021; Al Hasan *et al.*, 2022). The Beta variant, associated with increased transmissibility, led to a pronounced second wave. Successively, the emergence of the Delta and Omicron variants underscores the dynamic nature of viral evolution and its implication on public health strategies. The shifts in variant dominance also reflect the ongoing global narrative of the pandemic, where each variant's emergence could potentially alter the effectiveness of existing vaccines and therapeutic approaches, necessitating adjustments in public health responses.

The operational characteristics of the genomic surveillance system reflect both the regions sequencing capacity and its reliance on centralised infrastructure (Tegally *et al.*, 2020; Giandhari *et al.*, 2021). The dominance of Illumina-based platforms, particularly within KRISP/CERI, demonstrates a well-established and high-accuracy sequencing framework, while the supplementary use of Oxford Nanopore technology highlights the ability to rapidly generate genomes when rapid public health insights are needed (Quick *et al.*, 2016; Tyson *et al.*, 2020; Tegally, Wilkinson, Giovanetti, *et al.*, 2021). However, the presence of multiple sequencing technologies also underscores the importance of maintaining standardised analytical workflows to ensure consistency and comparability across laboratories (Cleemput *et al.*, 2020). The assessment of TAT trends further illustrates the adaptive performance of the surveillance network. Over the study period, there was a clear increase in the

proportion of genomes submitted within ≤ 21 days and 21–60 days, reflecting improved processing efficiency and responsiveness, consistent with the previously suggested 21-day benchmark for effective genomic surveillance (Brito *et al.*, 2022). The statistically significant differences observed between TAT categories indicate meaningful shifts in reporting dynamics over time. While no strong correlation was observed between total sequencing volume and TAT, the data suggest that periods of high sample intake may still contribute to delayed processing due to constraints in personnel and laboratory capacity. This highlights the need for continued investment in workforce development and resource scaling to sustain rapid genomic reporting during periods of increased epidemic pressure (de Oliveira and Baxter, 2024; Govender, 2024). Delays in data submission to GISAID, caused by factors also described in chapter 2, resulted in an average delay in genome reporting. This delay could have influenced the perceived emergence of VOCs, impacting public health interventions as observed with the Delta variant (Ramphal *et al.*, 2024). Sequencing limitations, including delayed turnaround times and possibly inadequate rural coverage, may have impacted the accuracy of SARS-CoV-2 lineage detection (Colton *et al.*, 2023). The consensus pertaining to delays in processing hinder real-time surveillance, may allow variants to spread unnoticed, while under-sampling in rural areas creates geographic biases and data gaps (Meckawy *et al.*, 2022). Such constraints could potentially weaken epidemiological trends, delay vaccine and treatment adaptations, and reduce the effectiveness of public health interventions, emphasising the need for improved sequencing infrastructure and faster genomic surveillance. Continuous monitoring and analysis are, therefore, essential in improving turnaround times and ensuring timely real-time surveillance.

This retrospective analysis underscores several points for enhancing future public health strategies and preparedness. The identification of operational challenges, such as the variability in genomic data coverage and the turnaround time from sample collection to data submission, is critical. Furthermore, convenience-based sampling methods can distort the actual distribution of circulating variants (Ling-Hu *et al.*, 2022). Addressing these issues will require strengthening the infrastructure for genomic surveillance, possibly through enhanced funding, better training for local health personnel and the integration of genomic data with electronic health records for real-time tracking. Moreover, the findings advocate for a more granular, district-level approach to public health planning and intervention. Tailoring strategies to fit the unique epidemiological and genomic contexts of each district could enhance the responsiveness and effectiveness of interventions, potentially curtailing transmission more efficiently and mitigating the impact of future waves. While this chapter did not include a formal phylogeographic analysis, future studies incorporating such approaches could provide greater insight into the spatiotemporal spread of SARS-CoV-2 lineages across districts and enhance understanding of regional transmission dynamics.

3.5 Conclusion

Genomic surveillance of SARS-CoV-2 in KZN has improved our understanding of how the virus evolves and spreads. This study highlights key challenges, including sequencing disparities across districts, missing data, and delays in processing. Such issues affect the accuracy of variant tracking and public health responses. While KZN serves as a model for genomic surveillance in South Africa, the high proportion of missing data suggests even greater challenges in other regions, underscoring the urgent need for standardised data capture protocols and equitable sequencing efforts. Integrating genomic data with demographic and epidemiological insights is crucial for refining public health strategies, both locally and globally. A structured and well-resourced genomic surveillance system is essential to improve data standardisation, logistical efficiency, and laboratory capabilities, thereby enhancing the accuracy and responsiveness of outbreak management. Strengthening these areas through increased funding, improved personnel training, and the integration of genomic data with electronic health records will enable more effective disease monitoring and intervention. This study underscores the necessity of a district-level approach to public health planning, tailored to the unique epidemiological contexts of different regions. Policymakers can use genomic insights to track viral evolution, allocate resources efficiently, and implement targeted measures to control outbreaks. A strong public health system that prioritises genomic surveillance is essential for pandemic preparedness and global health security.

Linking Chapters three and four

Chapter Three concludes with an in-depth analysis of the SARS-CoV-2 trends in KZN, highlighting the importance of understanding localised genomic surveillance within a well-annotated province. This analysis provides insights into demographic and transmission patterns, and identifies gaps in current strategies, setting a foundation for enhancing future surveillance efforts within urban and rural settings. Chapter Four transitions from a local to a global perspective, analysing South Africa's SARS-CoV-2 genomic data sharing within the broader context of global trends. It evaluates the quality, quantity, and timeliness of shared genomic data across selected high-, middle-, and low-income countries. This chapter benchmarks South Africa's genomic surveillance efforts against global trends, highlighting achievements, challenges, and strategic insights for improvement.

Chapter Four: Unveiling South Africa's Triumphs: An Exploration of SARS-CoV-2 Genome Sharing Trends and Impacts, 2020–2024

4.1 Background

The timely and widespread sharing of genomic sequences has become a cornerstone in managing global health crises, particularly in the context of the SARS-CoV-2 pandemic (Baker *et al.*, 2023; Struelens *et al.*, 2024). This exchange of data provides critical insights into viral evolution, supports the development of diagnostic tools, vaccines, and therapeutics, and strengthens international collaboration efforts to control outbreaks (Figure 4.1) (Ramesh *et al.*, 2021; Brito *et al.*, 2022; Carter *et al.*, 2022; Rabaan *et al.*, 2023). The COVID-19 pandemic has reinforced the necessity of rapid genomic surveillance, with global data repositories such as GISAID playing a central role in tracking viral spread and mutations (Hamzaoui *et al.*, 2024). Within this evolving landscape, South Africa has emerged as a key contributor to genomic surveillance, leveraging its historical expertise in handling infectious diseases such as HIV, tuberculosis, and COVID-19 through innovative techniques and robust public-private collaborations (Karim *et al.*, 2020; Bekker and Ntusi, 2021; Pritchard *et al.*, 2022).

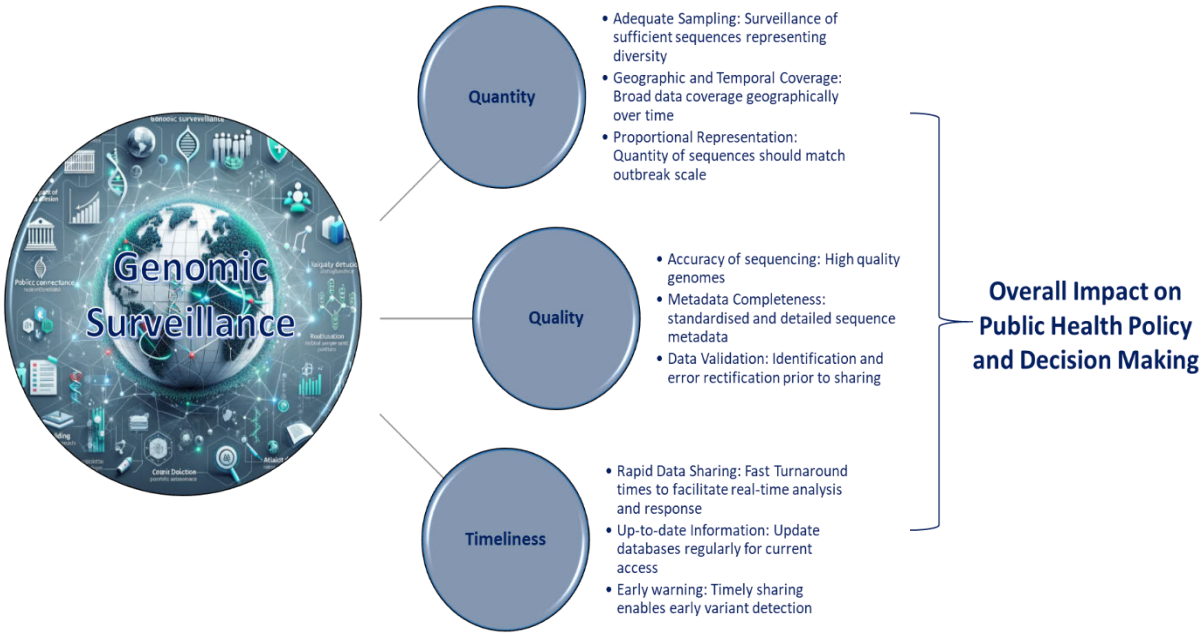


Figure 4.1: An adapted overview of the SARS-CoV-2 Genomic Surveillance System highlighting key metrics in genome sharing (Carter *et al.*, 2022). Image created using Dall-E (*Open AI*, 2023)

Since the early stages of the pandemic, South Africa has been at the forefront of the genomic sequencing efforts in Africa, significantly contributing SARS-CoV-2 genomes to international databases (Giandhari *et al.*, 2021). This early engagement was pivotal in tracking viral transmission patterns and identifying variants of concern, such as the Beta (B.1.351) and Omicron (B.1.1.529) variants, first detected in South Africa (Tegally *et al.*, 2021; Al Hasan *et al.*, 2022). The establishment of the Network for Genomic Surveillance in South Africa (NGS-SA) in May 2020 (Msomi *et al.*, 2020) further strengthened the country's capacity to generate and analyse SARS-CoV-2 genomic data in a timely manner, demonstrating the importance of localised genomic surveillance systems in informing public health interventions (Hill *et al.*, 2023). However, despite these achievements, challenges persist

in genomic sequencing infrastructure, particularly concerning disparities in sequencing capabilities, data quality, and timeliness across different regions (Brito *et al.*, 2022; Feng *et al.*, 2024).

The rapid evolution of global genomic systems has been driven by advancements in sequencing technologies, with high-income countries leveraging these sequencing platforms to inform public health responses through high sequencing coverage and rapid turnaround times (Pronyk *et al.*, 2023). Middle-income countries, including South Africa, have faced resource constraints but have made notable progress through international collaborations and public-private partnerships (Hossain, Abdulla and Rahman, 2022). Meanwhile, low-income countries continue to struggle with limited sequencing capacity, inadequate infrastructure, and slower data-sharing processes, highlighting global inequities in genomic surveillance (Belman, Saha and Beale, 2022). Addressing these disparities is crucial for enhancing global pandemic preparedness, ensuring equitable access to sequencing technologies, and strengthening international data-sharing frameworks (Brito *et al.*, 2022; Chen *et al.*, 2022).

This study aims to analyse trends in SARS-CoV-2 genomic surveillance data sharing in South Africa from 2020 to 2024, focusing on key metrics such as data quality, quantity, and timeliness. By benchmarking South Africa's contributions to those of countries across various socioeconomic levels and global trends, this research highlights the nation's achievements and challenges in supporting the global pandemic response. The findings will provide strategic insights to strengthen genomic surveillance within the region, contributing to broader efforts aimed at improving pandemic preparedness and response capabilities worldwide.

4.2 Methods

4.2.1 Data source and collection

Genomic surveillance and SARS-CoV-2 case data for South Africa, specified countries and globally were sourced from GISAID and Our World in Data (OWID) on 19 February 2024, respectively. Additional information was accessed from GISAID on 27 January 2025. Case data was crosschecked with WHO, CDC, and the COVID-19 Data Portal. The dataset collection dates spanned from 1 January 2020 to 31 January 2024. Additionally, publication data for South Africa and related regions was retrieved from Scopus and Web of Science (WoS) on 24 January 2024.

4.2.1.1 Selection Criteria and Data Management of GISAID Data

South African and global data used in the study was collected between 1 January 2020 and 31 January 2024. To ensure completeness, only whole genomes were filtered, retaining those with a length exceeding 29,000 nucleotides and containing no more than 5% ambiguous bases (Ns) across the entire genome. High coverage Genomes were further selected based on metadata completeness. Genomic metadata containing age listed in numerical values between 0 and 150 were selected, as well as gender

that was defined as “Male” or “Female”. The genomes were classified according to the continent and country/region, based on the information provided in the “Location” field of the metadata associated with each genome, indicating the geographical locations where samples were collected. Any discrepancies and duplicates were excluded. Data was plotted against cumulative global data for comparison. This study relies primarily on SARS-CoV-2 genomic data submitted to GISAID, which is the main data-sharing platform used in South Africa and collaborating regions. However, because some laboratories deposit sequences in other repositories or submit data with delays, it is possible that certain genomes were not captured, and the dataset may not fully represent all sequencing efforts conducted during the study period.

4.2.2 Selection Criteria and Approach for the Comparative Assessment of SARS-CoV-2 Genomic Data Sharing

To assess trends in SARS-CoV-2 genomic data sharing in South Africa, we focused on key metrics, including data quality, quantity, and timeliness across continental and specific countries. To ensure a comprehensive and representative analysis, countries were selected from each continent with diverse economic statuses and geographical locations, based on World Bank classifications. This selection included nations with well-established genomic surveillance systems (example: the United Kingdom and Denmark (Europe), the United States (North America), and Australia (Oceania)), countries comparable to South Africa in terms of economic and regional context (example: Brazil (South America), India and Thailand (Asia), Kenya and Nigeria (Africa)), and those facing distinctive genomic surveillance challenges (example: Singapore (Asia) and Canada (North America)). The genomic surveillance trends of each selected country and globally were analysed based on the components outlined in Figure 4.1, utilising data from Our World in Data (OWID) for each country’s genomic surveillance statistics. Particular emphasis was placed on comparing these approaches to those adopted in South Africa, enabling a reciprocal evaluation of genomic data-sharing practices.

4.2.2.1 Justification of Country selection

The selection of countries for comparison with South Africa’s SARS-CoV-2 genomic surveillance was based on their representation of the continent they belonged to. Countries with advanced genomic surveillance infrastructures—such as the United Kingdom, United States, Australia, and Denmark—were included due to their high sequencing capacity, rapid data dissemination, and significant contributions to global databases like GISAID. Brazil, India, Thailand, Kenya, and Nigeria were selected for their socioeconomic and regional comparability to South Africa, facing similar healthcare and resource challenges. Singapore and Canada were included to illustrate unique surveillance challenges: Singapore’s intensive but small-scale approach and Canada’s coordination across vast provinces. This selection ensures a diverse global perspective, enabling an assessment of

South Africa's standing, identifying strengths, and informing strategies to enhance its genomic surveillance capacity.

4.2.3 Publication Output

Guided by a study published in Scientometrics (Teixeira da Silva, Tsigaris and Erfanmanesh, 2021), we performed a basic search for publications on Clarivate's Web of Science (WoS) and Elsevier's Scopus to quantify research output. We examined documents published from 1 January 2020, to 24 January 2024, using search queries "SARS-CoV-2," "COVID-19," "Coronavirus 2019," "2019-nCoV," "novel coronavirus," "SARS-CoV-2 infection," and "COVID" in the topic and keywords. Data was downloaded on 24 January 2024. The searches on WoS and Scopus encompassed all databases. We retrieved a total of 640,967 documents from WoS and 412,113 documents from Scopus. The data was analysed based on various fields, including publication years, document types, organisations/institutes, countries of interest and funders. The data was curated manually to identify recurring funding sponsors and organisations/institutes.

4.2.4 Data analysis and visualisation

Data was analysed and visualised using R version 4.3.1 using available packages and scripts. Where applicable, other visualisation tools such as Tableau (online on https://www.tableau.com/) and Matplotlib (Python) were used. Quantitative statistical analysis was conducted using R's statistical packages to provide a detailed landscape of genomic surveillance.

4.3 Results

4.3.1 Global Variation in quantity, quality and timeliness of SARS-CoV-2 Genomic Data sharing efforts

An overview of the SARS-CoV-2 genomic surveillance system was highlighted in Figure 4.1. Data sharing efforts were trended across selected countries and globally. Figure 4.2 depicts a timeline of the first cases and total cases for the selected countries over time. The total cumulative confirmed COVID-19 cases highlight the overall impact of the pandemic on each country. The results show clear differences in the progression of total COVID-19 cases between high-income and low- and middle-income countries over time. High-income countries such as the United States, Brazil, and the United Kingdom experienced sustained and substantial increases in cumulative cases, characterised by multiple waves of heightened transmission. These steep and continuous rises reflect prolonged community spread, influenced in part by higher levels of population mobility and earlier international exposure. In contrast, countries including Kenya, Nigeria, and Thailand reported considerably lower case totals with

slower growth over the same period. Middle-income countries such as India and South Africa exhibited more abrupt, concentrated periods of increasing cases, suggesting shorter but intense outbreak phases followed by periods of relative stabilisation. The widening divergence in cumulative case trajectories highlights how economic capacity, public health system readiness, testing availability, and policy responses shaped the scale and timing of transmission. Overall, the data indicate that the burden and progression of the pandemic were not uniform globally, but were instead stratified along socioeconomic lines, with high-income countries experiencing more prolonged and larger outbreaks than many lower-income settings.

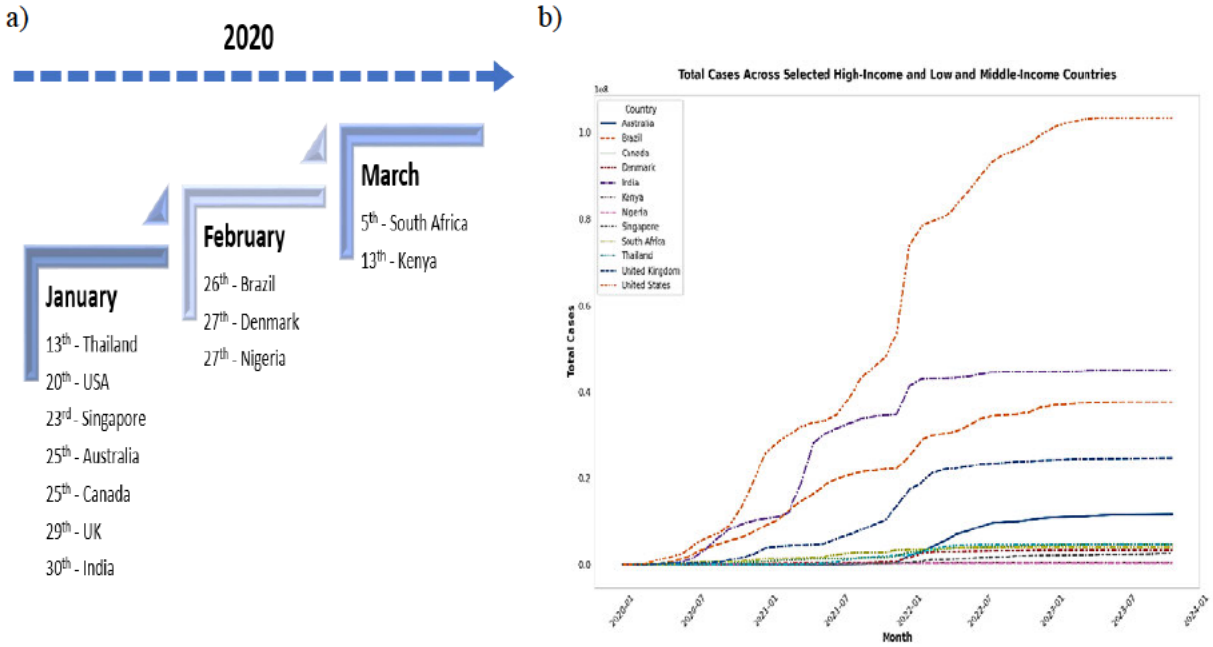


Figure 4.2: a) A timeline of the first detected cases of SARS-CoV-2 and b) total cases (x10^8) reported in selected countries of diverse economic standings and geographical locations, sourced from GISAID and OWID (date accessed: 19/02/2024).

The map depicted in Figure 4.3 provides an overview of the disease burden across select countries. The mapping of COVID-19 cases, deaths, population size, testing volume, and genomic submissions reveals significant disparities in both the scale of the pandemic and the capacity to monitor it across countries. High-income countries such as the United States, the United Kingdom, and Canada reported very high case counts and deaths, accompanied by extensive testing and comparatively large numbers of genome sequences submitted to global surveillance databases. This pattern reflects both a higher confirmed disease burden and a strong ability to detect, report, and genetically characterise circulating variants. In contrast, low- and middle-income countries, including Kenya and Nigeria, reported substantially lower total case counts and fewer deaths, but these figures must be interpreted alongside markedly lower testing rates and minimal genomic sequencing. This suggests that their recorded burden likely underestimates true infection levels due to limited diagnostic and genomic surveillance infrastructure. Middle-income countries such as South Africa, India, and Brazil fall

between these two extremes: they experienced large-scale outbreaks with notable mortality, but their genomic sequencing contributions, although present, remained considerably lower than those of high-income countries. Overall, the analysis indicates that global patterns of COVID-19 case detection, mortality reporting, and genomic surveillance are strongly influenced by national economic capacity and health system resources. Countries with greater resources were able to identify more cases and contribute more actively to global genomic monitoring, whereas resource-constrained settings faced challenges that likely obscured the full scale and evolution of the pandemic within their borders

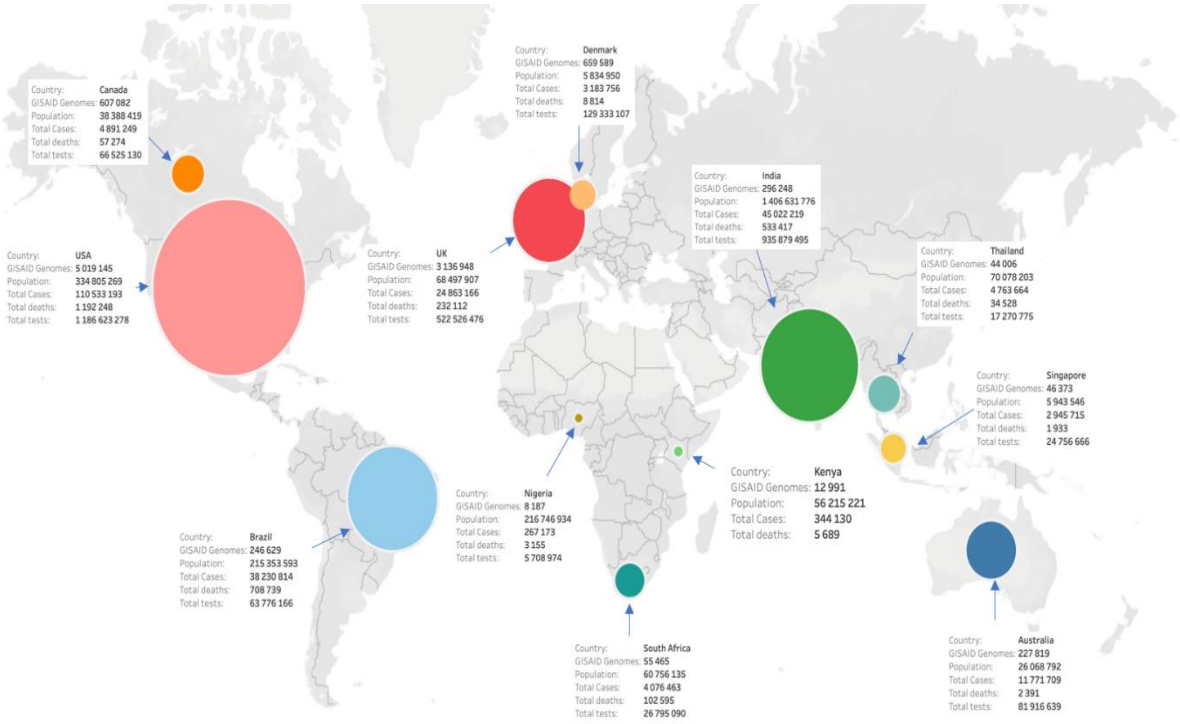


Figure 4.3: World map indicating selected countries and their respective number of genomes available on GISAID as at 19/02/2024, country population, total COVID-19 cases and deathsand total tests performed. Each circle is indicative of the total cases reported per country.

As of January 27, 2025, a total of 17,176,323 genomes from 223 countries have been submitted to GISAID since the first genome was recorded on 20 January 2020 (Wu *et al.*, 2020). In applying GISAID filters to the 16,718,242 (97,3%) genomes collected between 01 January 2020 to 31 January 2024, 16,347,201 (97,8%) were complete genomes with 5,848,259 (35,8%) genomes being high coverage. Data is plotted per continent distribution (Figure 4.4a). Selected countries (Denmark, Australia, UK, USA, Singapore, Brazil, Thailand, India. Kenya, Nigeria and South Africa) contributed to a total of 10,514,653 (64,3%) of total genomes for the period above, with 10,450,617 (99,4%) having complete genomes and 3,841,562 (36,8%) high coverage genomes. Data plotted per individual country (Figure 4.4b). The annual counts were as follows: 145, 206, 208, 184 and 105 countries/regions with genome contributions amounting to 314,451; 6,383,583; 7,804,540; 1,931,326 and 115,976 respectively.

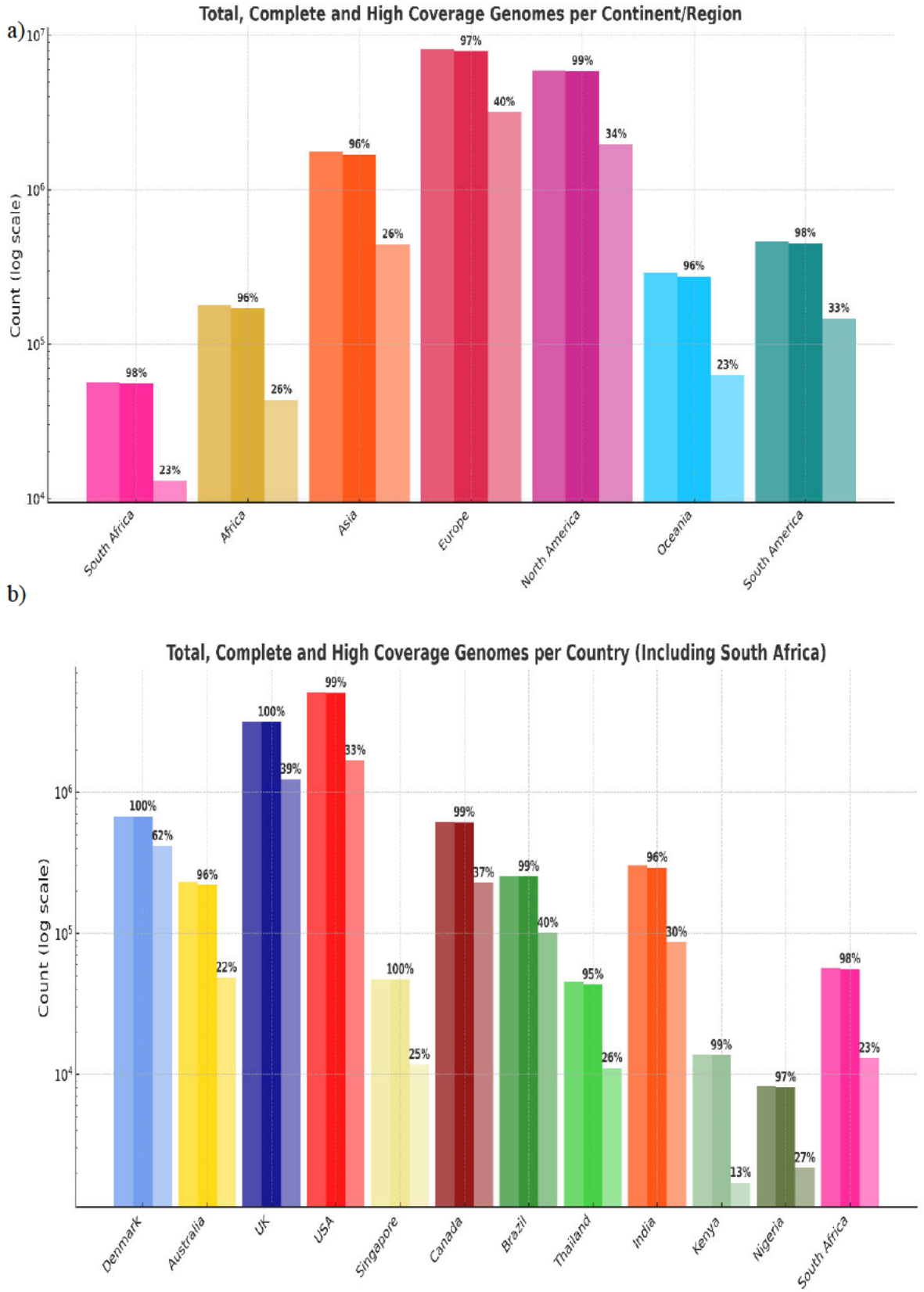


Figure 4.4: Total, Complete and High Coverage Genomic Distribution across a) Continents/Regions and b) Selected Countries

Figure 4.5 illustrates clear temporal and geographical trends in genome sharing and deposition time from 2020 to 2024. Overall, median genome deposition times declined across continents, with the most notable reductions in 2024, indicating improvements in sequencing and reporting efficiency. South Africa, highlighted in pink, follows this global pattern, showing progressively shorter deposition times over the study period. The box plots of total shared genomes show that North America, Asia, and Europe consistently contributed the highest number of sequences, while Africa and South America displayed greater variability and generally lower submission counts. Despite this, the box plots of median deposition time reveal a gradual convergence between continents, with Africa showing initially longer delays that decrease over time. These trends suggest that although substantial disparities in genomic surveillance capacity were evident in the early stages of the pandemic, reporting practices and turnaround times have improved globally, narrowing some of the initial gaps.

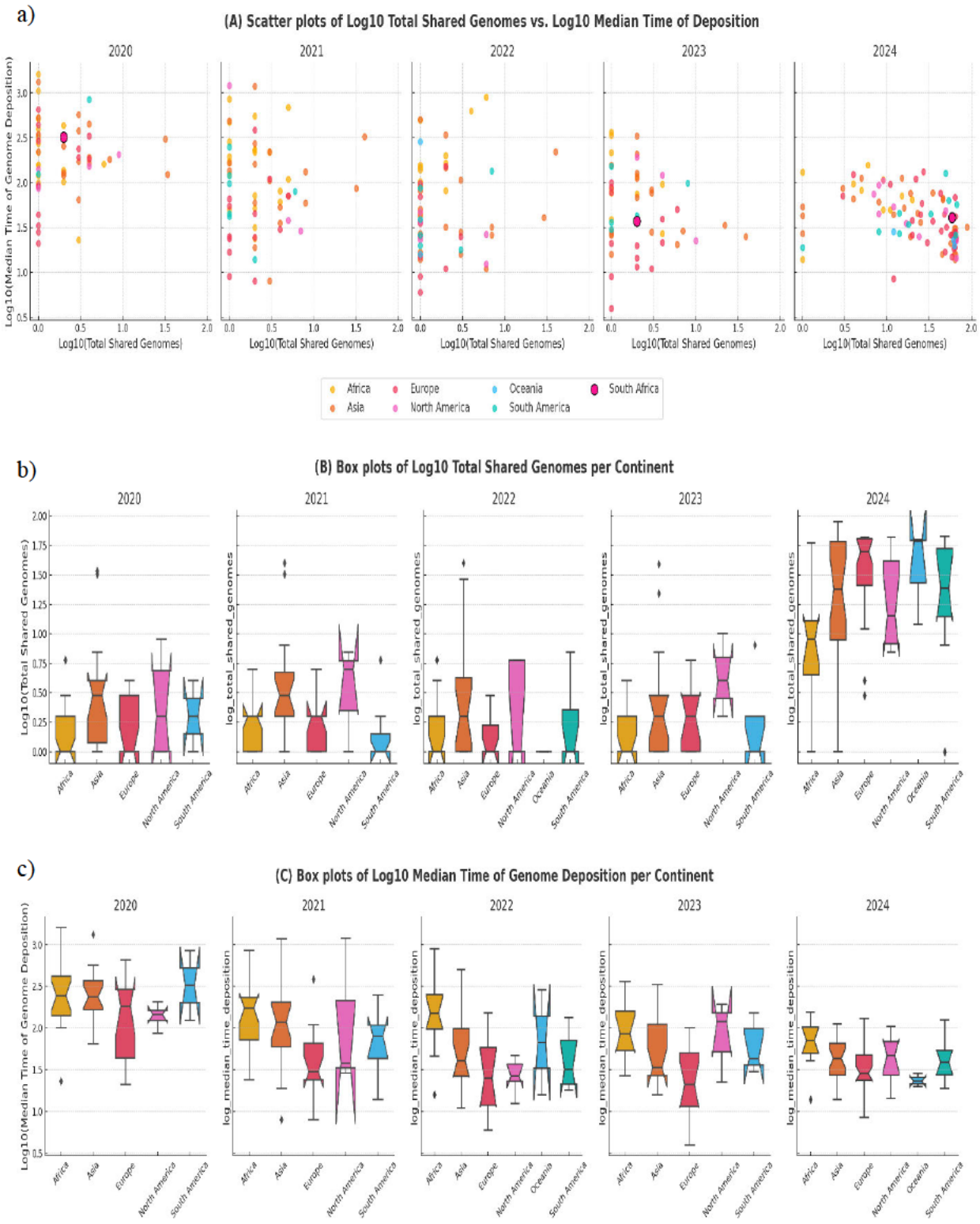


Figure 4.5: The visualisation presents data on the total shared genomes and the median genome deposition (in days) for various countries and regions between 2020 and 2024. (a) Scatter plots that depict the relationship between the total number of shared genomes and the median deposition time for each country or region across each year, (b) Box plots illustrate the log10 distribution of total shared genomes across six continents, and (c) Box plots that present the median genome deposition time categorised by continent. Each plot represents a different year, with continents/regions on the x-axis and

the log-transformed total number of shared genomes on the y-axis. The notched boxes indicate the interquartile range (IQR), with the median represented by the horizontal line inside each box, while the whiskers extend to show variability, and individual diamond markers signify outliers. South Africa is highlighted in pink for emphasis. The median deposition time represents the duration between sample collection and genome submission.

The temporal patterns in genome submissions show distinct shifts in sequencing activity over the course of the pandemic. Monthly trends indicate that 2021 and 2022 were periods of the highest global genome sharing, coinciding with major variant waves, followed by a marked decline in 2023 and 2024 as case numbers stabilised and global sequencing demand reduced. South Africa follows the global trajectory but at a much smaller scale, reflecting more limited sequencing capacity despite consistent participation. The comparison of deposition delays across countries highlights substantial variation in how quickly genomes were uploaded to GISAID. High-income countries such as Denmark, the United Kingdom, and Singapore show short median deposition times, suggesting efficient sequencing pipelines and data-sharing practices. In contrast, countries such as Kenya and Nigeria exhibit substantially longer delays, indicating resource and infrastructural constraints that affect real-time surveillance capacity. South Africa's median deposition time sits in the mid-range, demonstrating improved turnaround relative to some African peers but still slower than regions with stronger sequencing infrastructure. Overall, the results in Figure 4.6 show that while global sequencing effort peaked early in the pandemic, disparities in the speed of genome reporting persisted, shaped largely by differences in laboratory capacity and public health resourcing.

Considering that the median days to deposition on GISAID were plotted since 10 January 2020, the distribution of data clearly indicates a reduction in the duration between sampling and genomic data sharing to approximately 10 days, which is significantly shorter than the estimated average of 37 days.

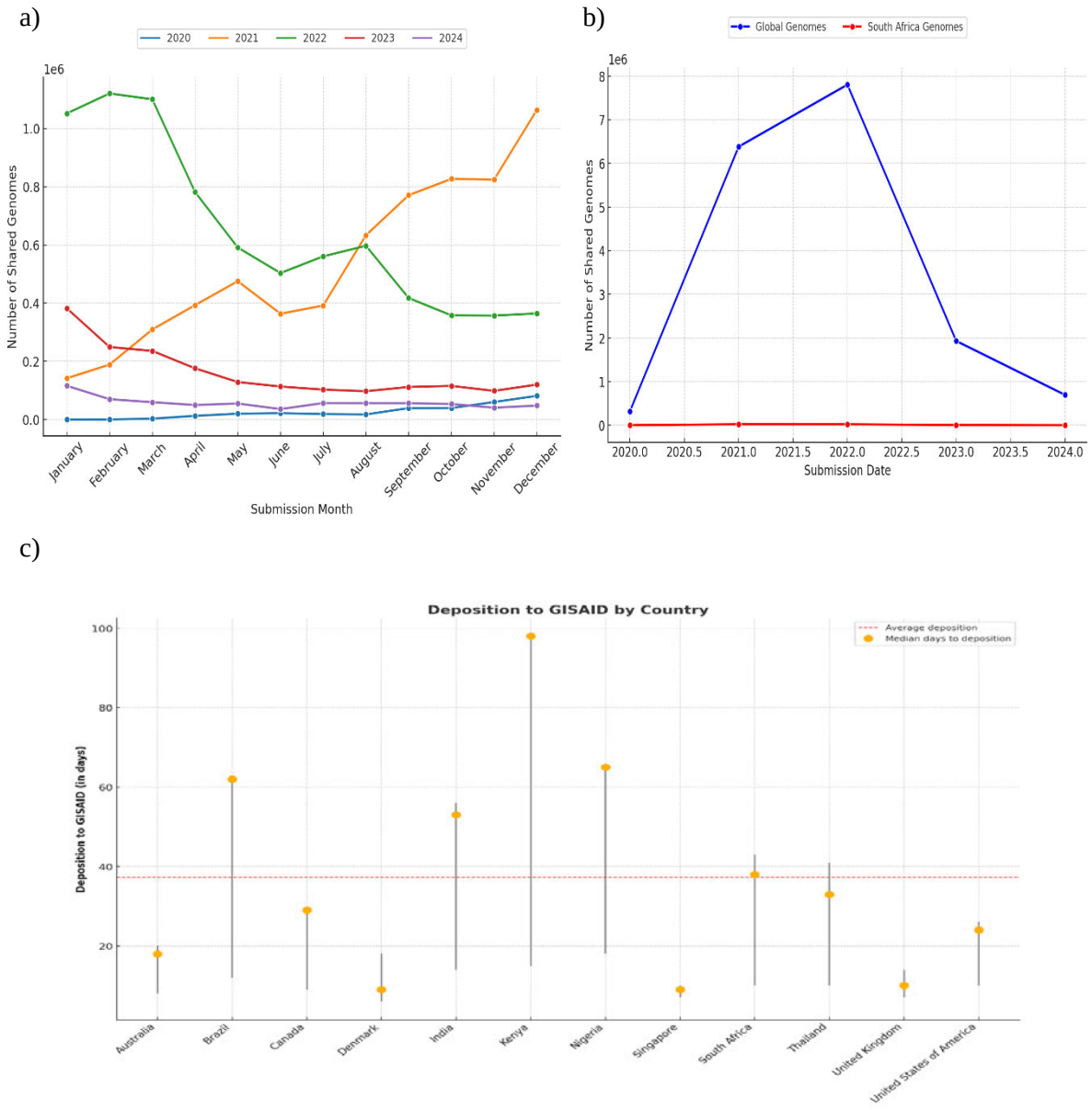
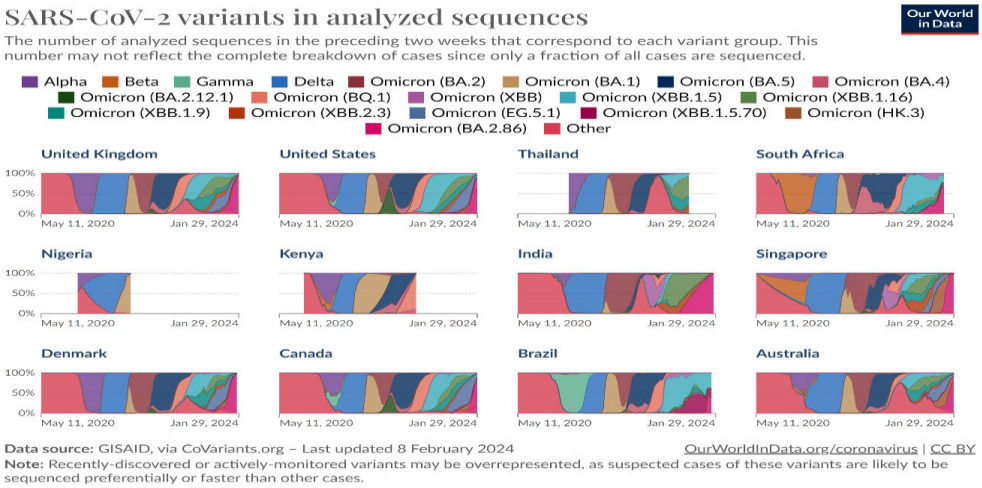


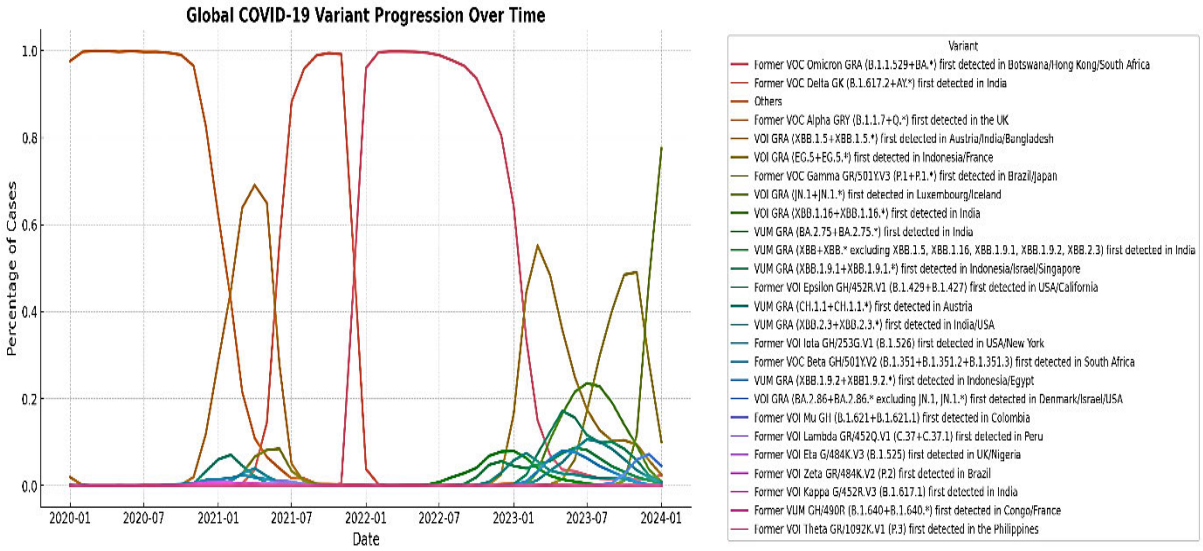
Figure 4.6: The number of shared genomes in South Africa with global trends and turnaround times per country. a) The number of shared genomes in South Africa each month during 2020–2024, b) The median number of shared genomes in South Africa and global trends from 2020 to 2024, and c) Average sequencing turnaround time per country as per GISAID submission tracker (date accessed: 27 January 2025. The circles represent the median days from specimen collection to GISAID deposition and the bars indicate value distribution per country. The red dashed line represents an overall average turnaround time across all countries listed. Data obtained from GISAID submission tracker.

SARS-CoV-2 variants shifted in clear waves globally, with Delta replacing earlier variants in 2021 and Omicron subsequently displacing Delta and diversifying into multiple sublineages that have remained dominant through 2024 (Figure 4.7).

a)



b)



c)

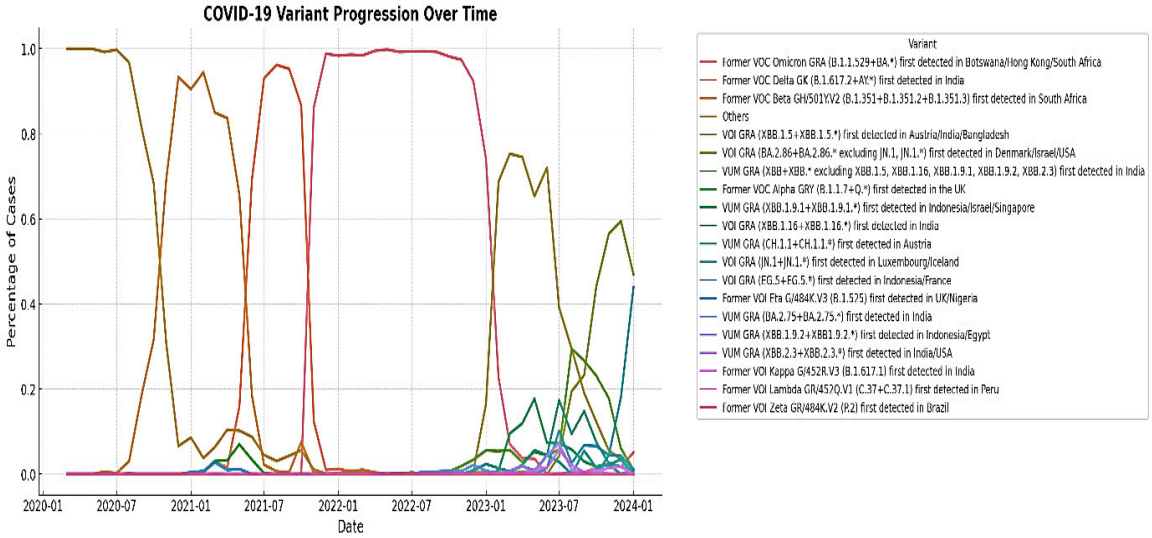


Figure 4.7: SARS-CoV-2 variant prevalence by a) country based on genomic sequencing data from OWID at https://ourworldindata.org/ (date accessed: 19/02/2024), and b) Variant progression over time globally and c) in South Africa (data from GISAID)

The variant progression plots show clear shifts in SARS-CoV-2 lineage dominance over time across countries and globally. Early in the pandemic, multiple lineages circulated at low levels, before the emergence of highly transmissible variants led to rapid replacement patterns. The Alpha and Beta variants dominated in early 2021 in regions such as the United Kingdom and South Africa, respectively, followed by the global rise of Delta in mid-2021. In late 2021, Delta was rapidly displaced by Omicron, which has remained the predominant variant group through 2022 and 2023, although with substantial diversification into multiple Omicron sublineages. Countries such as the United States, United Kingdom, Denmark, and Singapore show rapid turnover between Omicron subvariants, reflecting more intensive genomic surveillance. In contrast, countries with lower sequencing frequency, including Kenya, Nigeria, and Thailand, display broader periods of apparent variant dominance, likely reflecting lower resolution in detection..

Table S4.1 and Figure 4.8 provides insight into testing rates, case prevalence, the proportion of global cases, and sequencing intensity across countries between January 2020 and 2024. In Figure 4.8a, a moderate to strong positive correlation ($r = 0.72$, $p = 0.008$) is observed between prevalence rate and testing rate per 1,000 persons, indicating that countries conducting more tests detected more cases and therefore reported higher prevalence. Figure 4.8b shows that the United States accounted for the largest share of global reported cases, followed by India, Brazil and the United Kingdom. In Figure 4.8c, sequencing rates reveal notable differences in genomic surveillance capacity: Denmark sequenced over 20% of all reported cases, with the United Kingdom and Canada also sequencing comparatively high proportions. In contrast, the higher sequencing percentages in Kenya and Nigeria likely reflect the combination of low testing coverage and smaller numbers of reported cases, resulting in a higher proportion of those limited cases being sequenced rather than indicating extensive genomic surveillance. Overall, these findings show that countries with higher testing capacity not only detect more infections but also sequence a greater proportion of cases, while lower-resourced settings tend to have both reduced case detection and more restricted genomic surveillance efforts.

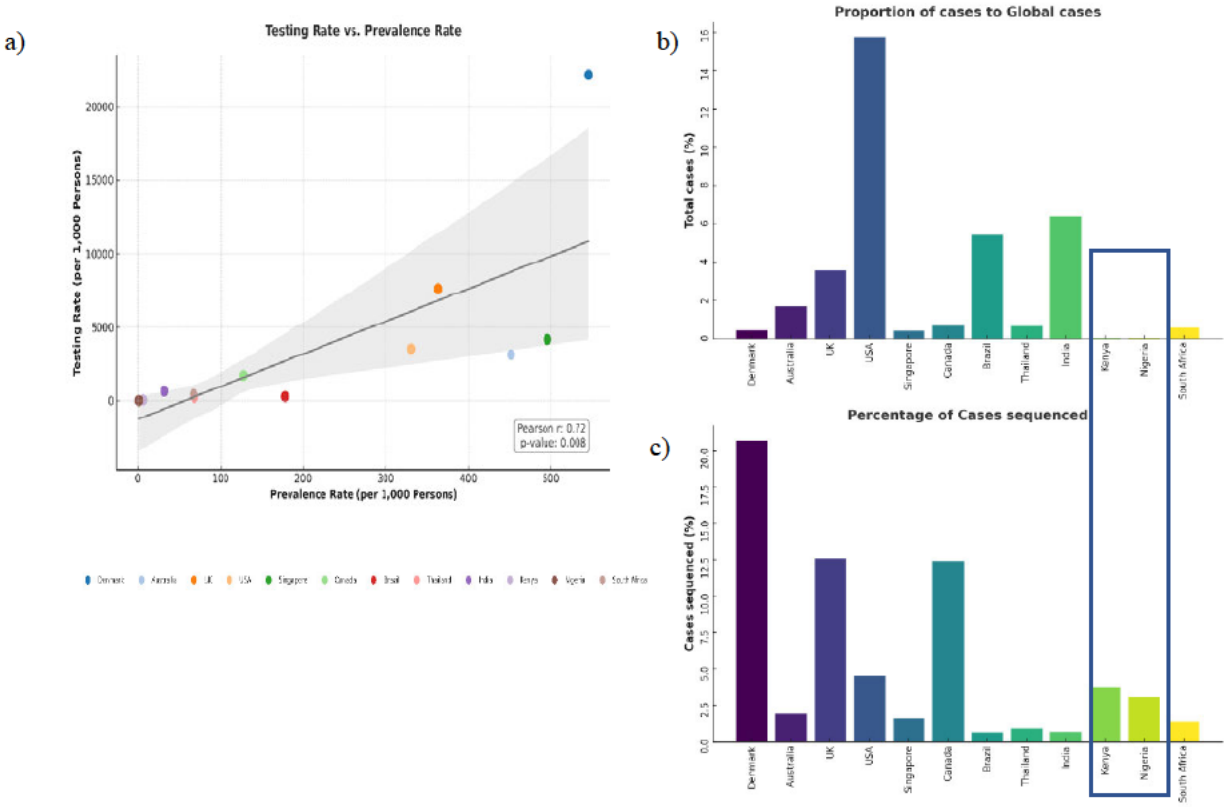


Figure 4.8: a) Correlation of SARS-CoV-2 testing and prevalence rates per 1,000 person across the different socioeconomic countries from January 2020 to January 2024, b) Proportion of global cases and c) genomic surveillance efforts (sequencing rates). Higher sequencing percentages in Kenya and Nigeria reflect low case detection rather than extensive genomic surveillance.

4.3.2 Overview of SARS-CoV-2 Publication data

The subsequent figures encapsulate our findings regarding the global publication landscape, South African contributions, the diversity of funding sources and the leading academic institutions with respect to SARS-CoV-2 research publications. These illustrations aimed to underscore the achievements and obstacles encountered by South Africa in the realm of scientific research, particularly from the standpoint of a lower-middle-income country (LMIC). In both, Scopus and WoS, the most common type of publications were research articles, comprising of 69,20% and 74,37% of the total, respectively (Figure 4.7a). In figure 4.7b) a comparative analysis of the number of publications in Scopus versus WoS reveals an initial rise in publications from 2020 to 2022, followed by a decline by 2024 in both databases. Figure 4.7c) extends the comparison to the selected countries. Review papers, editorial material, letters, books and others followed in lower volumes.

As tabulated in Table 4.1, we provide a snapshot of the most prominent documentation types as well the journal sources of South African publications. Figure 4.8 a) shows the top 10 funding sources that supported South African research and b) provides an overview of the top achieving academic institutions in terms of SARS-CoV-2 publication output in South African.

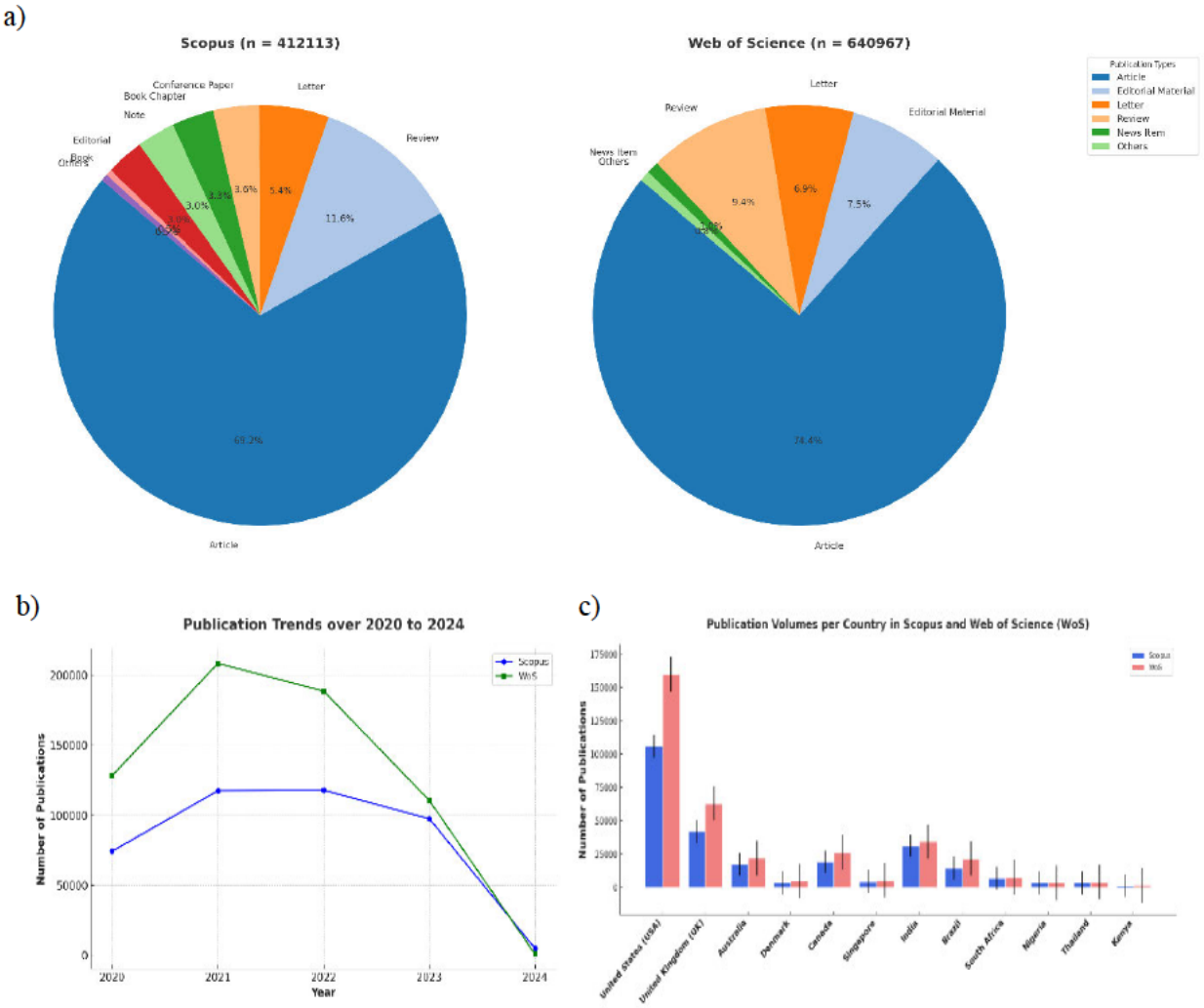


Figure 4.9: Distribution of Global COVID-19 research output, as defined by Scopus and WoS where data is represented by a) document types b) year and c) country.

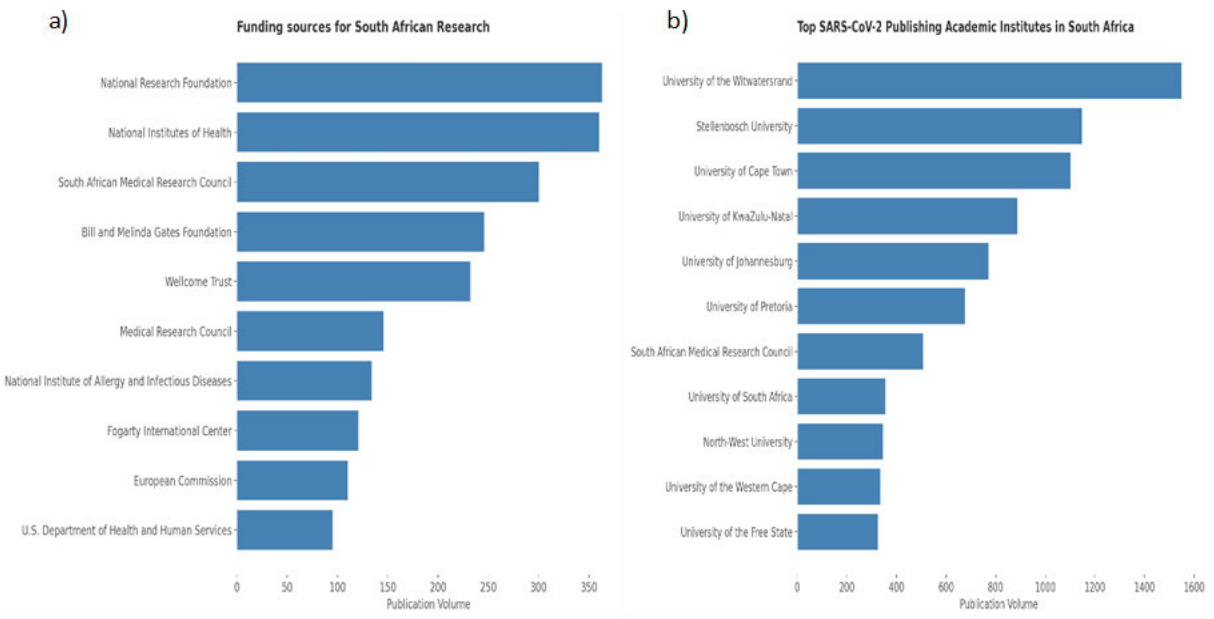


Figure 4.10: a) Top ten funding sponsors and b) academic institutions in terms of SARS-CoV-2 research in South Africa until January 2024

Table 4.1: Document type and sources published in South Africa (WoS)

DOCUMENT TYPE	Publication volumes
Article	4518
Review	938
Book Chapter	469
Note	265
Letter	187
Editorial	178
Conference Paper	127
Book	31
Short Survey	26
Data Paper	5
Retracted	2
Erratum	1
SOURCE TITLE	
South African Medical Journal	202
International Journal of Environmental Research And Public Health	122
Plos One	105
Lancet	80
Hts Teologiese Studies Theological Studies	68
Sustainability Switzerland	65
BMJ Global Health	59
Vaccines	59
Pan African Medical Journal	54
Frontiers In Public Health	52
South African Family Practice	47

1

4.4 Discussion

Although GISAID serves as the primary global repository for SARS-CoV-2 genomic data and remains the dominant platform used for data sharing in South Africa and many collaborating countries, reliance on GISAID alone introduces certain limitations. Not all genomic laboratories or surveillance networks upload data exclusively to GISAID; some sequences may also be deposited in alternative repositories such as NCBI GenBank or the European Nucleotide Archive (ENA). As a result, there is a possibility that some regional or temporally relevant data were not captured in this analysis. In addition, variations in institutional upload capacity, data-sharing policies, and delays in sequence submission may further influence the representativeness of the dataset. Therefore, while GISAID provides a robust and widely adopted framework for comparative genomic surveillance, the findings presented here should be interpreted with the understanding that the dataset may not reflect the full spectrum of sequencing activity across all laboratories and national platforms.

The findings of this study highlights the role of genomic surveillance in managing the SARS-CoV-2 pandemic, with South Africa playing a pivotal role in global efforts. A comparative analysis reveals disparities in genomic data sharing capabilities across different socioeconomic settings, emphasising the urgent need for equitable access to sequencing technologies and resources (Brito *et al.*, 2022). South Africa's genomic surveillance system, supported by organisations like the Network for Genomic Surveillance in South Africa (NGS-SA) (Msomi *et al.*, 2020), has demonstrated significant capacity in tracking viral evolution, identifying variants of concern (VOCs), and informing public health interventions. However, challenges remain in ensuring consistent data quality, quantity, and timely reporting across all regions.

High-income countries like as the United Kingdom, the United States, Australia, and Denmark benefit from advanced sequencing infrastructure, extensive government funding, and well-integrated healthcare systems that enable high sequencing coverage and rapid turnaround times (Burns *et al.*, 2019; Stark *et al.*, 2019; Struelens *et al.*, 2024). Their ability to generate and share high-quality genomic data in real-time has significantly influenced global pandemic responses (Tosta *et al.*, 2023). These countries have successfully integrated genomic surveillance into routine public health frameworks, demonstrating the impact of continuous investment in cutting-edge sequencing technologies (Gim, 2022; Hoang *et al.*, 2022; Akande *et al.*, 2023; Roman, 2023). For example, the UK's COG-UK consortium exemplifies this integrated approach (Marjanovic *et al.*, 2022). Middle-income countries like South Africa, Brazil, India, and Thailand exhibit substantial advancements in genomic surveillance despite facing resource constraints (Konono *et al.*, 2024). South Africa, in particular, has played a pivotal role in the early detection of key SARS-CoV-2 variants, including Beta and Omicron (Cele *et al.*, 2021; Tegally *et al.*, 2021; Ahmed *et al.*, 2022). Through strong public-private collaborations and international partnerships, South Africa has optimised its genomic sequencing capabilities, aligning itself with global best practices

(de Oliveira *et al.*, 2022). However, persistent challenges such as disparities in sequencing coverage across different provinces and delayed data reporting need to be addressed to further enhance the efficiency of its genomic surveillance system (Staunton, Swanepoel and Labuschaigne, 2020; Belman, Saha and Beale, 2022; Pronyk *et al.*, 2023). Low-income countries such as Kenya and Nigeria face more profound difficulties in genomic surveillance due to inadequate infrastructure, limited funding, and low sequencing capacities (A. Lambisia *et al.*, 2022; Khemchandani, 2022; Ozer *et al.*, 2022). These challenges lead to inconsistencies in genomic data quality and slower detection of emerging variants (Bosonkie *et al.*, 2023; Pronyk *et al.*, 2023). However, international collaborations and capacity-building initiatives are helping to bridge some of these gaps. South Africa's model of leveraging existing infectious disease expertise, particularly from its HIV and tuberculosis research programs, provides a framework that other low-resource settings could adopt to strengthen their genomic surveillance efforts (Eike *et al.*, 2022).

Analysis of SARS-CoV-2 genomic data sharing trends between 2020 and 2024 reveals significant improvements in sequencing efficiency and global data deposition times (Chen *et al.*, 2022; Ekusai-Sebatta *et al.*, 2023; Feng *et al.*, 2024). The global decline in genome deposition time suggests advancements in sequencing turnaround and reporting mechanisms. South Africa, while initially exhibiting higher deposition times, has progressively improved in this regard, aligning with global trends (Tegally *et al.*, 2023; *GISAID - Initiative*, 2024). Despite these gains, regional inconsistencies within South Africa highlight the need for uniform capacity-building strategies to ensure timely data sharing across all provinces. Genomic surveillance has been instrumental in tracking variant evolution and informing public health responses (Hughes *et al.*, 2022; Attar Cohen *et al.*, 2023). South Africa's ability to rapidly detect and report major variants, such as Beta and Omicron, underscores the effectiveness of its surveillance system. However, global trends indicate the emergence of newer sub-variants (e.g., XBB, BA.2.86, and EG.5), emphasising the necessity for continuous genomic monitoring (Esmaeilzadeh *et al.*, 2024). While high-income countries have the resources to conduct large-scale sequencing, middle- and low-income nations must strengthen their surveillance frameworks to avoid being disproportionately affected by delayed variant detection (De Jong *et al.*, 2023).

The study reveals a strong positive correlation between SARS-CoV-2 testing rates and reported prevalence. Countries with higher testing rates, such as Denmark and the United Kingdom, have been able to sequence and share a greater proportion of viral genomes (Brito *et al.*, 2022; Chen *et al.*, 2022). In contrast, lower testing rates in Kenya and Nigeria have led to an underestimation of actual case numbers, skewing sequencing rates (Jacobs and Okeke, 2022; Waitumbi *et al.*, 2022). This discrepancy underscores the need for enhanced diagnostic testing infrastructure alongside genomic surveillance to ensure accurate epidemiological assessments. Scientific research output related to SARS-CoV-2 has followed a distinct trajectory, with an initial surge in publications from 2020 to 2022, followed by a gradual decline by 2024. South Africa has contributed significantly to global research, particularly in

the fields of genomic surveillance, variant analysis, and public health interventions. Key funding sources and leading academic institutions have played a crucial role in driving research efforts, highlighting the importance of sustained investment in scientific innovation (Fu, Chou and Wang, 2022).

Several key strategies can strengthen South Africa's genomic surveillance framework. Investment in infrastructure, including expanding laboratory networks and improving data processing efficiency, will enhance real-time genomic surveillance capabilities (Inzaule *et al.*, 2021; Wilkinson *et al.*, 2021). Continued capacity-building initiatives, such as training scientists and technicians in bioinformatics and genomic sequencing, will ensure a skilled workforce (Mboowa, Sserwadda and Aruhomukama, 2021; Kotze, Mashamba-Thompson and Stephens, 2023; Uvere *et al.*, 2023). Strengthening regional and international collaborations through partnerships with high-income countries and international organisations can facilitate knowledge exchange, technology transfer, and funding support (Knoppers, Abdul-Rahman and Bédard, 2007; Varshney *et al.*, 2016). Integrating genomic data seamlessly into public health policies will improve pandemic preparedness and response (Mboowa and Sserwadda, 2019). Finally, addressing data quality and reporting timeliness by standardising sequencing protocols and improving data-sharing frameworks will enhance the reliability and accessibility of genomic information (Jamin *et al.*, 2021).

4.5 Conclusion

South Africa's genomic surveillance efforts during the SARS-CoV-2 pandemic provide a compelling case study, showcasing how strategic investments, capacity building, and data integration can yield significant achievements. While South Africa has made remarkable progress and contributed substantially to the global response, disparities in sequencing capacity, data quality, and reporting efficiency persist. Addressing these challenges, alongside those faced by other emerging economies, requires targeted investments, technological advancements, and robust capacity-building initiatives. By adopting best practices and fostering stronger collaboration networks, South Africa can solidify its leadership in African genomic surveillance. Bridging these gaps will not only enhance future responses to SARS-CoV-2 but also bolster global preparedness for emerging infectious diseases, ultimately improving the timeliness and effectiveness of pandemic responses worldwide.

Chapter Five: General Discussion, conclusions, future recommendations and study limitations

5.1 General Discussion

In this final chapter of the study, I explore the complex dimensions of genomic surveillance and the analysis of SARS-CoV-2, as presented across three preceding, investigative chapters. I proceed to outline future recommendations, highlighting areas primed for further exploration and proposing platforms that could enhance the depth of subsequent investigations. Finally, I acknowledge the limitations inherent in the study, reflecting on the constraints that bound the research scope and the potential implications these limitations may have on the generalisability and applicability of the findings. Through this discourse, I aim to make a substantial contribution to the field of genomic surveillance, laying a solid foundation for future research efforts. My goal is to support the enhancement and preparation for future outbreaks, ensuring that the scientific community is better equipped to respond effectively and in a timely manner.

Chapter two presents the study's first objective which is to conduct an assessment and comparison of SARS-CoV-2 genomic data. This chapter, entitled "Reflective evaluation of Next-Generation Sequencing data during early phase detection of the Delta variant," employs a retrospective data analysis from two NGS platforms, Illumina MiSeq and Ion Torrent (Ramphal *et al.*, 2024). These platforms Ire pivotal during the peak of the pandemic, with the analysis utilising online software tools, Genome Detective and NextClade, to examine data variability (Vilsker *et al.*, 2019; Aksamentov *et al.*, 2021). The fundamental paradigm in this chapter was of technological and analytical advancements. Utilising advanced sequencing technologies and bioinformatics tools is key to swiftly identifying VOCs and evaluating their potential public health impacts. Yet, the integrity and quality of the data generated by these technologies, critical for comparison and research, often goes under examined. Data quality discrepancies between technologies prompted us to analyse the same sample set on two NGS platforms: the Ion Torrent S5 and Illumina MiSeq (Gozashti and Corbett-Detig, 2021; Plitnick *et al.*, 2021). The analysis, both retrospective and comprehensive, highlights the crucial role of genomic surveillance in tracking and mitigating the spread of SARS-CoV-2, while pointing out the distinct strengths and weaknesses of each platform. The Ion Torrent S5 excels in sequencing samples with low viral loads, offering genomic coverage, which makes it particularly useful for early detection of infections. In contrast, the Illumina MiSeq stands out for its efficient library preparation, making it a steadfast tool for ongoing genomic surveillance efforts. The findings call for ongoing enhancements in sequencing technology, primer design and bioinformatics, to improve the precision and speed of genomic surveillance. Addressing identified shortcomings, such as the necessity for updated primer sets and overcoming hurdles in real-time variant tracking, is vital for a deeper understanding of the virus's evolution and prompt public health interventions. Ultimately, both the Ion Torrent S5 and Illumina MiSeq, when integrated with suitable software, provide complementary benefits for the genomic monitoring of SARS-CoV-2 (Pritchard *et al.*, 2022). This study provides critical insights into the

comparative efficacy and data quality of genomic epidemiology methods, serving as a foundational guide for enhancing future pandemic preparedness and response strategies.

Chapter three progresses to the study's second objective, which is to examine existing SARS-CoV-2 genomic surveillance trend data within a Ill-annotated South African province. This chapter entitled “Retrospective Evaluation of SARS-CoV-2 Genomic Surveillance Data from KwaZulu-Natal, 2020-2023” examines the geographical, demographical and epidemiological trends in the region. It emphasises sequencing trends, turnaround times and the distribution of SARS-CoV-2 lineages. This study aimed to shed light on district-level transmission patterns and pinpoint potential limitations, contributing to refining future genomic surveillance strategies. The study conducted on samples from KZN in South Africa provides a comprehensive analysis of SARS-CoV-2 genomic epidemiology through routine genomic surveillance data, emphasising the region's unique demographic and socio-economic challenges. This research underscores the importance of genomic surveillance in understanding the virus's transmission dynamics, the diversity of circulating lineages and the implications for public health strategies (Raghwani *et al.*, 2021; Lambrou *et al.*, 2022; Espinoza *et al.*, 2023; Tajudin *et al.*, 2024). The commitment to genomic surveillance in KZN, as evidenced by the collection of 6089 genomic sequences across all 11 districts, underlines the critical role of this approach in guiding public health responses to the pandemic. The identification of 134 individual lineages, including major variants such as Alpha, Beta, Delta and Omicron, along with their sub-lineages, highlights the viral diversity within KZN. This diversity is crucial for understanding the virus's evolution and the potential emergence of new variants that could influence public health strategies. The analysis of case fatality rates and genomic surveillance rates across districts reveals the complex interplay between demographics, healthcare capacity and surveillance efforts in determining the pandemic's impact. The urban district of eThekwini, for instance, demonstrated an effective healthcare response and robust genomic surveillance, despite a high incidence of COVID-19 cases. The demographic analysis, revealing a predominance of females and certain age groups in the genomic sequences, alongside the urban-rural divide in viral spread and diversity, underscores the need for tailored public health strategies. Urban centres like eThekwini, with their high population density and mobility, present different challenges and opportunities for viral containment compared to rural areas, which may face issues related to healthcare access and slower transmission rates (Robishaw *et al.*, 2021; Walker *et al.*, 2022; Napit *et al.*, 2023). The study recognises the limitations related to the exclusion of some samples and biases in sample collection. These challenges underscore the critical importance of complete, high-quality metadata, as Ill as mandatory quality checks on data capturing efforts (Gozashti and Corbett-Detig, 2021). Such standards are essential to ensure the data’s comprehensiveness and representativeness, thereby enabling more meaningful analysis and insights (Plasek *et al.*, 2020; Seitz, 2020). In conclusion, this research offered valuable contributions to the global body of knowledge on COVID-19, highlighting the critical role of genomic surveillance in informing and shaping effective

public health responses in urban and rural settings. The ongoing evolution of SARS-CoV-2 and the emergence of new variants require ongoing research and adaptable public health strategies to manage the pandemic and future outbreaks effectively.

The concluding investigative chapter addresses the study's third objective. The study's investigation into SARS-CoV-2 genomic surveillance in South Africa underscores the critical importance of data sharing in ensuring effective pandemic response strategies. Chapter 4, titled "Unveiling South Africa's Triumphs: An Exploration of SARS-CoV-2 Genome Sharing Trends and Impacts, 2020–2024," provided a comparative analysis of South Africa's genomic data sharing relative to global trends. The findings demonstrated both the successes and the limitations of South Africa's surveillance efforts, offering a roadmap for refining future outbreak preparedness strategies. South Africa has established itself as a leader in genomic surveillance within the African continent, contributing significantly to global databases such as GISAID. Early detection of key SARS-CoV-2 variants, including Beta (B.1.351) and Omicron (B.1.1.529), showcased the country's robust sequencing infrastructure and collaborative networks (Tegally *et al.*, 2021; Tegally *et al.*, 2021). However, the study also revealed disparities in sequencing coverage across different regions within South Africa, highlighting the need for uniformity in data collection and reporting (Moodley *et al.*, 2022). A comparative analysis with high-income countries such as the United Kingdom, United States, Australia, and Denmark illustrated how Ill-funded infrastructure and advanced sequencing capabilities facilitate more extensive and timely genomic surveillance (Lambrou *et al.*, 2022). In contrast, countries with similar economic standing to South Africa, such as Brazil, India, and Thailand, demonstrated varying degrees of success in genomic data sharing, influenced by factors such as resource allocation, healthcare infrastructure, and policy integration (Cimerman *et al.*, 2021; Giovanetti *et al.*, 2022; Horovitz, Félix and Ferraz, 2024). Additionally, low-income countries such as Kenya and Nigeria faced persistent challenges in maintaining consistent sequencing efforts due to financial and logistical constraints. Despite these challenges, the study observed increasing international collaborations aimed at addressing these disparities and promoting more inclusive global genomic surveillance frameworks. The study focused on three critical metrics of genomic surveillance: quantity, quality, and timeliness (Feng *et al.*, 2024). A) Quantity - South Africa has contributed a substantial number of SARS-CoV-2 sequences to international databases, aligning itself with middle-income countries that have made notable progress in genomic surveillance. However, sequencing efforts fluctuated throughout different pandemic waves, influenced by funding availability, sample collection constraints, and shifts in research priorities. B) Quality - Data completeness and metadata accuracy Ire identified as crucial factors in determining the utility of genomic information. While South Africa met global data quality standards, disparities in metadata completeness across provincial datasets indicated the need for enhanced training and standardisation. C) Timeliness - The chapter revealed a progressive reduction in genome deposition time, aligning with global trends of improved sequencing efficiency. However,

some regions within South Africa exhibited longer deposition times, which could be mitigated through increased automation and streamlined data submission processes (Brito *et al.*, 2022; Chen *et al.*, 2022; Walker *et al.*, 2022; Feng *et al.*, 2024). The COVID-19 pandemic underscored the indispensable role of genomic surveillance in real-time outbreak monitoring and response. The findings from Chapter 4 reinforce the necessity of sustained investment in genomic infrastructure, particularly in middle-income countries striving to balance resource limitations with global health contributions (Forero *et al.*, 2016). Strengthening regional and international collaborations is essential, as expanding cross-border partnerships will facilitate knowledge exchange, technology transfer, and financial support, helping countries with limited sequencing capacities to bridge existing gaps (Yeh *et al.*, 2019). Enhancing data quality and reporting standards through standardised protocols for metadata completeness and uniform sequencing practices will improve the reliability and comparability of genomic data across different regions (Gozashti and Corbett-Detig, 2021). Additionally, investing in workforce development by implementing training programs focused on bioinformatics and sequencing methodologies will equip local scientists with the expertise necessary to sustain and expand genomic surveillance initiatives (Attar Cohen *et al.*, 2023; Kotze, Mashamba-Thompson and Stephens, 2023; Uvere *et al.*, 2023). Furthermore, integrating genomic surveillance with public health decision-making is crucial to ensuring that sequencing efforts translate into actionable insights, guiding interventions such as vaccine development and targeted containment measures (Robishaw *et al.*, 2021; Domman *et al.*, 2022; Struelens *et al.*, 2024). The comparative analysis highlighted the substantial impact of economic disparities on genomic surveillance capacity. High-income nations demonstrated rapid turnaround times and extensive sequencing coverage due to Ill-funded public health infrastructures (De Jong *et al.*, 2023). In contrast, middle- and low-income countries faced challenges such as inconsistent funding, logistical constraints, and fragmented healthcare systems (Tekola-Ayele and Rotimi, 2015; Belman, Saha and Beale, 2022). To address these disparities, strategic interventions should include increasing government and private sector investment, as sustained funding from national budgets and private-public partnerships will enhance genomic surveillance infrastructure, making it more resilient to future public health emergencies (Maison *et al.*, 2021; Robishaw *et al.*, 2021). Leveraging cost-effective sequencing technologies, such as portable sequencing platforms and automated bioinformatics pipelines, could reduce the financial burden associated with large-scale sequencing initiatives. Lastly, developing equitable data-sharing policies will ensure that all countries, regardless of economic status, have access to global genomic data repositories, promoting collaborative research, transparency and improving global outbreak response mechanisms (Plasek *et al.*, 2020; Seitz, 2020; Robishaw *et al.*, 2021; Ling-Hu *et al.*, 2022). Chapter 4 provided a comprehensive evaluation of South Africa's SARS-CoV-2 genomic surveillance efforts, highlighting both achievements and areas requiring improvement (de Oliveira *et al.*, 2022). The synthesis of these findings reinforces the importance of genomic data sharing in guiding effective pandemic responses and emphasises the need for sustained investments in sequencing infrastructure, workforce training, and international collaboration. While South Africa has made

commendable strides in genomic surveillance, future efforts must focus on addressing regional disparities, enhancing data reporting standards, ensuring the seamless integration of genomic insights into public health policies (Inzaule *et al.*, 2021). By leveraging lessons learned from the COVID-19 pandemic, South Africa and other middle-income nations can build resilient genomic surveillance frameworks that not only address current challenges but also strengthen preparedness for future infectious disease outbreaks.

In support of the overall findings and observations, studies have highlighted several aspects in terms of improving science in Africa (de Oliveira *et al.*, 2022; Mfuh, Abanda and Titanji, 2023; de Oliveira and Baxter, 2024). The papers emphasise the critical need to strengthen genomic capacity in Africa, particularly in response to climate change affecting pathogen behaviour. It underscores the importance of involving the continent's predominantly young population in scientific research to tackle health challenges effectively. Despite Africa bearing a significant disease burden, its scientific output remains low, highlighting the urgency of boosting research capacity (Kramer and Libhaber, 2018; Confraria and Wang, 2020). The studies also underscore the importance of bolstering genomics infrastructure and empowering local scientists as crucial steps for enhancing global health security and addressing emerging pathogens (de Oliveira and Baxter, 2024).

South Africa played a pivotal role in genomic surveillance during the SARS-CoV-2 pandemic, contributing to our understanding of the virus's transmission dynamics and evolution. The NGS-SA led studies focusing on the early transmission dynamics of the virus, revealing a heterogeneous spread with multiple distinct introductions (Giandhari *et al.*, 2021). Notably, in KwaZulu-Natal, this phase was marked by hospital-acquired transmission and early mortality, characterised by numerous international introductions, predominantly involving lineages B.1 and B at the time (Lessells, Moosa and de Oliveira, 2020; Tegally *et al.*, 2021). This included instances of locally acquired infections within a Durban hospital, underlining the complexity of the pandemic's spread within the country (Lessells, Moosa and de Oliveira, 2020). Similarly, in Cape Town, genomic surveillance played a crucial role in tracking at least nine early introductions of the virus, including an early-localised transmission in a shopping centre (Engelbrecht *et al.*, 2021). This detailed tracking exemplified the effectiveness of genomic surveillance in understanding the initial spread events and circulating variants. Additionally, South African researchers successfully adapted Illumina protocols to improve the speed and accuracy of outbreak investigations (Pillay *et al.*, 2020). This adaptation exemplified the nation's agility in genomic research during urgent public health crises. Moreover, the identification of novel SARS-CoV-2 lineages, the rapid detection of new variants like Beta and Omicron and the implementation of commercial genotyping assays for monitoring circulating variants, all highlight the robustness of South Africa's genomic surveillance system. This system has not only been critical for tracking the virus's spread within South Africa but has also provided key insights into the pandemic's dynamics across the African continent (Tegally, Wilkinson, Lessells, *et al.*, 2021; Umunnakwe *et al.*, 2022). Additionally, the use

of tandem targeted RT-PCR-based genotyping assays and whole genome sequencing for variant surveillance reflects the adaptability of the surveillance system in South Africa (Pinkhover *et al.*, 2022). Furthermore, the study of reinfection trends in the context of the emergence of the Omicron variant provides crucial insights into the dynamics of viral spread and immune evasion in South Africa (Pulliam *et al.*, 2022). The nation's contributions have been instrumental in expanding our understanding of SARS-CoV-2's spread and evolution, both within South Africa and on a larger scale across Sub-Saharan Africa (Tegally, San, *et al.*, 2022). Through these collective efforts, South Africa has underscored its critical role in genomic surveillance, offering valuable insights into the pandemic's spread and evolution.

The history of coronaviruses, marked by outbreaks such as SARS-CoV (2002-2003), MERS-CoV (2012-present) and SARS-CoV-2 (2019-present), demonstrates the critical role of effective surveillance systems (Da Costa *et al.*, 2021). Poor surveillance, characterised by delayed detection, inadequate global health infrastructure, insufficient genomic monitoring and slow information sharing, significantly contributed to the spread and impact of these viruses (De Wit *et al.*, 2016; Yin and Wunderink, 2018; Da Costa *et al.*, 2021). The COVID-19 pandemic highlighted these deficiencies, demonstrating the urgent need for robust, coordinated and transparent global surveillance to swiftly detect, monitor and respond to emerging viral threats. Strengthening surveillance infrastructure and promoting international collaboration are essential to mitigate future pandemics (Faculty of Pharmacy Kampala International University Uganda and David, 2024). In summary, South Africa's genomic surveillance of SARS-CoV-2 stands as a testament to the tenacity of human spirit and the inexorable march of scientific progress. The nation's unwavering commitment, collaborative ethos and pioneering efforts have redefined the boundaries of what is achievable in the realm of viral genomics. The impact of this research transcends the immediate crisis, ushering in an era where proactive surveillance and rapid response mechanisms hold the key to averting future pandemics. As the world takes stock of the monumental strides made during the pandemic, South Africa's legacy will forever shine as a beacon of hope and inspiration, reminding us that in the face of adversity, scientific inquiry can illuminate the path to a safer and healthier future.

5.2 Conclusions

South Africa’s response to the SARS-CoV-2 pandemic exemplifies the country’s remarkable advancements in genomic surveillance, demonstrating a steadfast commitment to transparency, scientific excellence, and global collaboration. Through its robust and timely data-sharing strategies, South Africa has played a crucial role in enhancing the global understanding of the virus, tracking its evolution, and facilitating swift responses to emerging variants (Mahomed and Staunton, 2021; Tegally *et al.*, 2021a; Tegally *et al.*, 2021b). This proactive approach has not only strengthened global public

health efforts but also set a precedent for how nations can leverage genomic surveillance to mitigate the impact of infectious disease outbreaks. Despite facing socio-economic challenges, South Africa has navigated the complexities of public health crises with innovation and resilience. The nation's ability to optimise limited resources for high-impact genomic surveillance serves as an inspiring model for other countries with constrained healthcare infrastructure. Its experience highlights the indispensable role of equitable access to genomic technologies, reinforcing the need for stronger international partnerships and resource-sharing frameworks. The country's leadership in genomic data transparency has fostered a spirit of global cooperation, emphasising that collective action is vital in managing current and future pandemics. Looking ahead, building long-term resilience within the healthcare system must be a priority. This entails not only strengthening existing infrastructure but also fostering continuous innovation in public health practices. Ensuring equitable access to healthcare, regardless of socio-economic status, is paramount in creating robust and inclusive health systems. The lessons learned from South Africa's response to the pandemic provide invaluable insights for shaping future public health policies, offering a roadmap for regions worldwide to develop more resilient and adaptive healthcare frameworks. The findings of this thesis reinforce the pivotal role of genomic surveillance in infectious disease management. By evaluating global trends, comparing surveillance strategies, and identifying key challenges and opportunities, this research underscores the necessity of sustained investment, cross-border collaboration, and the integration of sequencing data into public health decision-making. These insights contribute to strengthening pandemic preparedness, ensuring timely responses to emerging pathogens, and enhancing global health security. Ultimately, this study advocates for a more equitable, cooperative, and science-driven approach to public health, ensuring that the world is better equipped to confront future health crises with resilience and efficiency.

5.3 Future recommendations and limitations

While this study provides valuable insights into SARS-CoV-2 genomic surveillance in South Africa and its global implications, several limitations should be acknowledged. First, the study relies on publicly available genomic data from GISAID and case data from global repositories, which may be subject to reporting biases, inconsistencies, and delays (Khare *et al.*, 2021; *GISAID - Initiative*, 2024). Variations in sequencing capacity across different regions within South Africa and other countries may have influenced comparative analyses. Second, while South Africa has made significant progress in genomic surveillance, limitations in sequencing infrastructure, funding availability, and skilled workforce distribution may have affected the completeness and timeliness of data reporting (Adebamowo *et al.*, 2018; Hill *et al.*, 2021). The study does not fully account for the operational and logistical challenges that may have influenced genome-sharing trends. Lastly, while the study provides a strong framework for genomic surveillance in the context of SARS-CoV-2, its applicability to future

pandemics remains uncertain. The evolution of pathogens, advancements in sequencing technologies, and changing public health priorities may require adaptations to the strategies outlined in this research.

To overcome these limitations and strengthen future genomic surveillance efforts, several targeted and innovative recommendations should be explored to improve effectiveness and pandemic preparedness. First, strengthening genomic surveillance infrastructure requires increased investment in high-throughput sequencing technologies, ensuring that resource-limited regions have access to next-generation sequencing (NGS) platforms for real-time pathogen monitoring (Pronyk *et al.*, 2023; Gerilovych *et al.*, 2024). Establishing decentralised sequencing laboratories outside major urban centres would improve data collection efficiency and reduce delays, while targeted workforce development programs in bioinformatics and genomic epidemiology would cultivate expertise in genomic surveillance (Abrudan *et al.*, 2021; de Oliveira *et al.*, 2022). To sustain these advancements, governments and funding agencies should implement long-term financial models to ensure continued investment beyond pandemic periods, preventing infrastructure degradation due to fluctuating resources. International collaboration and data sharing must also be improved through standardised and enforceable data-sharing policies that balance transparency with ethical concerns. Establishing regional genomic surveillance hubs, particularly in LMICs, could facilitate data exchange and real-time pathogen tracking. High-income nations with established genomic capabilities should engage in capacity-building partnerships, providing LMICs with access to technology, training, and infrastructure to develop independent genomic surveillance systems. Additionally, incentivising data sharing through grants, technical support, and scientific recognition would encourage more institutions and researchers to contribute genomic data to global repositories. Genomic surveillance should also be seamlessly integrated into public health decision-making to maximise its impact. Governments must implement automated bioinformatics pipelines for real-time genomic analysis, allowing for early detection of emerging variants and guiding immediate public health responses. Linking genomic surveillance with epidemiological databases would provide a more comprehensive understanding of pathogen transmission dynamics, informing targeted interventions. Moreover, genomic insights should be leveraged for vaccine development by predicting viral evolution, ensuring the timely formulation of updated vaccines. Public health agencies should also establish advisory panels composed of genomic scientists, epidemiologists, and policymakers to ensure that genomic surveillance findings directly influence policy decisions (Moodley *et al.*, 2022). To address socioeconomic barriers to genomic surveillance, cost-effective solutions should be prioritised. Research and development into portable sequencing technologies, such as nanopore sequencing, would enable rapid, field-based genomic analysis at a lower cost, reducing reliance on centralised laboratories. Public-private collaborations should be strengthened to facilitate affordable access to sequencing reagents, equipment, and analytical software for LMICs (Belman, Saha and Beale, 2022; Gomes, Rebelo and Alves De Sousa, 2022; Espinoza *et al.*, 2023). Furthermore, public health literacy on genomic surveillance must be improved

through targeted awareness campaigns and educational programs to ensure policymakers, healthcare professionals, and the public understand its significance, fostering long-term support and sustainability (de Oliveira, 2022). Future research must extend beyond SARS-CoV-2 to focus on genomic surveillance of other high-risk pathogens, particularly zoonotic diseases. Integrating human and animal genomic surveillance under a One Health framework would allow for early detection of potential spill-over events, mitigating future outbreaks (One Health High-Level Expert Panel (OHHLEP) *et al.*, 2022; Hayman *et al.*, 2023). Novel research should also explore the use of artificial intelligence and machine learning in genomic epidemiology to improve predictive modelling for pathogen evolution. Additionally, investigating how climate change influences the emergence and spread of infectious diseases using genomic data could provide valuable insights into long-term global health risks. Addressing antimicrobial resistance (AMR) through genomic surveillance is another critical avenue, enabling the tracking of resistance patterns and informing targeted antimicrobial stewardship programs (Abrudan *et al.*, 2021).

By advocating for these recommendations, future research may help establish a more resilient, globally integrated, and equitable genomic surveillance system, strengthening pandemic preparedness. These innovative approaches will strengthen real-time pathogen tracking while fostering proactive, data-driven public health strategies to mitigate future health crises effectively. Building on the lessons of the COVID-19 pandemic—guided by principles of equity, innovation, and collaboration—will ensure sustained preparedness and prevent the loss of invaluable knowledge and experience gained during this unprecedented time. In alignment with this study's rationale, I conclude with a powerful statement:

> "By empowering African scientists and building a strong genomic infrastructure, I can collectively pave the way to a healthier future for all"
>
> – Prof Tulio de Oliveira and Dr Cheryl Baxter
> (de Oliveira and Baxter, 2024)

References

Abbasian, M.H. *et al.* (2023) "Global landscape of SARS-CoV-2 mutations and conserved regions.," *Journal of translational medicine*, 21(1), p. 152. Available at: https://doi.org/10.1186/s12967-023-03996-w.

Abrudan, M. *et al.* (2021) "Train-the-Trainer as an Effective Approach to Building Global Networks of Experts in Genomic Surveillance of AMR." Scientific Communication and Education. Available at: https://doi.org/10.1101/2021.06.18.448940.

Adebamowo, S.N. *et al.* (2018) "Implementation of genomics research in Africa: challenges and recommendations," *Global Health Action*, 11(1), p. 1419033. Available at: https://doi.org/10.1080/16549716.2017.1419033.

Aggarwal, D. *et al.* (2022) "The role of viral genomics in understanding COVID-19 outbreaks in long-term care facilities," *The Lancet Microbe*, 3(2), pp. e151–e158. Available at: https://doi.org/10.1016/S2666-5247(21)00208-1.

Ahmad, A., Fawaz, M.A.M. and Aisha, A. (2022) "A comparative overview of SARS-CoV-2 and its variants of concern," *Le Infezioni in Medicina*, 30(3), pp. 328–343. Available at: https://doi.org/10.53854/liim-3003-2.

Ahmed, W. *et al.* (2022) "Detection of the Omicron (B.1.1.529) variant of SARS-CoV-2 in aircraft wastewater," *Science of The Total Environment*, 820, p. 153171. Available at: https://doi.org/10.1016/j.scitotenv.2022.153171.

Akande, O.W. *et al.* (2023) "Strengthening pathogen genomic surveillance for health emergencies: insights from the World Health Organization's regional initiatives," *Frontiers in Public Health*, 11, p. 1146730. Available at: https://doi.org/10.3389/fpubh.2023.1146730.

Akkiz, H. (2021) "Implications of the Novel Mutations in the SARS-CoV-2 Genome for Transmission, Disease Severity, and the Vaccine Development," *Frontiers in Medicine*, 8(May), pp. 1–10. Available at: https://doi.org/10.3389/fmed.2021.636532.

Akkız, H. (2022) "The Biological Functions and Clinical Significance of SARS-CoV-2 Variants of Corcern," *Frontiers in Medicine*, 9(May), pp. 1–15. Available at: https://doi.org/10.3389/fmed.2022.849217.

Aksamentov, I. *et al.* (2021) "Nextclade: clade assignment, mutation calling and quality control for viral genomes," *Zenodo* [Preprint]. Available at: https://doi.org/10.5281/zenodo.5607694.

Al Hasan, S.M. *et al.* (2022) "COVID-19 outbreak trends in South Africa: A comparison of Omicron (B.1.1.529), Delta (B.1.617.2), and Beta (B.1.351) variants outbreak periods," *Journal of Infection and Public Health*, 15(7), pp. 726–733. Available at: https://doi.org/10.1016/j.jiph.2022.05.011.

Alessandrini, F. *et al.* (2020) "Research Panel by Massive Parallel Sequencing," *Genes*, 11(December 2019), pp. 1–12.

Alexander, M. (2023) *The nine provinces of South Africa*. Available at: https://southafrica-info.com/ (Accessed: August 22, 2023).

Alkhatib, M. *et al.* (2021) "SARS-CoV-2 Variants and Their Relevant Mutational Profiles: Update Summer 2021," *Microbiology Spectrum*. Edited by D.R. Perez, 9(3), pp. e01096-21. Available at: https://doi.org/10.1128/Spectrum.01096-21.

Amicone, M. *et al.* (2022) "Mutation rate of SARS-CoV-2 and emergence of mutators during experimental evolution," *Evolution, Medicine and Public Health*, 10(1), pp. 142–155. Available at: https://doi.org/10.1093/emph/eoac010.

Andersen, K.G. *et al.* (2020) "The proximal origin of SARS-CoV-2," *Nature Medicine*, 26(4), pp. 450–452. Available at: https://doi.org/10.1038/s41591-020-0820-9.

Arndt, C. *et al.* (2020) "Covid-19 lockdowns, income distribution, and food security: An analysis for South Africa," *Global Food Security*, 26, p. 100410. Available at: https://doi.org/10.1016/j.gfs.2020.100410.

Arora, P. *et al.* (2021) "New Coronavirus (SARS-CoV-2) Crossing Borders Beyond Cities, Nations, and Continents: Impact of International Travel," *Balkan Medical Journal*, 38(4), pp. 205–211. Available at: https://doi.org/10.5152/balkanmedj.2021.21074.

Attar Cohen, H. *et al.* (2023) "Surveillance for variants of SARS-CoV-2 to inform risk assessments," *Bulletin of the World Health Organization*, 101(11), pp. 706–716. Available at: https://doi.org/10.2471/BLT.23.290093.

Atzrodt, C.L. *et al.* (2020) "A Guide to COVID-19: a global pandemic caused by the novel coronavirus SARS-CoV-2," *The FEBS Journal*, 287(17), pp. 3633–3650. Available at: https://doi.org/10.1111/febs.15375.

Ayouni, I. *et al.* (2021) "Effective public health measures to mitigate the spread of COVID-19: a systematic review," *BMC Public Health*, 21(1), pp. 1–14. Available at: https://doi.org/10.1186/s12889-021-11111-1.

Azuma, K. *et al.* (2020) "Environmental factors involved in SARS-CoV-2 transmission: effect and role of indoor environmental quality in the strategy for COVID-19 infection control," *Environmental Health and Preventive Medicine*, 25(1), p. 66. Available at: https://doi.org/10.1186/s12199-020-00904-2.

Babadi, A.A. *et al.* (2022) "SARS-CoV-2 detection by targeting four loci of viral genome using graphene oxide and gold nanoparticle DNA biosensor," *Scientific Reports*, 12(1), pp. 1–12. Available at: https://doi.org/10.1038/s41598-022-23996-y.

Babb de Villiers, C. *et al.* (2021) "Next Generation Sequencing for SARS-CoV-2," *Foundation for Innovative New Diagnostics* [Preprint].

Baker, K.S. *et al.* (2023) "Genomics for public health and international surveillance of antimicrobial resistance," *The Lancet Microbe*, 4(12), pp. e1047–e1055. Available at: https://doi.org/10.1016/S2666-5247(23)00283-5.

Bekker, L.G. and Ntusi, N.A.B. (2021) "Lessons from two SARS-CoV-2 waves in South Africa," *The Lancet Global Health*, 9(9), pp. e1177–e1178. Available at: https://doi.org/10.1016/S2214-109X(21)00313-2.

Belman, S., Saha, S. and Beale, M.A. (2022) "SARS-CoV-2 genomics as a springboard for future disease mitigation in LMICs," *Nature Reviews Microbiology*, 20(1), p. 3. Available at: https://doi.org/10.1038/s41579-021-00664-y.

Biancolella, M. *et al.* (2023) "COVID-19 annual update: a narrative review," *Human Genomics*, 17(1), p. 68. Available at: https://doi.org/10.1186/s40246-023-00515-2.

Billah, M.A., Miah, M.M. and Khan, M.N. (2020) "Reproductive number of coronavirus: A systematic review and meta-analysis based on global level evidence," *PLoS ONE*, 15(11 November), pp. 1–17. Available at: https://doi.org/10.1371/journal.pone.0242128.

Boehm, E. *et al.* (2021) "Novel SARS-CoV-2 variants: the pandemics within the pandemic," *Clinical Microbiology and Infection*, 27(8), pp. 1109–1117. Available at: https://doi.org/10.1016/j.cmi.2021.05.022.

Bosonkie, M. *et al.* (2023) "Improving testing capacity for COVID-19: experiences and lessons from Senegal, Uganda, Nigeria, and the Democratic Republic of Congo," *Frontiers in Public Health*, 11(November), pp. 1–15. Available at: https://doi.org/10.3389/fpubh.2023.1202966.

Boterman, W. (2023) "Population density and SARS-CoV-2 pandemic: Comparing the geography of different waves in the Netherlands," *Urban Studies*, 60(8), pp. 1377–1402. Available at: https://doi.org/10.1177/00420980221087165.

Bragg, L.M. *et al.* (2013) "Shining a Light on Dark Sequencing: Characterising Errors in Ion Torrent PGM Data," *PLoS Computational Biology*, 9(4). Available at: https://doi.org/10.1371/journal.pcbi.1003031.

Braun, M. and Benstoem, C. (2023) "How to never run out of words." Available at: profbenstoem.com.

Brito, A.F. *et al.* (2022) "Global disparities in SARS-CoV-2 genomic surveillance," *Nature Communications*, 13(1), p. 7003. Available at: https://doi.org/10.1038/s41467-022-33713-y.

Buermans, H.P.J. and den Dunnen, J.T. (2014) "Next generation sequencing technology: Advances and applications," *Biochimica et Biophysica Acta - Molecular Basis of Disease*, 1842(10), pp. 1932–1941. Available at: https://doi.org/10.1016/j.bbadis.2014.06.015.

Burns, B.L. *et al.* (2019) "Healthcare System Priorities for Successful Integration of Genomics: An Australian Focus," *Frontiers in Public Health*, 7, p. 41. Available at: https://doi.org/10.3389/fpubh.2019.00041.

Butera, Y. *et al.* (2021) "Genomic sequencing of SARS-CoV-2 in Rwanda reveals the importance of incoming travelers on lineage diversity," *Nature Communications*, 12(1), pp. 1–12. Available at: https://doi.org/10.1038/s41467-021-25985-7.

Carabelli, A.M. *et al.* (2023) "SARS-CoV-2 variant biology: immune escape, transmission and fitness," *Nature Reviews Microbiology*, 21(3), pp. 162–177. Available at: https://doi.org/10.1038/s41579-022-00841-7.

Carter, L. *et al.* (2022) "Global genomic surveillance strategy for pathogens with pandemic and epidemic potential 2022–2032," *Bulletin of the World Health Organization*, 100(04), pp. 239-239A. Available at: https://doi.org/10.2471/BLT.22.288220.

Cele, S. *et al.* (2021) "Escape of SARS-CoV-2 501Y.V2 from neutralization by convalescent plasma," *Nature*, 593(7857), pp. 142–146. Available at: https://doi.org/10.1038/s41586-021-03471-w.

Centers for Disease Control and Prevention (2023) *SARS-CoV-2 Variant Classifications and Definitions*, *COVID-19*. Available at: https://www.cdc.gov/coronavirus/2019-ncov/variants/variant-classifications.html#print (Accessed: June 23, 2023).

Chakraborty, C. *et al.* (2021) "Evolution, Mode of Transmission, and Mutational Landscape of Newly Emerging SARS-CoV-2 Variants," *mBio*. Edited by S.P. Goff, 12(4), pp. e01140-21. Available at: https://doi.org/10.1128/mBio.01140-21.

Challen, R. *et al.* (2021) "Early epidemiological signatures of novel SARS-CoV-2 variants: establishment of B.1.617.2 in England," *medRxiv*, p. 2021.06.05.21258365. Available at: https://www.medrxiv.org/content/10.1101/2021.06.05.21258365v1%0Ahttps://www.medrxiv.org/content/10.1101/2021.06.05.21258365v1.abstract.

Charre, C. *et al.* (2020) "Evaluation of NGS-based approaches for SARS-CoV- 2 whole genome characterisation," *Virus Evolution*, 6(2), pp. 1–8. Available at: https://doi.org/10.1093/ve/veaa075.

Chauhan, A.S. *et al.* (2023) "The Health System's Response to and the Impact of COVID-19 on Health Services, Providers, and Seekers: A Rapid Review in the Wake of the Pandemic," *COVID*, 3(8), pp. 1106–1157. Available at: https://doi.org/10.3390/covid3080081.

Chekol Abebe, E. *et al.* (2022) "Mutational Pattern, Impacts and Potential Preventive Strategies of Omicron SARS-CoV-2 Variant Infection," *Infection and Drug Resistance*, Volume 15(April), pp. 1871–1887. Available at: https://doi.org/10.2147/IDR.S360103.

Chen, Z. *et al.* (2022) "Global landscape of SARS-CoV-2 genomic surveillance and data sharing," *Nature Genetics*, 54(4), pp. 499–507. Available at: https://doi.org/10.1038/s41588-022-01033-y.

Cheng, Y. *et al.* (2023) "Web Resources for SARS-CoV-2 Genomic Database, Annotation, Analysis and Variant Tracking," *Viruses*, 15(5), pp. 1–17. Available at: https://doi.org/10.3390/v15051158.

Cherian, S. *et al.* (2021) "SARS-CoV-2 Spike Mutations, L452R, T478K, E484Q and P681R, in the Second Wave of COVID-19 in Maharashtra, India," *Microorganisms*, 9(7), p. 1542. Available at: https://doi.org/10.3390/microorganisms9071542.

Chirico, F. *et al.* (2022) "Safety & effectiveness of COVID-19 vaccines: A narrative review," *Indian Journal of Medical Research*, 155(1), p. 91. Available at: https://doi.org/10.4103/ijmr.ijmr_474_21.

Choi, J.H. *et al.* (2023) "Global lineage evolution pattern of sars-cov-2 in Africa, America, Europe, and Asia: A comparative analysis of variant clusters and their relevance across continents," *Journal of Translational Internal Medicine*, 11(4), pp. 410–422. Available at: https://doi.org/10.2478/jtim-2023-0118.

Chrysostomou, A.C. *et al.* (2022) "Genomic Epidemiology of the SARS-CoV-2 Epidemic in Cyprus from November 2020 to October 2021: The Passage of Waves of Alpha and Delta Variants of Concern," *Viruses*, 15(1), p. 108. Available at: https://doi.org/10.3390/v15010108.

Cimerman, S. *et al.* (2021) "One year after the arrival of COVID-19 in Latin America: what have we learned in Brazil and other countries?," *The Brazilian Journal of Infectious Diseases*, 25(2), p. 101571. Available at: https://doi.org/10.1016/j.bjid.2021.101571.

Clark, C. *et al.* (2022) "Validation of reduced S-gene target performance and failure for rapid surveillance of SARS-CoV-2 variants," *PLoS ONE*, 17(10), pp. 1–11. Available at: https://doi.org/10.1371/journal.pone.0275150.

Cleemput, S. *et al.* (2020) "Genome Detective Coronavirus Typing Tool for rapid identification and characterization of novel coronavirus genomes," *Bioinformatics*. Edited by P. Luigi Martelli, 36(11), pp. 3552–3555. Available at: https://doi.org/10.1093/bioinformatics/btaa145.

Colton, H. *et al.* (2023) "Factors affecting turnaround time of SARS-CoV-2 sequencing for inpatient infection prevention and control decision making: analysis of data from the COG-UK HOCI study," *Journal of Hospital Infection*, 131, pp. 34–42. Available at: https://doi.org/10.1016/j.jhin.2022.09.022.

Confraria, H. and Wang, L. (2020) "Medical research versus disease burden in Africa," *Research Policy*, 49(3), p. 103916. Available at: https://doi.org/10.1016/j.respol.2019.103916.

Corey, L. *et al.* (2021) "SARS-CoV-2 Variants in Patients with Immunosuppression," *New England Journal of Medicine*, 385(6), pp. 562–566. Available at: https://doi.org/10.1056/NEJMsb2104756.

COVID-19 South African Online Portal (2023). Available at: https://sacoronavirus.co.za/2023 (Accessed: October 1, 2023).

Da Costa, A.P.A. *et al.* (2021) "History and Global Status of the New Coronavirus Covid-2019 and Aspects of Previous Infections of SARS-CoV and MERS-CoV: A Systematic Review," *Coronaviruses*, 2(6), p. e130621190409. Available at: https://doi.org/10.2174/2666796702666210118153843.

Dandara, C. *et al.* (2014) "H3Africa and the African Life Sciences Ecosystem: Building Sustainable Innovation," *OMICS: A Journal of Integrative Biology*, 18(12), pp. 733–739. Available at: https://doi.org/10.1089/omi.2014.0145.

Davies, N.G. *et al.* (2021) "Estimated transmissibility and impact of SARS-CoV-2 lineage B.1.1.7 in England," *Science*, 372(6538). Available at: https://doi.org/10.1126/science.abg3055.

Davis, J.J. *et al.* (2021) "Analysis of the ARTIC version 3 and version 4 SARS-CoV-2 primers and their impact on the detection of the G142D amino acid substitution in the spike protein." Available at: https://doi.org/10.1101/2021.09.27.461949.

De Jong, S.P.J. *et al.* (2023) "Global solidarity in genomic surveillance improves early detection of respiratory virus threats." Infectious Diseases (except HIV/AIDS). Available at: https://doi.org/10.1101/2023.11.01.23297901.

De Wit, E. *et al.* (2016) "SARS and MERS: Recent insights into emerging coronaviruses," *Nature Reviews Microbiology*, 14(8), pp. 523–534. Available at: https://doi.org/10.1038/nrmicro.2016.81.

Dhama, K. *et al.* (2020) "Coronavirus Disease 2019–COVID-19," *Clinical Microbiology Reviews*, 33(4). Available at: https://doi.org/10.1128/CMR.00028-20.

Dhawan, M. *et al.* (2022) "Omicron variant (B.1.1.529) and its sublineages: What do we know so far amid the emergence of recombinant variants of SARS-CoV-2?," *Biomedicine & Pharmacotherapy*, 154, p. 113522. Available at: https://doi.org/10.1016/j.biopha.2022.113522.

Dhungel, B. *et al.* (2022) "Reliability of Early Estimates of the Basic Reproduction Number of COVID-19: A Systematic Review and Meta-Analysis," *International Journal of Environmental Research and Public Health*, 19(18). Available at: https://doi.org/10.3390/ijerph191811613.

Di Paola, N. *et al.* (2020) "Viral genomics in Ebola virus research," *Nature Reviews Microbiology*, 18(7), pp. 365–378. Available at: https://doi.org/10.1038/s41579-020-0354-7.

Domman, D. *et al.* (2022) "112 Genomic surveillance for SARS-CoV-2 for New Mexico and the Mountain West," *Journal of Clinical and Translational Science*, 6(s1), pp. 4–4. Available at: https://doi.org/10.1017/cts.2022.32.

Durizzo, K. *et al.* (2021) "Managing the COVID-19 pandemic in poor urban neighborhoods: The case of Accra and Johannesburg," *World Development*, 137, p. 105175. Available at: https://doi.org/10.1016/j.worlddev.2020.105175.

Eike, D. *et al.* (2022) "How the COVID-19 Pandemic Alters the Landscapes of the HIV and Tuberculosis Epidemics in South Africa: A Case Study and Future Directions," *Epidemiologia*, 3(2), pp. 297–313. Available at: https://doi.org/10.3390/epidemiologia3020023.

Ekusai-Sebatta, D. *et al.* (2023) "Data sharing practices in collaborative human genomic research in low- and middle-income countries: A systematic review protocol," *PLOS ONE*. Edited by V.E. Gibbon, 18(11), p. e0292996. Available at: https://doi.org/10.1371/journal.pone.0292996.

Elbe, S. and Buckland-Merrett, G. (2017) "Data, disease and diplomacy: GISAID's innovative contribution to global health," *Global Challenges*, 1(1), pp. 33–46. Available at: https://doi.org/10.1002/gch2.1018.

Engelbrecht, S. *et al.* (2021) "Multiple early introductions of sars-cov-2 to cape town, South Africa," *Viruses*, 13(3), pp. 1–11. Available at: https://doi.org/10.3390/v13030526.

Esmaeilzadeh, A. *et al.* (2024) "EG.5 (Eris) and BA.2.86 (Pirola) two new subvariants of SARS-CoV-2: a new face of old COVID-19," *Infection*, 52(2), pp. 337–343. Available at: https://doi.org/10.1007/s15010-023-02146-0.

Espinoza, B. *et al.* (2023) "Coupled models of genomic surveillance and evolving pandemics with applications for timely public health interventions," *Proceedings of the National Academy of Sciences*, 120(48), p. e2305227120. Available at: https://doi.org/10.1073/pnas.2305227120.

Evans, N. *et al.* (2021) "SARS-CoV-2 and human immunodeficiency virus: Pathogen pincer attack," *HIV/AIDS - Research and Palliative Care*, 13, pp. 361–375. Available at: https://doi.org/10.2147/HIV.S300055.

Faculty of Pharmacy Kampala International University Uganda and David, B.G. (2024) "Global Health Security: Preparing For Future Pandemics," *NEWPORT INTERNATIONAL JOURNAL OF RESEARCH IN MEDICAL SCIENCES*, 5(3), pp. 42–45. Available at: https://doi.org/10.59298/NIJRMS/2024/5.3.4245.

Fadaka, A.O. *et al.* (2020) "Understanding the epidemiology, pathophysiology, diagnosis and management of SARS-CoV-2," *Journal of International Medical Research*, 48(8). Available at: https://doi.org/10.1177/0300060520949077.

Faria, N.R. *et al.* (2016) "Zika virus in the Americas: Early epidemiological and genetic findings," *Science (New York, N.Y.)*, 352(6283), p. 345. Available at: https://doi.org/10.1126/SCIENCE.AAF5036.

Faria, N.R. *et al.* (2021) "Genomics and epidemiology of the P.1 SARS-CoV-2 lineage in Manaus, Brazil," *Science*, 372(6544), pp. 815–821. Available at: https://doi.org/10.1126/science.abh2644.

Feng, Y. *et al.* (2024) "Trends and impacts of SARS-CoV-2 genome sharing: a comparative analysis of China and the global community, 2020–2023," *Frontiers in Public Health*, 12, p. 1491623. Available at: https://doi.org/10.3389/fpubh.2024.1491623.

Foo, C.D. *et al.* (2022) "COVID-19 public health and social measures: a comprehensive picture of six Asian countries," *BMJ Global Health*, 7(11). Available at: https://doi.org/10.1136/bmjgh-2022-009863.

Forero, D.A. *et al.* (2016) "Current needs for human and medical genomics research infrastructure in low and middle income countries: Table 1," *Journal of Medical Genetics*, 53(7), pp. 438–440. Available at: https://doi.org/10.1136/jmedgenet-2015-103631.

Franceschi, V.B. *et al.* (2021) "Predominance of the SARS-CoV-2 Lineage P.1 and Its Sublineage P.1.2 in Patients from the Metropolitan Region of Porto Alegre, Southern Brazil in March 2021," *Pathogens*, 10(8), p. 988. Available at: https://doi.org/10.3390/pathogens10080988.

Fu, W., Chou, S.-Y. and Wang, L.-S. (2022) *NIH Grant Expansion, Ancestral Diversity and Scientific Discovery in Genomics Research*. w30155. Cambridge, MA: National Bureau of Economic Research, p. w30155. Available at: https://doi.org/10.3386/w30155.

Gao, W. *et al.* (2021) "Role of asymptomatic and pre-symptomatic infections in covid-19 pandemic," *BMJ*, p. n2342. Available at: https://doi.org/10.1136/bmj.n2342.

Gardy, J.L. and Loman, N.J. (2018) "Towards a genomics-informed, real-time, global pathogen surveillance system," *Nature Reviews Genetics*, 19(1), pp. 9–20. Available at: https://doi.org/10.1038/nrg.2017.88.

Genome Detective Virus Tool (2019). Available at: https://www.genomedetective.com/app/typingtool/virus/ (Accessed: December 7, 2021).

Gerilovych, A.P. *et al.* (2024) "APPLICATION OF THE NEXT GENERATION SEQUENCING IN BIOLOGY AND MEDICINE," *One Health Journal*, 2(I), pp. 32–44. Available at: https://doi.org/10.31073/onehealthjournal2024-I-05.

Giandhari, J. *et al.* (2021) "Early transmission of SARS-CoV-2 in South Africa: An epidemiological and phylogenetic report," *International Journal of Infectious Diseases*, 103, pp. 234–241. Available at: https://doi.org/10.1016/j.ijid.2020.11.128.

Gim, J.-A. (2022) "A Genomic Information Management System for Maintaining Healthy Genomic States and Application of Genomic Big Data in Clinical Research," *International Journal of Molecular Sciences*, 23(11), p. 5963. Available at: https://doi.org/10.3390/ijms23115963.

Giovanetti, M. *et al.* (2022) "Genomic epidemiology of the SARS-CoV-2 epidemic in Brazil," *Nature Microbiology*, 7(9), pp. 1490–1500. Available at: https://doi.org/10.1038/s41564-022-01191-z.

GISAID - Initiative (2024). Available at: https://www.gisaid.org/ (Accessed: May 8, 2023).

Gomes, B.M., Rebelo, C.B. and Alves De Sousa, L. (2022) "Public health, surveillance systems and preventive medicine in an interconnected world," in *One Health*. Elsevier, pp. 33–71. Available at: https://doi.org/10.1016/B978-0-12-822794-7.00006-X.

Gorbalenya, A.E. *et al.* (2020) "The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2," *Nature Microbiology*, 5(4), pp. 536–544. Available at: https://doi.org/10.1038/s41564-020-0695-z.

Govender, I. (2024) "Brain drain in South Africa is affecting health care," *South African Family Practice*, 66(1). Available at: https://doi.org/10.4102/safp.v66i1.5830.

Gozashti, L. and Corbett-Detig, R. (2021) "Shortcomings of SARS-CoV-2 genomic metadata," *BMC Research Notes*, 14(1), p. 189. Available at: https://doi.org/10.1186/s13104-021-05605-9.

Hadfield, J. *et al.* (2018) "Nextstrain: real-time tracking of pathogen evolution," *Bioinformatics*. Edited by J. Kelso, 34(23), pp. 4121–4123. Available at: https://doi.org/10.1093/bioinformatics/bty407.

Haider, N. *et al.* (2020) "Lockdown measures in response to COVID-19 in nine sub-Saharan African countries," *BMJ Global Health*, 5(10), pp. 1–10. Available at: https://doi.org/10.1136/bmjgh-2020-003319.

Hamidi, S., Sabouri, S. and Ewing, R. (2020) "Does Density Aggravate the COVID-19 Pandemic?: Early Findings and Lessons for Planners," *Journal of the American Planning Association*, 86(4), pp. 495–509. Available at: https://doi.org/10.1080/01944363.2020.1777891.

Hamzaoui, Z. *et al.* (2024) "Genomic surveillance of SARS-CoV-2 in North Africa: 4 years of GISAID data sharing," *IJID Regions*, 11, p. 100356. Available at: https://doi.org/10.1016/j.ijregi.2024.100356.

Harris, J.E. (2020) "Data from the COVID-19 epidemic in Florida suggest that younger cohorts have been transmitting their infections to less socially mobile older adults," *Review of Economics of the Household*, 18(4), pp. 1019–1037. Available at: https://doi.org/10.1007/s11150-020-09496-w.

Harrison, A.G., Lin, T. and Wang, P. (2020) "Mechanisms of SARS-CoV-2 Transmission and Pathogenesis," *Trends in Immunology*, 41(12), pp. 1100–1115. Available at: https://doi.org/10.1016/j.it.2020.10.004.

Hartenian, E. *et al.* (2020) "The molecular virology of coronaviruses," *Journal of Biological Chemistry*, 295(37), pp. 12910–12934. Available at: https://doi.org/10.1074/jbc.REV120.013930.

Harvey, W.T. *et al.* (2021) "SARS-CoV-2 variants, spike mutations and immune escape," *Nature Reviews Microbiology*, 19(7), pp. 409–424. Available at: https://doi.org/10.1038/s41579-021-00573-0.

Hayman, D.T.S. *et al.* (2023) "Developing One Health surveillance systems," *One Health*, 17, p. 100617. Available at: https://doi.org/10.1016/j.onehlt.2023.100617.

Hill, V. *et al.* (2021) "Progress and challenges in virus genomic epidemiology," *Trends in Parasitology*, 37(12), pp. 1038–1049. Available at: https://doi.org/10.1016/j.pt.2021.08.007.

Hill, V. *et al.* (2023) "Toward a global virus genomic surveillance network," *Cell Host & Microbe*, 31(6), pp. 861–873. Available at: https://doi.org/10.1016/j.chom.2023.03.003.

Hoang, T. *et al.* (2022) "AusTrakka: Fast-tracking nationalized genomics surveillance in response to the COVID-19 pandemic," *Nature Communications*, 13(1), p. 865. Available at: https://doi.org/10.1038/s41467-022-28529-9.

Hoffmann, M. *et al.* (2020) "SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor," *Cell*, 181(2), pp. 271-280.e8. Available at: https://doi.org/10.1016/j.cell.2020.02.052.

Horovitz, D.D.G., Félix, T.M. and Ferraz, V.E.D.F. (2024) "Medical Genetics in Brazil in the 21st Century: A Thriving Specialty and Its Incorporation in Public Health Policies," *Genes*, 15(8), p. 973. Available at: https://doi.org/10.3390/genes15080973.

Hossain, M., Abdulla, F. and Rahman, A. (2022) "Challenges and difficulties faced in low- and middle-income countries during COVID-19," *Health Policy OPEN*, 3, p. 100082. Available at: https://doi.org/10.1016/j.hpopen.2022.100082.

Hossain, M.S. *et al.* (2021) "Genome-wide identification and prediction of SARS-CoV-2 mutations show an abundance of variants: Integrated study of bioinformatics and deep neural learning," *Informatics in Medicine Unlocked*, 27, p. 100798. Available at: https://doi.org/10.1016/j.imu.2021.100798.

Hu, B. *et al.* (2021) "Characteristics of SARS-CoV-2 and COVID-19," *Nature Reviews Microbiology*, 19(3), pp. 141–154. Available at: https://doi.org/10.1038/s41579-020-00459-7.

Huang, S.-W. and Wang, S.-F. (2021) "SARS-CoV-2 Entry Related Viral and Host Genetic Variations: Implications on COVID-19 Severity, Immune Escape, and Infectivity," *International Journal of Molecular Sciences*, 22(6), p. 3060. Available at: https://doi.org/10.3390/ijms22063060.

Hughes, L. *et al.* (2022) "Outbreak.info genomic reports: scalable and dynamic surveillance of SARS-CoV-2 variants and mutations." In Review. Available at: https://doi.org/10.21203/rs.3.rs-1723829/v1.

Hurd, P.J. and Nelson, C.J. (2009) "Advantages of next-generation sequencing versus the microarray in epigenetic research," *Briefings in Functional Genomics and Proteomics*, 8(3), pp. 174–183. Available at: https://doi.org/10.1093/bfgp/elp013.

Inzaule, S.C. *et al.* (2021) "Genomic-informed pathogen surveillance in Africa: opportunities and challenges," *The Lancet. Infectious Diseases*, 21(9), pp. e281–e289. Available at: https://doi.org/10.1016/S1473-3099(20)30939-7.

Ion GeneStudio™ S5 System (no date). Available at: https://www.thermofisher.com/order/catalog/product/A38194 (Accessed: May 5, 2022).

Jackson, B. *et al.* (2021) "Generation and transmission of interlineage recombinants in the SARS-CoV-2 pandemic," *Cell*, 184(20), pp. 5179-5188.e8. Available at: https://doi.org/10.1016/j.cell.2021.08.014.

Jacobs, E.D. and Okeke, M.I. (2022) "A critical evaluation of Nigeria's response to the first wave of COVID-19," *Bulletin of the National Research Centre*, 46(1), p. 44. Available at: https://doi.org/10.1186/s42269-022-00729-9.

Jacot, D. *et al.* (2021) "Assessment of SARS-CoV-2 Genome Sequencing: Quality Criteria and Low-Frequency Variants," *Journal of Clinical Microbiology*. Edited by A. Mellmann, 59(10), pp. 1–10. Available at: https://doi.org/10.1128/JCM.00944-21.

Jamin, C. *et al.* (2021) "Harmonization of whole-genome sequencing for outbreak surveillance of Enterobacteriaceae and Enterococci," *Microbial Genomics*, 7(7). Available at: https://doi.org/10.1099/mgen.0.000567.

Jassat, W. *et al.* (2022) "Clinical severity of COVID-19 in patients admitted to hospital during the omicron wave in South Africa: a retrospective observational study," *The Lancet Global Health*, 10(7), pp. e961–e969. Available at: https://doi.org/10.1016/S2214-109X(22)00114-0.

Kakavandi, S. *et al.* (2023) "Structural and non-structural proteins in SARS-CoV-2: potential aspects to COVID-19 treatment or prevention of progression of related diseases," *Cell Communication and Signaling*, 21(1), p. 110. Available at: https://doi.org/10.1186/s12964-023-01104-5.

Karim, F. *et al.* (2020) "HIV infection alters SARS-CoV-2 responsive immune parameters but not clinical outcomes in COVID-19 disease," *medRxiv*, 4001, p. 2020.11.23.20236828. Available at: https://doi.org/10.1101/2020.11.23.20236828.

Kaushal, N. *et al.* (2020) "Mutational Frequencies of SARS-CoV-2 Genome during the Beginning Months of the Outbreak in USA," *Pathogens*, 9(7), p. 565. Available at: https://doi.org/10.3390/pathogens9070565.

Kemp, S.A. *et al.* (2021) "SARS-CoV-2 evolution during treatment of chronic infection," *Nature*, 592(7853), pp. 277–282. Available at: https://doi.org/10.1038/s41586-021-03291-y.

Khan, K. *et al.* (2022) "Omicron infection enhances Delta antibody immunity in vaccinated persons," *Nature* [Preprint]. Available at: https://doi.org/10.1038/s41586-022-04830-x.

Khare, S. *et al.* (2021) "GISAID's Role in Pandemic Response," *China CDC Weekly*, 3(49), pp. 1049–1051. Available at: https://doi.org/10.46234/ccdcw2021.255.

Kharsany, A.B.M. *et al.* (2019) "Trends in HIV Prevention, Treatment, and Incidence in a Hyperendemic Area of KwaZulu-Natal, South Africa," *JAMA Network Open*, 2(11), p. e1914378. Available at: https://doi.org/10.1001/jamanetworkopen.2019.14378.

Khemchandani, A. (2022) "World of Genomics: Kenya," *Front Line Genomics*, 8 November. Available at: https://frontlinegenomics.com/world-of-genomics-kenya/ (Accessed: October 9, 2024).

Klein, S. *et al.* (2020) "SARS-CoV-2 structure and replication characterized by in situ cryo-electron tomography," *Nature Communications*, 11(1), p. 5885. Available at: https://doi.org/10.1038/s41467-020-19619-7.

Kluge, H. *et al.* (2018) "Strengthening global health security by embedding the International Health Regulations requirements into national health systems," *BMJ global health*, 3(Suppl 1), p. e000656. Available at: https://doi.org/10.1136/bmjgh-2017-000656.

Knoppers, B.M., Abdul-Rahman, M.H. and Bédard, K. (2007) "Genomic Databases and International Collaboration," *King's Law Journal*, 18(2), pp. 291–311. Available at: https://doi.org/10.1080/09615768.2007.11427678.

Konings, F. *et al.* (2021) "SARS-CoV-2 Variants of Interest and Concern naming scheme conducive for global discourse," *Nature Microbiology*, 6(July), pp. 821–823. Available at: https://doi.org/10.1038/s41564-021-00932-w.

Konono, K.C.C. *et al.* (2024) "Technological Barriers to Routine Genomic Surveillance for Vaccine Development Against SARS-CoV-2 in Africa: A Systematic Review," *Influenza and Other Respiratory Viruses*, 18(11), p. e70047. Available at: https://doi.org/10.1111/irv.70047.

Kotze, M.J., Mashamba-Thompson, T.P. and Stephens, D. (2023) "Editorial: Implementation of genomic medicine in Africa: One continent, one vision," *Frontiers in Genetics*, 13, p. 1133118. Available at: https://doi.org/10.3389/fgene.2022.1133118.

Koyama, T., Platt, D. and Parida, L. (2020) "Variant analysis of SARS-CoV-2 genomes," *Bulletin of the World Health Organization*, 98(7), pp. 495–504. Available at: https://doi.org/10.2471/BLT.20.253591.

Kramer, B. and Libhaber, E. (2018) "Closing the barrier between disease and health outcomes in Africa through research and capacity development," *Global Health Action*, 11(1), p. 1425597. Available at: https://doi.org/10.1080/16549716.2018.1425597.

Kuchinski, K.S. *et al.* (2022) "Mutations in emerging variant of concern lineages disrupt genomic sequencing of SARS-CoV-2 clinical specimens," *International Journal of Infectious Diseases*, 114, pp. 51–54. Available at: https://doi.org/10.1016/j.ijid.2021.10.050.

Laehnemann, D., Borkhardt, A. and McHardy, A.C. (2016) "Denoising DNA deep sequencing data-high-throughput sequencing errors and their correction," *Briefings in Bioinformatics*, 17(1), pp. 154–179. Available at: https://doi.org/10.1093/bib/bbv029.

Lahens, N.F. *et al.* (2017) "A comparison of Illumina and Ion Torrent sequencing platforms in the context of differential gene expression," *BMC genomics*, 18(1), p. 602. Available at: https://doi.org/10.1186/s12864-017-4011-0.

Lambisia, A. *et al.* (2022) "Temporal distribution and clinical characteristics of the Alpha, Delta and Omicron SARS-CoV-2 variants of concern in Laikipia, Kenya: institutional and community-based genomic surveillance," *Wellcome Open Research*, 7(November), p. 235. Available at: https://doi.org/10.12688/wellcomeopenres.18306.1.

Lambisia, A.W. *et al.* (2022) "Optimization of the SARS-CoV-2 ARTIC Network V4 Primers and Whole Genome Sequencing Protocol," *Frontiers in Medicine*, 9(February), pp. 1–8. Available at: https://doi.org/10.3389/fmed.2022.836728.

Lambrou, A.S. *et al.* (2022) "Genomic Surveillance for SARS-CoV-2 Variants: Predominance of the Delta (B.1.617.2) and Omicron (B.1.1.529) Variants — United States, June 2021–January 2022," *MMWR. Morbidity and Mortality Weekly Report*, 71(6), pp. 206–211. Available at: https://doi.org/10.15585/mmwr.mm7106a4.

Lessells, R., Moosa, Y. and de Oliveira, T. (2020) *Report into a nosocomial outbreak of coronavirus disease 2019 (COVID-19) at Netcare St. Augustine's Hospital*. Durban, South Africa. Available at: https://doi.org/10.1126/science.abc9593.

Lewis, H.C. *et al.* (2022) "SARS-CoV-2 infection in Africa: a systematic review and meta-analysis of standardised seroprevalence studies, from January 2020 to December 2021," *BMJ Global Health*, 7(8), p. e008793. Available at: https://doi.org/10.1136/bmjgh-2022-008793.

Li, Q. *et al.* (2020) "Early transmission dynamics in Wuhan, China, of novel coronavirus-infected pneumonia," *New England Journal of Medicine*, 382(13), pp. 1199–1207. Available at: https://doi.org/10.1056/NEJMoa2001316.

Li, X. *et al.* (2021) "Genomic Feature Analysis of Betacoronavirus Provides Insights Into SARS and COVID-19 Pandemics," *Frontiers in Microbiology*, 12, p. 614494. Available at: https://doi.org/10.3389/fmicb.2021.614494.

Ling-Hu, T. *et al.* (2022) "Challenges and Opportunities for Global Genomic Surveillance Strategies in the COVID-19 Era," *Viruses*, 14(11), pp. 1–21. Available at: https://doi.org/10.3390/v14112532.

Liu, J. *et al.* (2021) "Characterization of SARS-CoV-2 worldwide transmission based on evolutionary dynamics and specific viral mutations in the spike protein," *Infectious Diseases of Poverty*, 10(1), pp. 1–15. Available at: https://doi.org/10.1186/s40249-021-00895-4.

Liu, L. *et al.* (2016) "Comparison of nextgeneration sequencing systems," *Advances in Biofuel Production: Algae and Aquatic Plants*, (February), pp. 279–303. Available at: https://doi.org/10.1201/b16568-3.

Liu, T. *et al.* (2021) "A benchmarking study of SARS-CoV-2 whole-genome sequencing protocols using COVID-19 patient samples," *iScience*, 24(8), p. 102892. Available at: https://doi.org/10.1016/j.isci.2021.102892.

Loman, N.J. *et al.* (2012) "Performance comparison of benchtop high-throughput sequencing platforms," *Nature Biotechnology*, 30(5), pp. 434–439. Available at: https://doi.org/10.1038/nbt.2198.

Lu, R. *et al.* (2020) "Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding," *The Lancet*, 395(10224), pp. 565–574. Available at: https://doi.org/10.1016/S0140-6736(20)30251-8.

Ma, L. *et al.* (2023) "On the collection and integration of SARS-CoV-2 genome data," *Biosafety and Health*, 5(4), pp. 204–210. Available at: https://doi.org/10.1016/j.bsheal.2023.07.004.

Madhi, S.A. *et al.* (2022) "Population Immunity and Covid-19 Severity with Omicron Variant in South Africa," *New England Journal of Medicine*, 386(14), pp. 1314–1326. Available at: https://doi.org/10.1056/NEJMoa2119658.

Mahomed, S. and Staunton, C. (2021) "Ethico-legal analysis of international sample and data sharing for genomic research during COVID-19: A South African perspective," *BioLaw Journal - Rivista di BioDiritto*, pp. 261-276 Paginazione. Available at: https://doi.org/10.15168/2284-4503-785.

Maimela, G. *et al.* (2024) "Household transmission of SARS-CoV-2 in a rural area in South Africa," *South African Medical Journal*, 114(2), p. e1159. Available at: https://doi.org/10.7196/SAMJ.2024.v114i2.1159.

Maison, D. *et al.* (2021) "The challenges arising from the COVID-19 pandemic and the way people deal with them. A qualitative longitudinal study," *PLOS ONE*. Edited by S.M. Atiqul Haq, 16(10), p. e0258133. Available at: https://doi.org/10.1371/journal.pone.0258133.

Majam, M. *et al.* (2021) "International citizen project to assess early stage adherence to public health measures for COVID-19 in South Africa," *PLOS ONE*. Edited by P.A. Newman, 16(3), p. e0248055. Available at: https://doi.org/10.1371/journal.pone.0248055.

Malone, B. *et al.* (2022) "Structures and functions of coronavirus replication–transcription complexes and their relevance for SARS-CoV-2 drug design," *Nature Reviews Molecular Cell Biology*, 23(1), pp. 21–39. Available at: https://doi.org/10.1038/s41580-021-00432-z.

Marine, R.L. *et al.* (2020) "Comparison of Illumina MiSeq and the Ion Torrent PGM and S5 platforms for whole-genome sequencing of picornaviruses and caliciviruses," *Journal of Virological Methods*, 280(February), p. 113865. Available at: https://doi.org/10.1016/j.jviromet.2020.113865.

Marjanovic, S. *et al.* (2022) "COVID-19 Genomics UK (COG-UK) Consortium: Final Report."

Markov, P.V. *et al.* (2023) "The evolution of SARS-CoV-2," *Nature Reviews Microbiology*, 21(6), pp. 361–379. Available at: https://doi.org/10.1038/s41579-023-00878-2.

Marta, R.A. *et al.* (2022) "COVID-19 vaccines: Update of the vaccines in use and under development," *Vacunas*, 23, pp. S88–S102. Available at: https://doi.org/10.1016/j.vacun.2022.06.003.

Martin, R. *et al.* (2021) "Genetic conservation of SARS-CoV-2 RNA replication complex in globally circulating isolates and recently emerged variants from humans and minks suggests minimal pre-existing resistance to remdesivir," *Antiviral Research*, 188(January), p. 105033. Available at: https://doi.org/10.1016/j.antiviral.2021.105033.

Mathieu, E. *et al.* (2021) "A global database of COVID-19 vaccinations," *Nature Human Behaviour*, 5(7), pp. 947–953. Available at: https://doi.org/10.1038/s41562-021-01122-8.

Mathieu, E. *et al.* (2024) "Coronavirus (COVID-19) Vaccinations," *OurWorldinData.org*. [Preprint]. Available at: https://ourworldindata.org/covid-vaccinations (Accessed: November 26, 2024).

Mboowa, G. and Sserwadda, I. (2019) "Role of genomics literacy in reducing the burden of common genetic diseases in Africa," *Molecular Genetics & Genomic Medicine*, 7(7), p. e00776. Available at: https://doi.org/10.1002/mgg3.776.

Mboowa, G., Sserwadda, I. and Aruhomukama, D. (2021) "Genomics and bioinformatics capacity in Africa: no continent is left behind," *Genome*, 64(5), pp. 503–513. Available at: https://doi.org/10.1139/gen-2020-0013.

McBroome, J. *et al.* (2023) "Automated Agnostic Designation of Pathogen Lineages." Evolutionary Biology. Available at: https://doi.org/10.1101/2023.02.03.527052.

Meckawy, R. *et al.* (2022) "Effectiveness of early warning systems in the detection of infectious diseases outbreaks: a systematic review," *BMC Public Health*, 22(1), p. 2216. Available at: https://doi.org/10.1186/s12889-022-14625-4.

Mei, S. *et al.* (2015) "Simulating city-level airborne infectious diseases," *Computers, Environment and Urban Systems*, 51, pp. 97–105. Available at: https://doi.org/10.1016/j.compenvurbsys.2014.12.002.

Mercatelli, D. and Giorgi, F.M. (2020) "Geographic and Genomic Distribution of SARS-CoV-2 Mutations," *Frontiers in Microbiology*, 11(July), pp. 1–13. Available at: https://doi.org/10.3389/fmicb.2020.01800.

Merriman, B., Torrent, I. and Rothberg, J.M. (2012) "Progress in Ion Torrent semiconductor chip based sequencing," *Electrophoresis*, 33(23), pp. 3397–3417. Available at: https://doi.org/10.1002/elps.201200424.

Mfuh, K.O., Abanda, N.N. and Titanji, B.K. (2023) "Strengthening diagnostic capacity in Africa as a key pillar of public health and pandemic preparedness," *PLOS global public health*, 3(6), p. e0001998. Available at: https://doi.org/10.1371/journal.pgph.0001998.

Mishchenko, E.L. and Ivanisenko, V.A. (2022) "Replication-transcription complex of coronaviruses: functions of individual viral non-structural subunits, properties and architecture of their complexes," *Vavilov Journal of Genetics and Breeding*, 26(2), pp. 121–127. Available at: https://doi.org/10.18699/VJGB-22-15.

Modisenyane, M. *et al.* (2023) "Test-Trace strategy for disease control and management: South Africa's control measures to contain the spread of COVID-19," *South African Health Review*, 24. Available at: https://doi.org/10.61473/001c.74342.

Mofijur, M. *et al.* (2021) "Impact of COVID-19 on the social, economic, environmental and energy domains: Lessons learnt from a global pandemic," *Sustainable Production and Consumption*, 26, pp. 343–359. Available at: https://doi.org/10.1016/j.spc.2020.10.016.

Moir, M. and Amoako, D.G. (2022) "Continued Emergence and Evolution of Omicron in South Africa : New BA . 4 and," (May). Available at: https://doi.org/10.1101/2022.05.01.22274406.

Moodley, K. *et al.* (2022) "Ethics and governance challenges related to genomic data sharing in southern Africa: the case of SARS-CoV-2.," *The Lancet. Global health*, (22), pp. 1–5. Available at: https://doi.org/10.1016/S2214-109X(22)00417-X.

Moyo, S. *et al.* (2023) "Seroprevalence survey of anti-SARS-CoV-2 antibody and associated factors in South Africa: Findings of the 2020–2021 population-based household survey," *PLOS Global Public Health*. Edited by J. Croda, 3(9), p. e0002358. Available at: https://doi.org/10.1371/journal.pgph.0002358.

Msomi, N. *et al.* (2020) "A genomics network established to respond rapidly to public health threats in South Africa," *The Lancet Microbe*, 1(6), pp. e229–e230. Available at: https://doi.org/10.1016/S2666-5247(20)30116-6.

Msomi, N. *et al.* (2021) "The implementation of SARS-CoV-2 genomic surveillance in South Africa," (May). Available at: http://dx.doi.org/10.2471/BLT.21.286011.

Municipal Demarcation Board (2021) *MDB Spatial Knowledge Hub*. Available at: www.demarcation.org.za; https://spatialhub-mdb-sa.opendata.arcgis.com/ (Accessed: February 22, 2024).

Municipalities of South Africa (2023). Available at: https://municipalities.co.za/provinces/view/4/kwazulu-natal (Accessed: April 17, 2023).

Munung, N.S., Mayosi, B.M. and de Vries, J. (2018) "Genomics research in Africa and its impact on global health: insights from African researchers," *Global Health, Epidemiology and Genomics*, 3. Available at: https://doi.org/10.1017/gheg.2018.3.

Napit, R. *et al.* (2023) "Rapid genomic surveillance of SARS-CoV-2 in a dense urban community of Kathmandu Valley using sewage samples," *PLOS ONE*. Edited by B. Giri, 18(3), p. e0283664. Available at: https://doi.org/10.1371/journal.pone.0283664.

National Institute for Communicable Diseases (2020) *FIRST CASE OF COVID-19 CORONAVIRUS REPORTED IN SA*. Available at: https://www.nicd.ac.za/first-case-of-covid-19-coronavirus-reported-in-sa/ (Accessed: October 12, 2021).

NCBI (2021) *Sequence Read Archive (SRA)*. Available at: https://www.ncbi.nlm.nih.gov/sra (Accessed: December 6, 2021).

Nicola, M. *et al.* (2020) "The socio-economic implications of the coronavirus pandemic (COVID-19): A review," *International Journal of Surgery*, 78, pp. 185–193. Available at: https://doi.org/10.1016/j.ijsu.2020.04.018.

Niemi, M.E.K. *et al.* (2021) "Mapping the human genetic architecture of COVID-19," *Nature*, 600(7889), pp. 472–477. Available at: https://doi.org/10.1038/s41586-021-03767-x.

de Oliveira, T. *et al.* (2022) "Building genomic sequencing capacity in Africa to respond to the SARS-CoV-2 pandemic." Available at: https://www.science.org/do/10.1126/collection.ade0606/full/abbott_booklet_2022_07_29-1686067072893.pdf.

de Oliveira, T. (2022) "Infectious diseases science in Africa takes a leading place in the world," *The Lancet*, 6736(22), pp. 10–11. Available at: https://doi.org/10.1016/S0140-6736(22)00977-1.

de Oliveira, T. and Baxter, C. (2024) "Investing in Africa's scientific future," *Science*, 383(6678). Available at: https://doi.org/10.1126/science.adn4168.

Omotayo, A.O. *et al.* (2021) "Clean water, sanitation and under-five children diarrhea incidence: Empirical evidence from the South Africa's General Household Survey," *Environmental Science and Pollution Research*, 28(44), pp. 63150–63162. Available at: https://doi.org/10.1007/s11356-021-15182-w.

One Health High-Level Expert Panel (OHHLEP) *et al.* (2022) "One Health: A new definition for a sustainable and healthy future," *PLOS Pathogens*. Edited by J.D. Dvorin, 18(6), p. e1010537. Available at: https://doi.org/10.1371/journal.ppat.1010537.

Open AI (2023) *ChatGPT*. Available at: https://chat.openai.com/.

Ortiz-Pineda, P.A. and Sierra-Torres, C.H. (2022) "Evolutionary Traits and Genomic Surveillance of SARS-CoV-2 in South America," *Global Health*. Edited by M. Gadji, 2022, pp. 1–9. Available at: https://doi.org/10.1155/2022/8551576.

O'Toole, Á. *et al.* (2021) "Tracking the international spread of SARS-CoV-2 lineages B.1.1.7 and B.1.351/501Y-V2 with grinch," *Wellcome Open Research*, 6(September), p. 121. Available at: https://doi.org/10.12688/wellcomeopenres.16661.2.

Ozer, E.A. *et al.* (2022) "Multiple expansions of globally uncommon SARS-CoV-2 lineages in Nigeria," *Nature Communications*, 13(1). Available at: https://doi.org/10.1038/s41467-022-28317-5.

Patrick, H.J.H., Chomič, A. and Armstrong, K.F. (2016) "Cooled Propylene Glycol as a Pragmatic Choice for Preservation of DNA From Remote Field-Collected Diptera for Next-Generation Sequence Analysis," *Journal of Economic Entomology*, 109(3), pp. 1469–1473. Available at: https://doi.org/10.1093/jee/tow047.

Paul, P. *et al.* (2023) "Effectiveness of the pre-Omicron COVID-19 vaccines against Omicron in reducing infection, hospitalization, severity, and mortality compared to Delta and other variants: A systematic review," *Human Vaccines & Immunotherapeutics*, 19(1), p. 2167410. Available at: https://doi.org/10.1080/21645515.2023.2167410.

Perez-Gomez, R. (2021) "The Development of SARS-CoV-2 Variants: The Gene Makes the Disease," *Journal of Developmental Biology*, 9(4), p. 58. Available at: https://doi.org/10.3390/jdb9040058.

Pillay, S. *et al.* (2020) "Whole Genome Sequencing of SARS-CoV-2: Adapting Illumina Protocols for Quick and Accurate Outbreak Investigation during a Pandemic," *Genes*, 11(8), p. 949. Available at: https://doi.org/10.3390/genes11080949.

Pillay, S. *et al.* (2023) "Evaluation of miniaturized Illumina DNA preparation protocols for SARS-CoV-2 whole genome sequencing," *PLOS ONE*. Edited by R. Kalendar, 18(4), p. e0283219. Available at: https://doi.org/10.1371/journal.pone.0283219.

Pinkhover, N.P. *et al.* (2022) "SARS-CoV-2 Variant Surveillance Using Tandem Targeted RT-PCR-based Genotyping Assays and Whole Genome Sequencing," *MedRxiv*, (September 2020), pp. 1–7. Available at: https://doi.org/10.1101/2022.01.31.22270226.

Plasek, J.M. *et al.* (2020) "Following data as it crosses borders during the COVID-19 pandemic," *Journal of the American Medical Informatics Association*, 27(7), pp. 1139–1141. Available at: https://doi.org/10.1093/jamia/ocaa063.

Play Your Part - Brand South Africa (2021). Available at: https://brandsouthafrica.com/play-your-part/ (Accessed: August 29, 2023).

Plitnick, J. *et al.* (2021) "Whole-genome sequencing of sars-cov-2: Assessment of the ion torrent ampliseq panel and comparison with the illumina miseq artic protocol," *Journal of Clinical Microbiology*, 59(12), pp. 1–8. Available at: https://doi.org/10.1128/JCM.00649-21.

Potdar, V. *et al.* (2021) "Phylogenetic classification of the whole-genome sequences of SARS-CoV-2 from India & evolutionary trends," *Indian Journal of Medical Research*, 153(1), p. 166. Available at: https://doi.org/10.4103/ijmr.IJMR_3418_20.

Pritchard, E. *et al.* (2022) "Monitoring populations at increased risk for SARS-CoV-2 infection in the community using population-level demographic and behavioural surveillance," *The Lancet Regional Health - Europe*, 13, p. 100282. Available at: https://doi.org/10.1016/j.lanepe.2021.100282.

Pronyk, P.M. *et al.* (2023) "Advancing pathogen genomics in resource-limited settings," *Cell Genomics*, 3(12), p. 100443. Available at: https://doi.org/10.1016/j.xgen.2023.100443.

Pulliam, J.R.C. *et al.* (2022) "Increased risk of SARS-CoV-2 reinfection associated with emergence of Omicron in South Africa," *Science*, 376(6593). Available at: https://doi.org/10.1126/science.abn4947.

Purasram, M. *et al.* (2022) "Knowledge, attitudes and perceptions of a patient population on the COVID-19 vaccine rollout," *Health SA Gesondheid*, 27. Available at: https://doi.org/10.4102/hsag.v27i0.1845.

Quail, M.A. *et al.* (2012) "A tale of three NGS sequencing platforms," *BMC Genomics*, 13(341), p. 13.

Quick, J. *et al.* (2016) "Real-time, portable genome sequencing for Ebola surveillance," *Nature*, 530(7589), pp. 228–232. Available at: https://doi.org/10.1038/nature16996.

Quick, J. *et al.* (2017) "Multiplex PCR method for MinION and Illumina sequencing of Zika and other virus genomes directly from clinical samples," *Nature Protocols*, 12(6), pp. 1261–1266. Available at: https://doi.org/10.1038/nprot.2017.066.

Rabaan, A.A. *et al.* (2021) "Viral dynamics and real-time rt-pcr ct values correlation with disease severity in covid-19," *Diagnostics*, 11(6). Available at: https://doi.org/10.3390/diagnostics11061091.

Rabaan, A.A. *et al.* (2023) "Variants of SARS-CoV-2: Influences on the Vaccines' Effectiveness and Possible Strategies to Overcome Their Consequences," *Medicina (Kaunas, Lithuania)*, 59(3), p. 507. Available at: https://doi.org/10.3390/medicina59030507.

Rachiglio, A.M. *et al.* (2021) "SARS-CoV-2 complete genome sequencing from the Italian Campania region using a highly automated next generation sequencing system," *Journal of Translational Medicine*, 19(1), pp. 1–10. Available at: https://doi.org/10.1186/s12967-021-02912-4.

Raghwani, J. *et al.* (2021) "Genomic epidemiology of early SARS-CoV-2 transmission dynamics in Gujarat, India." Infectious Diseases (except HIV/AIDS). Available at: https://doi.org/10.1101/2021.08.31.21262680.

Rahimi, A., Mirzazadeh, A. and Tavakolpour, S. (2021) "Genetics and genomics of SARS-CoV-2: A review of the literature with the special focus on genetic diversity and SARS-CoV-2 genome detection," *Genomics*, 113(1P2), pp. 1221–1232. Available at: https://doi.org/10.1016/j.ygeno.2020.09.059.

Rambaut, A. *et al.* (2020) "A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology," *Nature Microbiology*, 5(11), pp. 1403–1407. Available at: https://doi.org/10.1038/s41564-020-0770-5.

Ramesh, S. *et al.* (2021) "Emerging SARS-CoV-2 Variants: A Review of Its Mutations, Its Implications and Vaccine Efficacy," *Vaccines*, 9(10), p. 1195. Available at: https://doi.org/10.3390/vaccines9101195.

Ramphal, U. *et al.* (2024) "Reflective Evaluation of Next-Generation Sequencing Data during Early Phase Detection of the Delta Variant," *OBM Genetics*, 08(02), pp. 1–42. Available at: https://doi.org/10.21926/obm.genet.2402239.

Ren, X. *et al.* (2022) "Reinfection in patients with COVID-19: a systematic review," *Global Health Research and Policy*, 7(1), p. 12. Available at: https://doi.org/10.1186/s41256-022-00245-3.

Robishaw, J.D. *et al.* (2021) "Genomic surveillance to combat COVID-19: challenges and opportunities," *The Lancet Microbe*, 2(9), pp. e481–e484. Available at: https://doi.org/10.1016/S2666-5247(21)00121-X.

Roman, A. (2023) "A Closer Look Into Brazil's Healthcare System: What Can We Learn?," *Cureus* [Preprint]. Available at: https://doi.org/10.7759/cureus.38390.

Salehi-Vaziri, M. *et al.* (2022) "The ins and outs of SARS-CoV-2 variants of concern (VOCs)," *Archives of Virology*, 167(2), pp. 327–344. Available at: https://doi.org/10.1007/s00705-022-05365-2.

Salipante, S.J. *et al.* (2014) "Performance comparison of Illumina and Ion Torrent next-generation sequencing platforms for 16S rRNA-based bacterial community profiling," *Applied and Environmental Microbiology*, 80(24), pp. 7583–7591. Available at: https://doi.org/10.1128/AEM.02206-14.

"SARS-CoV-2 Genomic Surveillance Update - September 2024" (2024) *NICD*, 9 September. Available at: https://www.nicd.ac.za/diseases-a-z-index/disease-index-covid-19/sars-cov-2-genomic-surveillance-update/ (Accessed: November 22, 2024).

Schriml, L.M. *et al.* (2020) "COVID-19 pandemic reveals the peril of ignoring metadata standards," *Scientific Data*, 7(1), p. 188. Available at: https://doi.org/10.1038/s41597-020-0524-5.

Schröder, M. *et al.* (2021) "COVID-19 in South Africa: outbreak despite interventions," *Scientific Reports*, 11(1), pp. 1–9. Available at: https://doi.org/10.1038/s41598-021-84487-0.

Seitz, C. (2020) "Genetic Material and Sequence Data to Protect Global Health in the Light of Pandemic Outbreaks: Mapping the Legal Landscape under European and International Law," *European Journal of Health Law*, 27(3), pp. 232–241. Available at: https://doi.org/10.1163/15718093-BJA10017.

Shahzad, F. *et al.* (2020) "Asymmetric nexus between temperature and COVID-19 in the top ten affected provinces of China: A current application of quantile-on-quantile approach," *Science of the Total Environment*, 736, p. 139115. Available at: https://doi.org/10.1016/j.scitotenv.2020.139115.

Shu, Y. and McCauley, J. (2017) "GISAID: Global initiative on sharing all influenza data – from vision to reality," *Eurosurveillance*, 22(13), pp. 2–4. Available at: https://doi.org/10.2807/1560-7917.ES.2017.22.13.30494.

Sofi, M.S., Hamid, A. and Bhat, S.U. (2020) "SARS-CoV-2: A critical review of its history, pathogenesis, transmission, diagnosis and treatment," *Biosafety and Health*, 2(4), pp. 217–225. Available at: https://doi.org/10.1016/j.bsheal.2020.11.002.

Sohrabi, C. *et al.* (2020) "World Health Organization declares global emergency: A review of the 2019 novel coronavirus (COVID-19)," *International Journal of Surgery*, 76(January), pp. 71–76. Available at: https://doi.org/10.1016/j.ijsu.2020.02.034.

South Africa National Department of Health (2024). Available at: https://www.health.gov.za/ (Accessed: May 7, 2024).

South Africa: Urbanization from 2012-2022 (2023). Available at: https://www.statista.com (Accessed: August 22, 2023).

Stark, Z. *et al.* (2019) "Integrating Genomics into Healthcare: A Global Responsibility," *The American Journal of Human Genetics*, 104(1), pp. 13–20. Available at: https://doi.org/10.1016/j.ajhg.2018.11.014.

Statistics South Africa (2020) *COVID-19 Pandemic in South Africa – Demography Volume, Stats SA Library Cataloguing-in-Publication (CIP) Data*. Available at: ISBN 978-0-621-48827-2.

Staunton, C., Swanepoel, C. and Labuschaigne, M. (2020) "Between a rock and a hard place: COVID-19 and South Africa's response," *Journal of Law and the Biosciences*, 7(1), pp. 1–12. Available at: https://doi.org/10.1093/jlb/lsaa052.

Stewart-Ibarra, A.M. *et al.* (2018) "The Burden of Dengue Fever and Chikungunya in Southern Coastal Ecuador: Epidemiology, Clinical Presentation, and Phylogenetics from the First Two Years of a Prospective Study," *The American journal of tropical medicine and hygiene*, 98(5), pp. 1444–1459. Available at: https://doi.org/10.4269/AJTMH.17-0762.

Struelens, M.J. *et al.* (2024) "Real-time genomic surveillance for enhanced control of infectious diseases and antimicrobial resistance," *Frontiers in Science*, 2, p. 1298248. Available at: https://doi.org/10.3389/fsci.2024.1298248.

Subissi, L. *et al.* (2022) "An early warning system for emerging SARS-CoV-2 variants," *Nature Medicine* [Preprint]. Available at: https://doi.org/10.1038/s41591-022-01836-w.

Subramoney, K. *et al.* (2022) "Identification of SARS-CoV-2 Omicron variant using spike gene target failure and genotyping assays, Gauteng, South Africa, 2021," *Journal of Medical Virology*, 94(8), pp. 3676–3684. Available at: https://doi.org/10.1002/jmv.27797.

Szargut, M. *et al.* (2022) "SARS-CoV-2 Whole-Genome Sequencing by Ion S5 Technology—Challenges, Protocol Optimization and Success Rates for Different Strains," *Viruses*, 14(6). Available at: https://doi.org/10.3390/v14061230.

Tajudin, N. *et al.* (2024) "Genomic surveillance and sequencing of SARS-CoV-2 in Malaysia," *Asia Pacific Journal of Molecular Biology and Biotechnology*, pp. 71–83. Available at: https://doi.org/10.35118/apjmbb.2024.032.2.09.

Tao, K. *et al.* (2021) "The biological and clinical significance of emerging SARS-CoV-2 variants," *Nature Reviews Genetics*, 22(12), pp. 757–773. Available at: https://doi.org/10.1038/s41576-021-00408-x.

Tay, J.H. *et al.* (2022) "The Emergence of SARS-CoV-2 Variants of Concern Is Driven by Acceleration of the Substitution Rate.," *Molecular biology and evolution*, 39(2). Available at: https://doi.org/10.1093/molbev/msac013.

Tegally, H. *et al.* (2020) "Emergence and rapid spread of a new severe acute respiratory syndrome-related coronavirus 2 (SARS-CoV-2) lineage with multiple spike mutations in South Africa," *medRxiv*, 2. Available at: https://doi.org/10.1101/2020.12.21.20248640.

Tegally, H., Wilkinson, E., Giovanetti, M., *et al.* (2021) "Detection of a SARS-CoV-2 variant of concern in South Africa," *Nature*, 592(7854), pp. 438–443. Available at: https://doi.org/10.1038/s41586-021-03402-9.

Tegally, H., Wilkinson, E., Althaus, C.L., *et al.* (2021) "Rapid replacement of the Beta variant by the Delta variant in South Africa," *medRxiv*, p. 2021.09.23.21264018. Available at: https://doi.org/10.1101/2021.09.23.21264018.

Tegally, H., Wilkinson, E., Lessells, R.J., *et al.* (2021) "Sixteen novel lineages of SARS-CoV-2 in South Africa," *Nature Medicine*, 27(3), pp. 440–446. Available at: https://doi.org/10.1038/s41591-021-01255-3.

Tegally, H., Moir, M., *et al.* (2022) "Emergence of SARS-CoV-2 Omicron lineages BA.4 and BA.5 in South Africa," *Nature Medicine*, 28(9), pp. 1785–1790. Available at: https://doi.org/10.1038/s41591-022-01911-2.

Tegally, H., San, J.E., *et al.* (2022) "The evolving SARS-CoV-2 epidemic in Africa: Insights from rapidly expanding genomic surveillance," *Science*, 5358. Available at: https://doi.org/10.1126/science.abq5358.

Tegally, H. *et al.* (2023) "Dispersal patterns and influence of air travel during the global expansion of SARS-CoV-2 variants of concern," *Cell*, 186(15), pp. 3277-3290.e16. Available at: https://doi.org/10.1016/j.cell.2023.06.001.

Teixeira da Silva, J.A., Tsigaris, P. and Erfanmanesh, M. (2021) "Publishing volumes in major databases related to Covid-19," *Scientometrics*, 126(1), pp. 831–842. Available at: https://doi.org/10.1007/s11192-020-03675-3.

Tekola-Ayele, F. and Rotimi, C.N. (2015) "Translational Genomics in Low- and Middle-Income Countries: Opportunities and Challenges," *Public Health Genomics*, 18(4), pp. 242–247. Available at: https://doi.org/10.1159/000433518.

Thermo Fisher Scientific (2021) *Target Selection for Ion Torrent Next-Generation Sequencing*. Available at: https://www.thermofisher.com/za/en/home/life-science/sequencing/next-generation-sequencing/ion-torrent-next-generation-sequencing-workflow/ion-torrent-next-generation-sequencing-select-targets.html (Accessed: December 8, 2021).

Tosta, S. *et al.* (2023) "Global SARS-CoV-2 genomic surveillance: What we have learned (so far)," *Infection, Genetics and Evolution*, 108(January), p. 105405. Available at: https://doi.org/10.1016/j.meegid.2023.105405.

Tshiabuila, D. *et al.* (2022) "Comparison of SARS - CoV - 2 sequencing using the ONT GridION and the Illumina MiSeq," *BMC Genomics*, pp. 1–17. Available at: https://doi.org/10.1186/s12864-022-08541-5.

Tyson, J.R. *et al.* (2020) "Improvements to the ARTIC multiplex PCR method for SARS-CoV-2 genome sequencing using nanopore." Genomics. Available at: https://doi.org/10.1101/2020.09.04.283077.

Umunnakwe, C.N. *et al.* (2022) "Evaluation of a commercial SARS-CoV-2 multiplex PCR genotyping assay for variant identification in resource-scarce settings," *PLOS ONE*. Edited by P. Roques, 17(6), p. e0269071. Available at: https://doi.org/10.1371/journal.pone.0269071.

Uvere, E.O. *et al.* (2023) "Capacity-Building for Stroke Genomic Research Data Collection: The African Neurobiobank Ethical, Legal, and Social Implications Project Experience," *Biopreservation and Biobanking*, 21(2), pp. 158–165. Available at: https://doi.org/10.1089/bio.2021.0144.

Varshney, D. *et al.* (2016) "Understanding collaboration in a multi-national research capacity-building partnership: a qualitative study," *Health Research Policy and Systems*, 14(1), p. 64. Available at: https://doi.org/10.1186/s12961-016-0132-1.

Viana, R. *et al.* (2022) "Rapid epidemic expansion of the SARS-CoV-2 Omicron variant in southern Africa," *Nature* [Preprint]. Available at: https://doi.org/10.1038/d41586-021-03832-5.

Vilsker, M. *et al.* (2019) "Genome Detective: an automated system for virus identification from high-throughput sequencing data," *Bioinformatics*, 35(5), pp. 871–873. Available at: https://doi.org/10.1093/bioinformatics/bty695.

Vogels, C.B.F. *et al.* (2021) "Multiplex qPCR discriminates variants of concern to enhance global surveillance of SARS-CoV-2," *PLoS Biology*, 19(5 May), pp. 1–12. Available at: https://doi.org/10.1371/journal.pbio.3001236.

Waitumbi, J.N. *et al.* (2022) "COVID-19 mass testing and sequencing: Experiences from a laboratory in Western Kenya," *African Journal of Laboratory Medicine*, 11(1). Available at: https://doi.org/10.4102/ajlm.v11i1.1737.

Walker, A. *et al.* (2022) "Characterization of Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) Infection Clusters Based on Integrated Genomic Surveillance, Outbreak Analysis and Contact Tracing in an Urban Setting," *Clinical Infectious Diseases*, 74(6), pp. 1039–1046. Available at: https://doi.org/10.1093/cid/ciab588.

Wang, C. *et al.* (2020) "A novel coronavirus outbreak of global health concern," *The Lancet*, 395(10223), pp. 470–473. Available at: https://doi.org/10.1016/S0140-6736(20)30185-9.

Wang, Y. *et al.* (2021) "NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice. 1," *medRxiv*, 1(165), pp. 1–13.

Wilkinson, E. *et al.* (2021) "A year of genomic surveillance reveals how the SARS-CoV-2 pandemic unfolded in Africa," *Science*, 374(6566), pp. 423–431. Available at: https://doi.org/10.1126/science.abj4336.

Wise, J. (2020) "Covid-19: New coronavirus variant is identified in UK.," *BMJ (Clinical research ed.)*, 371, p. m4857. Available at: https://doi.org/10.1136/bmj.m4857.

Wolter, N. *et al.* (2021) "Early assessment of the clinical severity of the SARS-CoV-2 Omicron variant in South Africa," pp. 1–23.

Wolter, N. *et al.* (2022) "Early assessment of the clinical severity of the SARS-CoV-2 omicron variant in South Africa: a data linkage study," *The Lancet*, 399(10323), pp. 437–446. Available at: https://doi.org/10.1016/S0140-6736(22)00017-4.

World Health Organisation (2021) "COVID-19 new variants : Knowledge gaps and research."

World Health Organisation (2023a) *COVAX*. Available at: https://www.who.int/initiatives/act-accelerator/covax (Accessed: August 30, 2023).

World Health Organisation (2023b) *Historical working definitions and primary actions for SARS-CoV-2 variants*. Available at: https://www.who.int/publications/m/item/historical-working-definitions-and-primary-actions-for-sars-cov-2-variants (Accessed: June 23, 2023).

World Health Organisation (2023c) *Technical Advisory Group on SARS-CoV-2 Virus Evolution, World Health Organisation*. Available at: https://www.who.int/groups/technical-advisory-group-on-sars-cov-2-virus-evolution (Accessed: August 22, 2023).

World Health Organisation (2024a) *International Pathogen Surveillance Network launches catalytic grant fund for pathogen genomics*. Available at: https://www.who.int/news/item/23-02-2024-international-pathogen-surveillance-network-launches-catalytic-grant-fund-for-pathogen-genomics (Accessed: May 8, 2024).

World Health Organisation (2024b) *Launch of the Global Initiative on Digital Health*. Available at: https://www.who.int/news-room/events/detail/2024/02/20/default-calendar/launch-of-the-global-initiative-on-digital-health (Accessed: May 8, 2024).

World Health Organisation (2025) *WHO Coronavirus (COVID-19) Dashboard*. Available at: https://covid19.who.int/ (Accessed: February 21, 2024).

Worldometers (2025). Available at: https://www.worldometers.info/coronavirus/ (Accessed: February 7, 2025).

Worobey, M. *et al.* (2022) "The Huanan Seafood Wholesale Market in Wuhan was the early epicenter of the COVID-19 pandemic," *Science*, 377(6609), pp. 951–959. Available at: https://doi.org/10.1126/science.abp8715.

Wu, F. *et al.* (2020) "A new coronavirus associated with human respiratory disease in China," *Nature*, 579(7798), pp. 265–269. Available at: https://doi.org/10.1038/s41586-020-2008-3.

Yang, D. and Leibowitz, J.L. (2015) "The structure and functions of coronavirus genomic 3' and 5' ends," *Virus Research*, 206, pp. 120–133. Available at: https://doi.org/10.1016/j.virusres.2015.02.025.

Ye, Z.-W. *et al.* (2020) "Zoonotic origins of human coronaviruses.," *International journal of biological sciences*, 16(10), pp. 1686–1697. Available at: https://doi.org/10.7150/ijbs.45472.

Yeh, K. *et al.* (2019) "Achieving Health Security and Threat Reduction through Sharing Sequence Data," *Tropical Medicine and Infectious Disease*, 4(2), p. 78. Available at: https://doi.org/10.3390/tropicalmed4020078.

Yin, Y. and Wunderink, R.G. (2018) "MERS, SARS and other coronaviruses as causes of pneumonia," *Respirology*, 23(2), pp. 130–137. Available at: https://doi.org/10.1111/resp.13196.

Zhou, P. *et al.* (2020) "A pneumonia outbreak associated with a new coronavirus of probable bat origin," *Nature*, 579(7798), pp. 270–273. Available at: https://doi.org/10.1038/s41586-020-2012-7.

Zuckerman, N.S. *et al.* (2021) "Prolonged detection of complete viral genomes demonstrated by SARS-CoV-2 sequencing of serial respiratory specimens," *PLoS ONE*, 16(8 August), pp. 1–7. Available at: https://doi.org/10.1371/journal.pone.0255691.

Appendices

Appendix A – Accession numbers

Chapter 2 dataset

Epidemic data is freely available on the open-source Global Initiative on Sharing All Influenza Data (GISAID) database and accessed using the GISAID Identifier: EPI_SET_230203ym; doi: https://doi.org/10.55876/gis8.230203ym.

Chapter 3 dataset

South African genomic sequences dataset (EPI_SET_230908yt, doi: https://doi.org/10.55876/gis8.230908yt) and the KZN sample cohort (EPI_SET_230914wa, doi: https://doi.org/10.55876/gis8.230914wa) and corresponding metadata files can be accessed from GISAID.

Chapter 4 dataset

South African dataset used for comparison with global trends (EPI_SET_250217no, doi: https://doi.org/10.55876/gis8.250217no). Other GISAID data was obtained from "EpiCoV™" - select "Transmission tracker", and "Downloads" tab – under "Submission and Variant statistics", select "Global by month (xlsx)", and "Genomic epidemiology", select "nextregions".

Appendix B – Supplementary information

Chapter 2 Supplementary Tables

Table S2.1: Summary of sample set used for comparison of NGS platforms

Run no.	No. of samples	Collection dates	Extraction date	Location	Uploaded to GISAID
Run A	92	May 2021	15 June 2021	NHLS Eastern Cape	86 (93,5%)
Run B	91	March to June 2021	15 June 2021	NHLS and Private labs in Eastern Cape	70 (76,9%)

Table S2.2: GISAID Accession identifiers (EPI_SET ID: EPI_SET_230203ym**,** doi:10.55876/gis8.230203ym)

GISAID accession ID			
EPI_ISL_3275348	EPI_ISL_2693858	EPI_ISL_2693894	EPI_ISL_2727200
EPI_ISL_2693824	EPI_ISL_2693859	EPI_ISL_2693895	EPI_ISL_2727201
EPI_ISL_2693825	EPI_ISL_2693860	EPI_ISL_2693896	EPI_ISL_2727202
EPI_ISL_2693826	EPI_ISL_2693861	EPI_ISL_2693897	EPI_ISL_2727203
EPI_ISL_2693827	EPI_ISL_2693862	EPI_ISL_2693898	EPI_ISL_2727204
EPI_ISL_2693828	EPI_ISL_2693863	EPI_ISL_2693899	EPI_ISL_2727205
EPI_ISL_3275349	EPI_ISL_2693864	EPI_ISL_2693900	EPI_ISL_2727206
EPI_ISL_2693829	EPI_ISL_2693865	EPI_ISL_2693901	EPI_ISL_2727207
EPI_ISL_2693830	EPI_ISL_2693866	EPI_ISL_3275374	EPI_ISL_2727208
EPI_ISL_2693831	EPI_ISL_2693867	EPI_ISL_3275375	EPI_ISL_2727209
EPI_ISL_3275350	EPI_ISL_2693868	EPI_ISL_2727175	EPI_ISL_2727210
EPI_ISL_2693832	EPI_ISL_2693869	EPI_ISL_2727176	EPI_ISL_2727211
EPI_ISL_2693833	EPI_ISL_2693870	EPI_ISL_2727177	EPI_ISL_2727212
EPI_ISL_2693834	EPI_ISL_2693871	EPI_ISL_2727178	EPI_ISL_2727213
EPI_ISL_2693835	EPI_ISL_2693872	EPI_ISL_2727179	EPI_ISL_2727214
EPI_ISL_2693836	EPI_ISL_2693873	EPI_ISL_2727180	EPI_ISL_2727215
EPI_ISL_2693837	EPI_ISL_2693874	EPI_ISL_3275376	EPI_ISL_2727216
EPI_ISL_2693838	EPI_ISL_2693875	EPI_ISL_3275377	EPI_ISL_2727217
EPI_ISL_3275351	EPI_ISL_2693876	EPI_ISL_3275378	EPI_ISL_2727218
EPI_ISL_2693839	EPI_ISL_2693877	EPI_ISL_2727181	EPI_ISL_2727219
EPI_ISL_2693840	EPI_ISL_2693878	EPI_ISL_2727182	EPI_ISL_2727220
EPI_ISL_2693841	EPI_ISL_2693879	EPI_ISL_2727183	EPI_ISL_2727221
EPI_ISL_2693842	EPI_ISL_2693880	EPI_ISL_2727184	EPI_ISL_2727222
EPI_ISL_3275352	EPI_ISL_2693881	EPI_ISL_2727185	EPI_ISL_2727223
EPI_ISL_2693843	EPI_ISL_2693882	EPI_ISL_2727186	EPI_ISL_2727224
EPI_ISL_2693844	EPI_ISL_2693883	EPI_ISL_2727187	EPI_ISL_2727225
EPI_ISL_2693845	EPI_ISL_2693884	EPI_ISL_2727188	EPI_ISL_2727226
EPI_ISL_2693846	EPI_ISL_3275353	EPI_ISL_3275379	EPI_ISL_2727227
EPI_ISL_2693847	EPI_ISL_2693885	EPI_ISL_2727189	EPI_ISL_2727228
EPI_ISL_2693848	EPI_ISL_2693886	EPI_ISL_2727190	EPI_ISL_2727229
EPI_ISL_2693849	EPI_ISL_2693887	EPI_ISL_2727191	EPI_ISL_2727230
EPI_ISL_2693850	EPI_ISL_3275354	EPI_ISL_2727192	EPI_ISL_2727231
EPI_ISL_2693851	EPI_ISL_2693888	EPI_ISL_2727193	EPI_ISL_2727232
EPI_ISL_2693852	EPI_ISL_2693889	EPI_ISL_2727194	EPI_ISL_2727233
EPI_ISL_2693853	EPI_ISL_2693890	EPI_ISL_2727195	EPI_ISL_2727234
EPI_ISL_2693854	EPI_ISL_2693891	EPI_ISL_2727196	EPI_ISL_2727235
EPI_ISL_2693855	EPI_ISL_3275355	EPI_ISL_2727197	EPI_ISL_2727236
EPI_ISL_2693856	EPI_ISL_2693892	EPI_ISL_2727198	EPI_ISL_2727237
EPI_ISL_2693857	EPI_ISL_2693893	EPI_ISL_2727199	EPI_ISL_2727238

Table S2.3: Sequencing data for runs performed on the Illumina MiSeq

Run no.	Date of Sequencing	Success Rate of Run (%)	Greater than 80% coverage	Cluster Density (K/mm^2)	Cluster Passing filter (%)	Mean read Length (bp)	Coverage Depth	Q30 Score (%)	Data Obtained (Mb)
Run A	June 2021	93,5 %	90,7 %	1219	88,7	400	50X	N/A*	10485
Run B	June 2021	94,5 %	69,8 %	1293	92,3	400	50X	79,6	11681,3

* Not captured upon run completion

Table S2.4: Sequencing data for runs performed on the Ion Torrent S5

Run no.	Date of Sequencing	Success Rate of Run (%)	Greater than 80% coverage	Total Reads	Mean Read length (bp)	Average Coverage depth of Reference
Run A	29 June 2021	100 %	100 %	97,119,395	192	235.4X
Run B	07 July 2021	91,2 %	98,8 %	81,275,308	185	176.5X

Table S2.5: Absence of S-gene coverage on the Illumina MiSeq due to low viral load

Sample ID	Estimated Viral load	MiSeq				Ion Torrent S5			
		GC (%)	S-gene coverage	Clade	Mutations (S-gene)	GC (%)	S-gene coverage	Clade	Mutation (S-gene)
K016879	low	2,6	None	19A	-	100,0	100,0%	19A	*S:H66Y, S:T478K, S:N501Y, S:D614G*
K016899	-	13,5	None	20C	-	98,1	98,6%	19A	*S:A701V*
K016931	low	9,3	None	21A (Delta)	-	99,7	100,0%	20H (Beta, V2)	*S:K417N, S:E484K, S:N501Y, S:D614G, S:A701V*
K016938	low	3,3	None	19A	-	82,4	100,0%	19A	-
K016943	low	5,2	None	19B	-	97,7	93,1%	20A	*S:D614G*
K016950	low	6,8	None	19A	-	99,7	100,0%	19A	*S:D614G*

Table S2.6: Total mutations obtained betIen samples sequenced on the Ion torrent S5 and Illumina MiSeq platforms

Sample Lims ID	Ion Torrent S5	Illumina Miseq
	No. of mutations	No. of mutations
K016680	50	31
K016681	45	18
K016682	46	43
K016683	51	43
K016684	47	44
K016685	51	47
K016686	45	42
K016687	50	
K016688	45	43
K016689	53	48
K016690	51	42
K016691	48	1
K016692	50	41
K016693	51	47
K016694	14	48
K016695	53	48
K016696	49	48
K016697	49	48
K016698	49	48
K016699	50	
K016700	55	53
K016701	51	39
K016702	51	51
K016703	55	50
K016704	51	
K016705	47	45
K016706	50	47
K016707	49	47
K016708	50	49
K016709	17	45
K016710	51	48
K016711	48	48
K016712	48	46
K016713	47	47
K016714	51	41
K016715	14	5
K016716	44	43
K016717	44	42
K016718	30	20
K016719	20	1
K016720	53	47
K016721	54	48
K016722	46	44
K016723	46	43
K016724	36	42
K016725	44	43
K016726	46	43
K016727	54	50
K016728	52	47
K016729	47	48
K016730	52	48
K016731	50	52
K016732	44	42
K016733	54	51
K016734	47	46
K016735	52	49
K016736	44	43

K016737	49	45
K016738	54	53
K016739	44	44
K016740	53	48
K016741	53	49
K016742	53	48
K016743	52	48
K016744	51	48
K016745	52	49
K016746	47	45
K016747	48	46
K016748	45	43
K016749	50	47
K016750	52	
K016751	53	47
K016752	53	49
K016753	51	48
K016754	20	25
K016755	50	
K016756	49	47
K016757	46	44
K016758	53	73
K016759	54	50
K016760	29	1
K016761	52	49
K016762	64	63
K016763	43	42
K016764	42	40
K016765	55	48
K016766	53	49
K016767	53	48
K016768	54	49
K016769	38	
K016770	53	48
K016771	53	48
K016864	50	
K016865	53	35
K016866	50	46
K016867	51	47
K016868	64	53
K016869	48	45
K016870	15	5
K016871	46	44
K016872	51	44
K016873	46	
K016874	49	38
K016875	44	
K016876	50	40
K016877	51	40
K016878	51	40
K016879	24	0
K016880	51	39
K016881	52	47
K016882	80	66
K016883	69	18
K016884	37	33
K016885		40
K016886	58	
K016887	49	44
K016888	48	45
K016889	21	
K016890	39	47
K016891	44	48
K016892	52	48

K016893	94	92
K016894	45	45
K016895	41	23
K016896	50	48
K016897		51
K016898	46	44
K016899	11	9
K016900	47	44
K016901	52	46
K016902	13	37
K016903	54	53
K016904	53	46
K016905	45	44
K016906	49	46
K016907	46	30
K016908	22	22
K016909		45
K016910	47	46
K016911	37	41
K016912	48	44
K016913	49	47
K016914	49	47
K016915	10	0
K016916	49	45
K016917	32	42
K016918	52	47
K016919	56	49
K016920	14	23
K016921		39
K016922		46
K016923	74	2
K016924	50	45
K016925	48	44
K016926	11	1
K016927	53	48
K016928	72	38
K016929	54	52
K016930	51	51
K016931	38	2
K016932	45	26
K016933		45
K016934	47	37
K016935	46	43
K016936	69	28
K016937	48	45
K016938	0	0
K016939	52	48
K016940	25	12
K016941	50	36
K016942	53	50
K016943	14	3
K016944		6
K016945	6	22
K016946	48	47
K016947	53	47
K016948		51
K016949	56	51
K016950	21	1
K016951	54	49
K016952	54	50
K016953	56	50
K016954	52	49

Table S3.1: KwaZulu-Natal demographic representation across the first three years of the SARS-CoV-2 pandemic (March 2020 to March 2023)

Demographic Characteristic	No. of Genomes per Year				Total Genomes (*n* = 6089)
	2020 (*n* = 1940)	2021 (*n* = 2584)	2022 (*n* = 1381)	2023 (*n* = 184)	
Gender (*n* (%))					
Female	1226 (63,2%)	1496 (57,9%)	777 (56,3%)	104 (56,5%)	3603 (59,2%)
Male	625 (32,2%)	998 (38,6%)	540 (39,1%)	73(39,7%)	2236 (36,7%)
Unknown	89 (4,6%)	90 (3,5%)	64 (4,6%)	7 (3,8%)	250 (4,1%)
Age Groups (*n* (%))					
<0 - 18	214 (11,0%)	485 (18,8%)	168 (12,2%)	32 (17,4%)	899 (14,8%)
19 - 30	413 (21,3%)	583 (22,6%)	291 (21,1%)	29 (15,8%)	1316 (21,6%)
31 - 40	429 (22,1%)	505 (19,5%)	278 (20,1%)	35 (19,0%)	1247 (20,5%)
41 - 50	310 (16,0%)	404 (15,6%)	221 (16,0%	25 (13,6%)	960 (15,8%)
51 - 60	241 (12,4%)	295 (11,4%)	180 (13,0%)	29 (15,8%)	745 (12,2%)
>60	275 (16,8%)	276 (10,7%)	195 (14,1%)	32 (17,4%)	778 (12,8%)
Unknown	58 (3,0%)	36 (1,4%)	48 (3,5%)	2 (1,1%)	144 (2,4%)
KwaZulu-Natal Districts (*n* (%))					
Amajuba	77 (4,0%)	42 (1,6%)	5 (0,4%)	0 (0,0%)	124 (2,0%)
eThekwini	641 (33,0%)	989 (38,3%	1123 (81,3%)	117 (63,6%)	2870 (47,1%)
Harry Gwala	87 (4,5%)	64 (2,5%)	4 (0,3%)	3 (1,6%)	158 (2,6%)
iLembe	263 (13,6%)	261 (10,1%)	33 (2,4%)	36 (19,6%)	593 (9,7%)
King Cetshwayo	84 (4,3%)	445 (17,2%)	16 (1,2%)	1 (0,5%)	546 (9,0%)
Ugu	205 (10,6%)	205 (7,9%)	108 (7,8%)	22 (12,0%)	540 (8,9%)
uMgungundlovu	125 (6,4%)	91 (3,5%)	67 (4,9%)	0 (0,0%)	283 (4,6%)
uMkanyakude	202 (10,4%)	152 (5,9%)	3 (0,2%)	1 (0,5%)	358 (5,9%)
uMzinyathi	52 (2,7%)	11 (0,4%)	1 (0,1%)	0 (0,0%)	64 (1,1%)
uThukela	86 (4,4%)	21 (0,8%)	10 (0,7%)	0 (0,0%)	117 (1,9%)
Zululand	118 (6,1%)	303 (11,7%)	11 (0,8%)	4 (2,2%)	436 (7,2%)

Unknown represents those individuals for which gender and age was not captured at the time of specimen collection

Table S3.2: Genome distribution across variants per district in KwaZulu-Natal

	Amajuba	eThekwini	Harry Gwala	iLembe	King Cetshwayo	Ugu	uMgungundlovu	uMkanyakude	uMzinyathi	uThukela	Zululand	Total genomes
Alpha	0	2	0	2	3	0	0	0	0	1	1	**9**
Beta	30	214	60	170	131	143	26	151	3	9	230	**1167**
Delta And Sublineages (B.1.617.2 And AY.X)	0	441	21	66	237	87	52	15	3	7	53	**982**
Omicron And Sublineages (B.1.1.529 And Ba.X)	15	1257	8	72	92	112	88	24	6	21	30	**1725**
C.1.X	14	32	1	15	12	5	7	11	8	14	23	**142**
B Lineages	63	548	67	224	70	149	99	155	43	64	94	**1576**
Others	2	376	1	44	1	44	11	2	1	1	5	**488**
Total Variants	**124**	**2870**	**158**	**593**	**546**	**540**	**283**	**358**	**64**	**117**	**436**	**6089**

Table S3.3: Lineage distribution across genomes per district in KwaZulu-Natal

	Amajuba	eThekwini	Harry Gwala	iLembe	King Cetshwayo	Ugu	uMgungundlovu	uMkanyakude	uMzinyathi	uThukela	Zululand	Total genomes
Alpha	0	1	0	1	1	0	0	0	0	1	1	**1**
Beta	1	2	1	2	1	2	1	1	1	1	1	**2**
Delta And Sublineages (B.1.617.2 And AY.X)	0	16	4	7	9	9	6	2	1	2	5	**17**
Omicron And Sublineages (B.1.1.529 And Ba.X)	3	38	6	16	17	20	15	10	2	8	11	**42**
C.1.X	1	2	1	2	2	2	2	1	1	1	1	**2**
B Lineages	9	17	10	13	12	15	9	12	8	9	11	**23**
Others	2	40	1	19	1	12	3	2	1	1	4	**47**
Total lineages	**16**	**116**	**23**	**60**	**43**	**60**	**36**	**28**	**14**	**23**	**34**	**134**

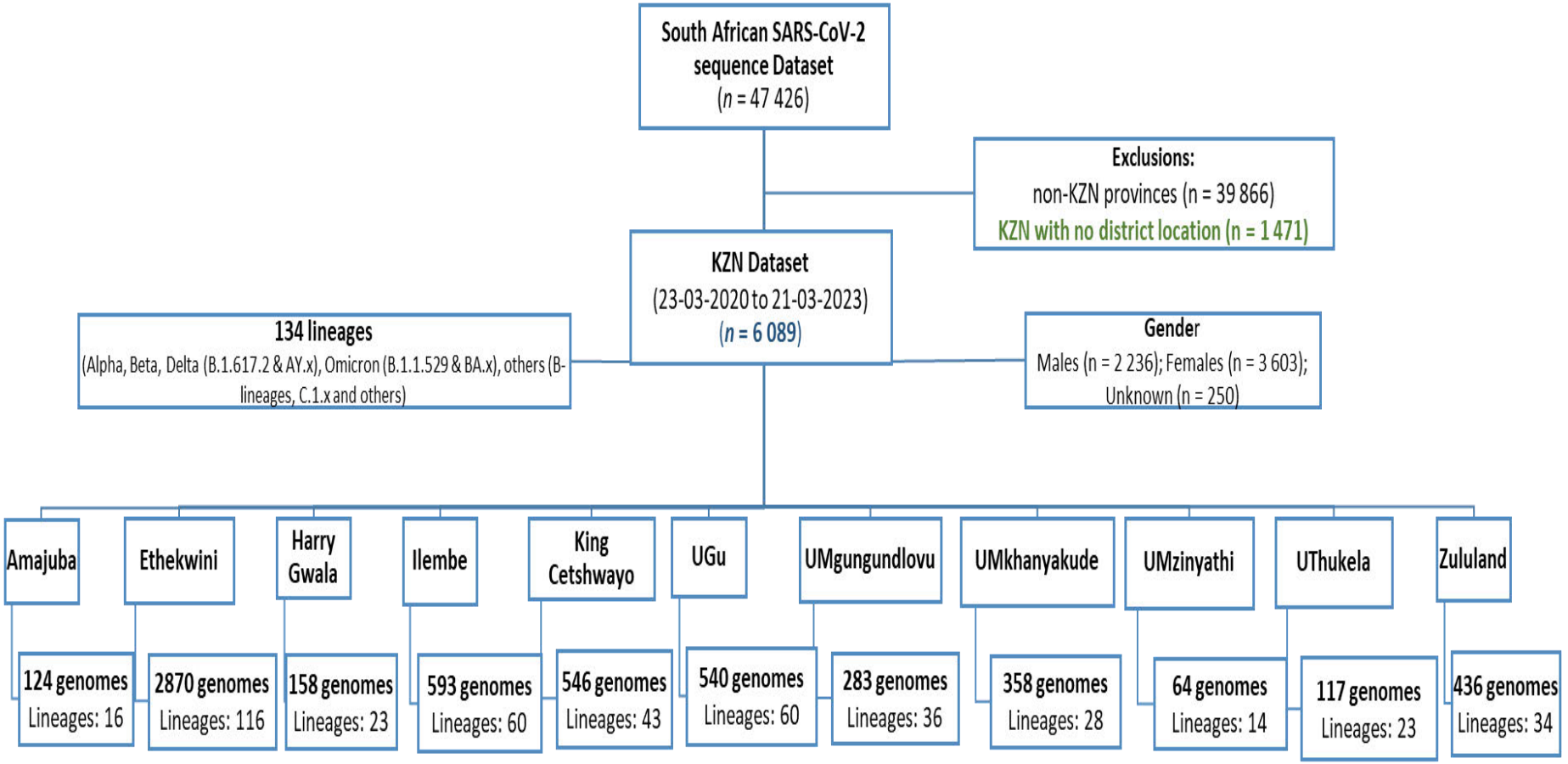


Figure S3.1: Consortium diagram of KZN Cohort of SARS-CoV-2 in KZN, 2020 – 2023. The illustration displays the distribution of 6089 genomes and 134 lineages across the 11 districtsand presents the VOCs that Ire in circulation within the district throughout 2020 and 2023.

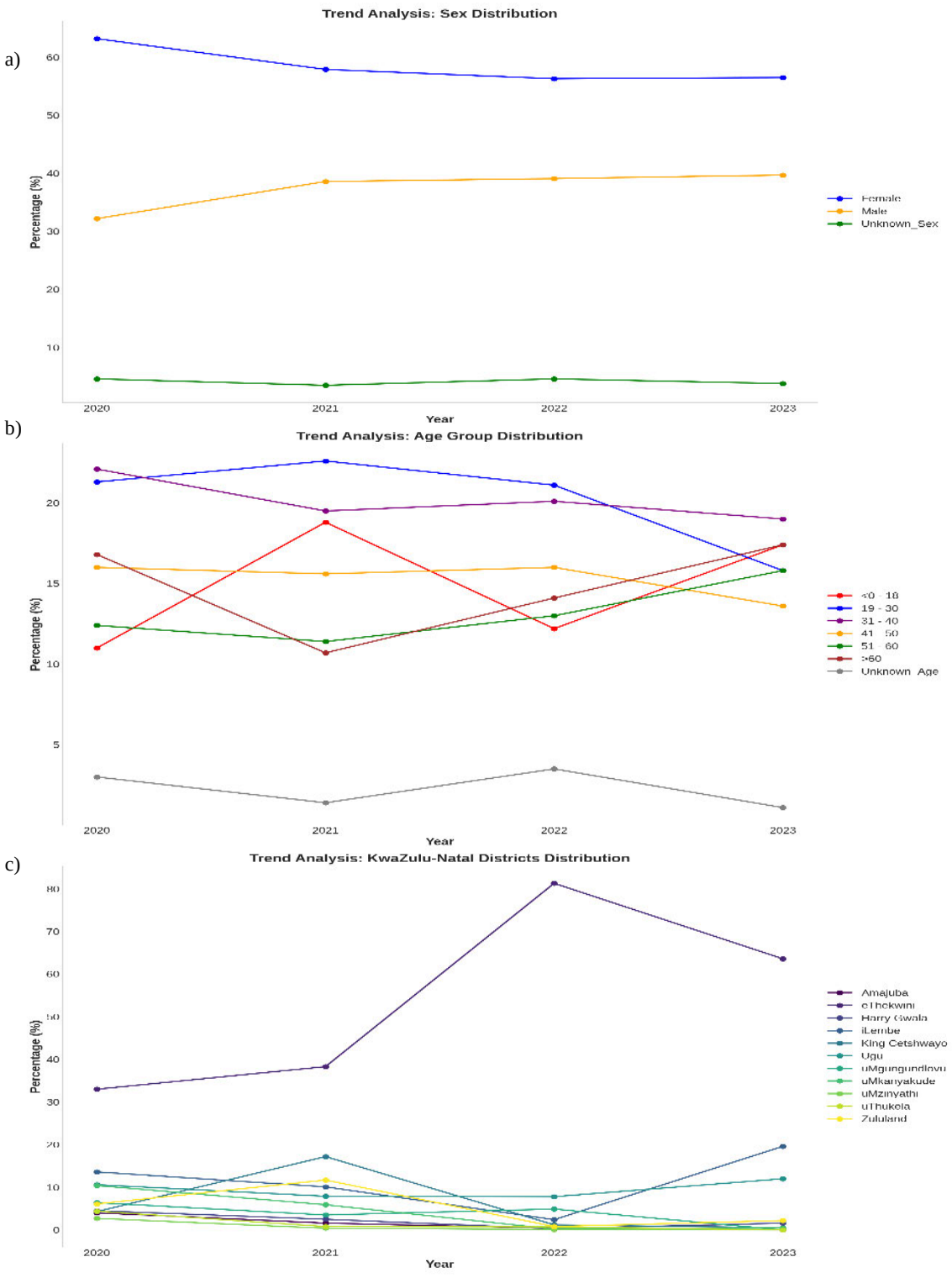


Figure S3.2: Valuable insights into the distribution trends of genomes across a) gender, b) age and c) districts over the first three years of the endemic

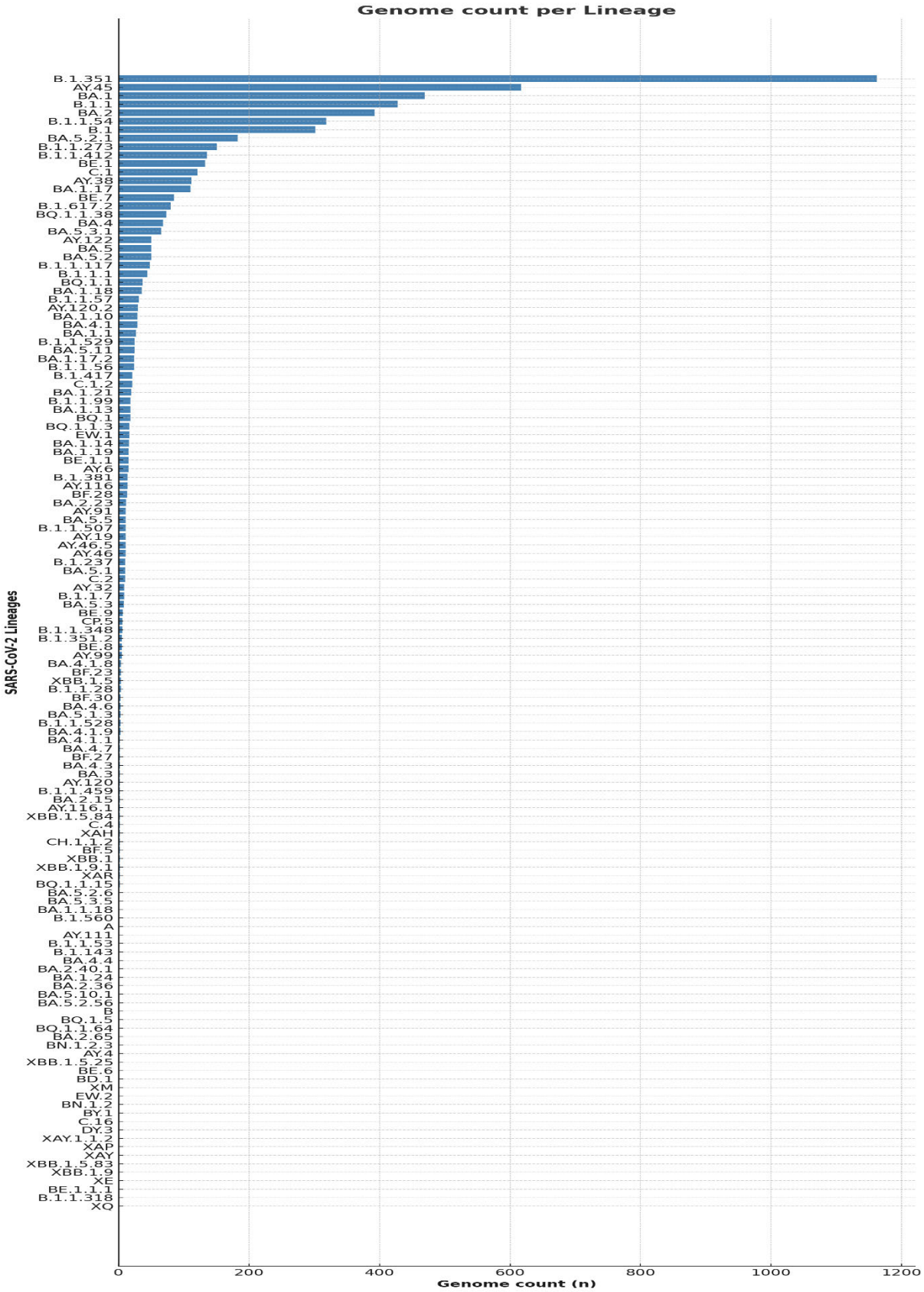


Figure S3.3: SARS-CoV-2 Genome count by lineage (2020 – 2023). All lineages that Ire present in the KZN genome dataset are presented (n = 1 to 1162).

Figure S3.4: Lineage distribution across individual districts within KwaZulu-Natal (March 2020 – March 2023)

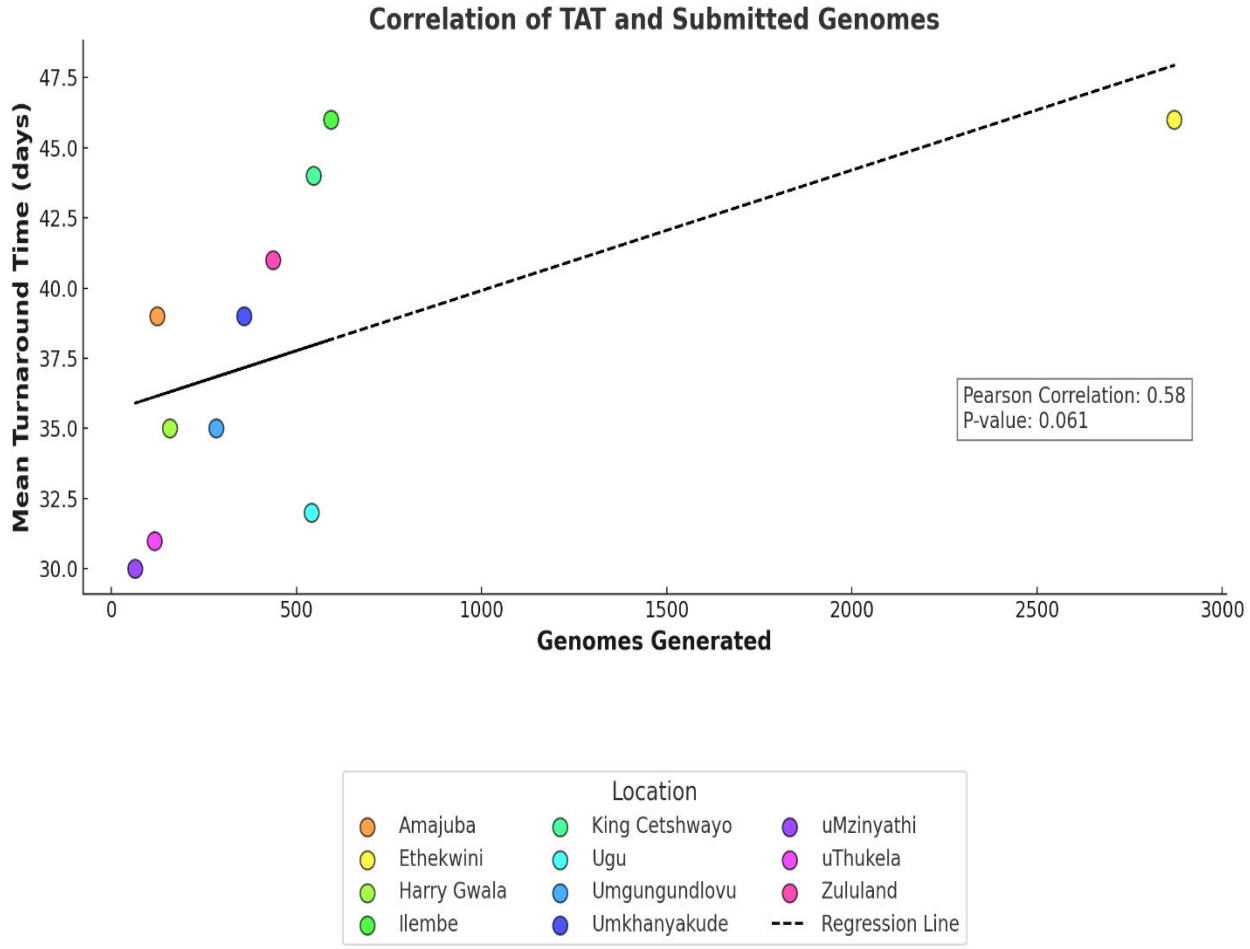


Figure S3.5: Linear Regression of the Mean turnaround time and no. of genomes submitted per district. (Persons correlation: 0.58; p-value: 0.061 *ns*)

Table S4.1: Country distribution of publications, genomes, total cases, total deaths, total tests, population and prevalence

Country	*Publications[1]*	*Genomes on GISAID[2]*	*Total Cases[3]*	*Total deaths[3]*	*Total tests[3]*	*Population[3]*	*Prevalence (per 100 persons)*
USA	133 040	5 019 145	110 533 193	1 192 248	1 186 623 278	334 805 269	33,0
UK	52 579	3 136 948	24 863 166	232 112	522 526 476	68 497 907	36,3
Australia	19 642	227 819	11 771 709	2 391	81 916 639	26 068 792	45,2
Denmark	4 148	659 589	3 183 756	8 814	129 333 107	5 834 950	54,6
Canada	22 676	607 082	4 891 249	57 274	66 525 130	38 388 419	12,7
Singapore	4 838	46 373	2 945 715	1 933	24 756 666	5 943 546	49,6
India	32 803	296 248	45 022 219	533 417	935 879 495	1 406 631 776	3,2
Brazil	17 980	246 629	38 230 814	708 739	63 776 166	215 353 593	17,8
Thailand	3 693	44 006	4 763 664	34 528	17 270 775	70 078 203	6,8
Kenya	1 357	12 991	344 130	5 689	3 967 062	56 215 221	0,6
Nigeria	3 470	8 187	267 173	3 155	5 708 974	216 746 934	0,1
South Africa	7 106	55 465	4 076 463	102 595	26 795 090	60 756 135	6,7
World	526 540	16 446 862	701 872 986	6 969 354		8 100 000 000	8,7
**Greenland*	-	No data	11 971	21	No data	56 856	21,1
**Uzbekistan*	-	325	175 082	1 016	No data	36 410 000	0,5
**Myanmar*	-	328	643 241	19 494	8 330 000	54 580 000	1,2

[1] Mean publication counts obtained from Scopus and WoS (date accessed: 24/01/2024)

[2] Total genomes publicly available on GISAID (dated accessed: 19/02/2024)

[3] Data obtained from https://www.worldometers.info/coronavirus (dated accessed: 19/02/2024)

**Total genomes publicly available on GISAID (date accessed: 27/01/2025) and total cases and deaths obtained from OWID (date accessed: 26/01/2025)*

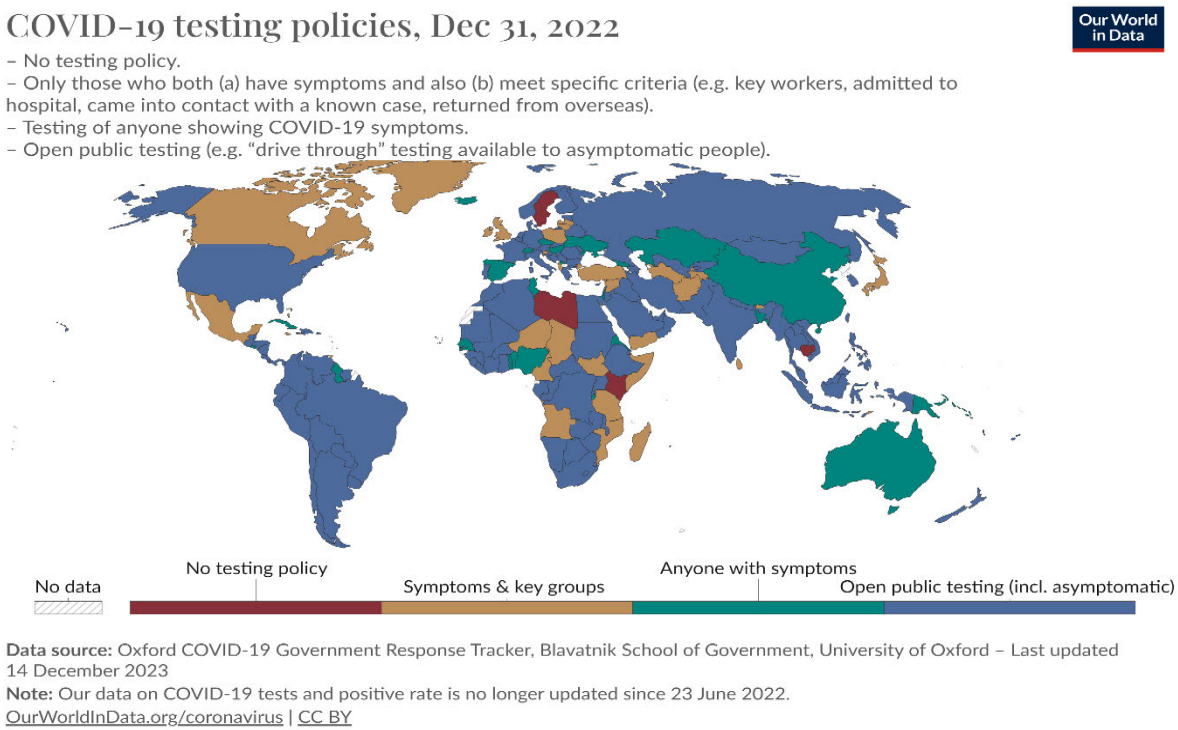


Figure S4.1: Global data related to COVID-19 testing policies sourced from Our World in Data (OWID) at https://ourworldindata.org/ (date accessed: 19/02/2024)

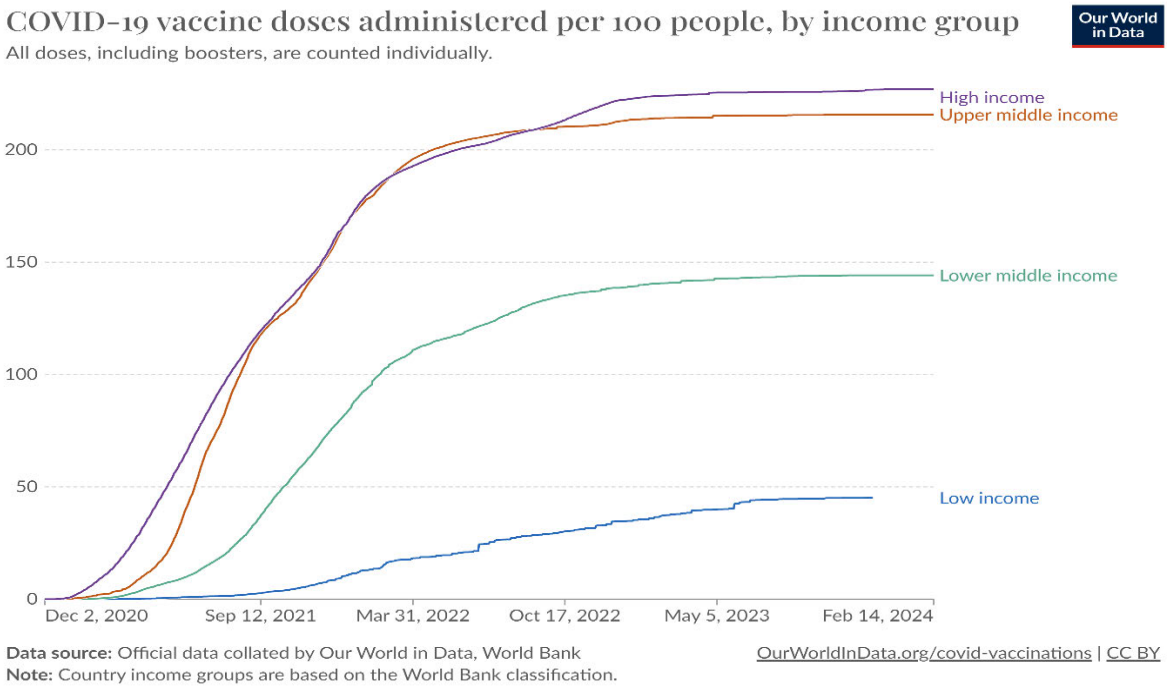


Figure S4.2: Vaccine administration based on income groups, sourced from Our World in Data (OWID) at https://ourworldindata.org/ (date accessed 19/02/2024)

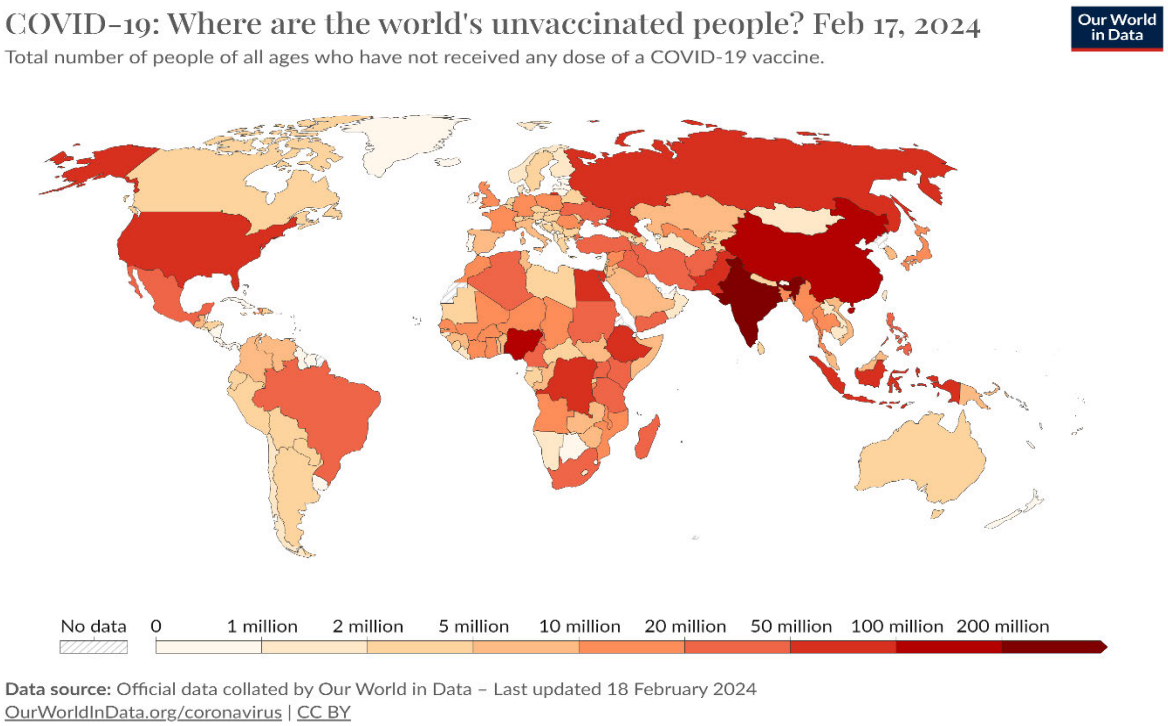


Figure S4.3: Global distribution of unvaccinated people, sourced from Our World in Data (OWID) at https://ourworldindata.org/ (date accessed: 19/02/2024)

Supplementary Files

Supplementary File S1:

SUPPLEMENTAL TABLE

Data Availability

GISAID Identifier: EPI_SET_230908yt
doi: 10.55876/gis8.230908yt

All genome sequences and associated metadata in this dataset are published in GISAID's EpiCoV database. To view the contributors of each individual sequence with details such as accession number, Virus name, Collection date, Originating Lab and Submitting Lab and the list of Authors, visit 10.55876/gis8.230908yt

Data Snapshot

- EPI_SET_230908yt is composed of 47,426 individual genome sequences.
- The collection dates range from 2020-03-06 to 2023-03-28;
- Data were collected in 1 countries and territories;
- All sequences in this dataset are compared relative to hCoV-19/Wuhan/WIV04/2019 (WIV04), the official reference sequence employed by GISAID (EPI_ISL_402124). Learn more at https://gisaid.org/WIV04.

Supplementary File S2:

SUPPLEMENTAL TABLE

Data Availability

GISAID Identifier: EPI_SET_230914wa
doi: 10.55876/gis8.230914wa

All genome sequences and associated metadata in this dataset are published in GISAID's EpiCoV database. To view the contributors of each individual sequence with details such as accession number, Virus name, Collection date, Originating Lab and Submitting Lab and the list of Authors, visit 10.55876/gis8.230914wa

Data Snapshot

- EPI_SET_230914wa is composed of 6,089 individual genome sequences.
- The collection dates range from 2020-03-23 to 2023-03-21;
- Data were collected in 1 countries and territories;
- All sequences in this dataset are compared relative to hCoV-19/Wuhan/WIV04/2019 (WIV04), the official reference sequence employed by GISAID (EPI_ISL_402124). Learn more at https://gisaid.org/WIV04.

Appendix C – Ethics certificates

UNIVERSITY OF KWAZULU-NATAL
INYUVESI YAKWAZULU-NATALI

28 October 2022

Miss Upasana Ramphal (200303380)
School of Lab Med & Medical Sc
Medical School

Dear Miss Ramphal,

Protocol reference number: BREC/00004745/2022
Project title: Benchmarking South African genomic surveillance strategies: A look at genomic surveillance and epidemiological evaluation of SARS-CoV-2 transmission within Southern Africa
Degree: PhD

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 28 October 2022. Please ensure that any outstanding site permissions are obtained and forwarded to BREC for approval before commencing research at a site.

This approval is valid for one year from 28 October 2022. To ensure uninterrupted approval of this study beyond the approval expiry date, an application for recertification must be submitted to BREC on RIG 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 13 December 2022.

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

UNIVERSITY OF KWAZULU-NATAL
INYUVESI YAKWAZULU-NATALI

24 August 2023

Miss Upasana Ramphal (200303380)
School of Laboratory Medicine & Medical Science
Medical School

Dear Miss Ramphal,

Protocol reference number: BREC/00004745/2022
Project title: Benchmarking South African genomic surveillance strategies: A look at genomic surveillance and epidemiological evaluation of SARS-CoV-2 transmission within Southern Africa
Degree: PhD

RECERTIFICATION APPLICATION APPROVAL NOTICE

Approved:	28 October 2023
Expiration of Ethical Approval:	28 October 2024

I wish to advise you that your application for recertification 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 12 September 2023.

Yours sincerely

[redacted]

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

Appendix D – Prompts Used for Generative AI Tools

Assistance with sentence construction was partially informed by the guidance provided in (Braun and Benstoem, 2023) and by prompt usage suggestions obtained from social platforms. Traditional AI tools (Grammarly and QuillBot) Ire employed to enhance the clarity and grammatical accuracy of poorly structured sentences through general rewording. Below is a comprehensive record of the prompts utilised during interactions with generative AI tools in the development of this thesis:

1. **Refinement of Writing and Concepts within Chapters 1 - 5**:

- "Suggest improvements in sentence construction while maintaining my original research and meaning."
- "Highlight any grammatical errors or other mistakes in the following sections of my thesis"
- "Does this sentence make sense and does it read Ill?"
- "Suggest how to define the following aims of my research in a better manner in my thesis background while maintaining its originality"
- "Recommend strategies and techniques to refine my flow and clarity, ensuring it remains my original idea."
- "What are some effective strategies for improving the clarity and coherence of my thesis structure while ensuring it remains true to my original research and findings?"
- "Shorten the following content for my acknowledgments page and make it more concise"
- "This is the draft of my main abstract. My aim is to refine the structure to ensure it develops logically and adheres to academic standards. Can you assist me in improving the transitions, logical flow and sentence clarity, while maintaining my original content and data interpretation?"
- "Suggest five alternative words for the following without altering the meaning of the sentence: Significant, hypothesise, retrospective, highlight, significance, improve, explore, mitigate, enhance, rectify, substantiate."
- "Shorten the following conclusion I have written and polish the language used"

2. **Creation of Visual Content**:

- Figure 1.4: Use of specialised design tool to design and create a visual overview of VOCs
- Figure 4.1: "Generate a visual interpretation of a genomic surveillance system for SARS-CoV-2. Include visual representation of components such as sequencing platforms, genomes, resources, data sharing networks, sequences, graphs and plots, data and bioinformatics pipelines. Keep the tone blue with clear and concise images in similar shades that provide contrast."

3. **Assistance with R and Python Programming code**:

- "Explain the error and/or warning message displayed and advise how I can correct it."

‘Warning message: In plot.xy (xy.coords(data, lineages), lineage = lineage): "lineage” not implemented’

‘The following object is masked from “package:stats”: filter’

‘Error in ‘[.data.frame`(df, , “missing_col”): undefined columns selected’

‘Error: package or namespace load failed for “pkg”: object “func” is not exported by 'namespace:other_pkg'

‘as.string(c (“Thekwini”, “ILembe”, “Ugu”,….., “zululand)), # Warning: NAs introduced by coercion’

‘Error: Aesthetics must be either length 1 or the same as the data (5): fill’

‘Error: stat_count() must not be used with a y aesthetic.’

‘Warning: Removed 1 rows containing missing values (geom_point)’

‘Warning: Removed 1 rows containing non-finite values (stat_smooth).’

‘Warning: Removed 1 rows containing missing values (geom_point).’