Molecular epidemiology and characteristics of immune adaptations across the SARS-CoV-2 spike glycoproteins from Gauteng, South Africa, 2020 to 2022

Kathleen Subramoney

This thesis was submitted to the Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, in fulfilment of the requirement for the degree of Doctor of Philosophy

Johannesburg, June 2024

Declaration

I, Kathleen Subramoney, declare that this thesis is my own, unaided work. It is being submitted for the degree of Doctor of Philosophy at the University of the Witwatersrand, Johannesburg. It has not been submitted before for any degree or examination at any other University.

_____

27 June 2024, Johannesburg

International and national presentations from this study

1. Kathleen Subramoney, Florette K. Treurnicht, Beauty Kalenga, Koketso Letsholo, Nkhensani Shadrack Mtileni. Real-time PCR-based genotyping of SARS-CoV-2 to identify variants of concern, Johannesburg, South Africa, 2021. Oral presentation - ASLM conference, Virtual platform, 15-18 November 2021

2. Kathleen Subramoney, Arshad Ismail, Dr. Florette K. Treurnicht. SARS-CoV-2 lineages among clinical and community surveillance samples submitted to an academic Virology laboratory, Gauteng, South Africa, March 2020-February 2021. Oral presentation - ASLM conference. Virtual platform, 15-18 November 2021

3. Kathleen Subramoney, Dr. Florette K. Treurnicht, Arshad Ismail. SARS-CoV-2 lineages among clinical and community surveillance samples submitted to an academic Virology laboratory, Gauteng, South Africa, March 2020-April 2021. Oral presentation - PathRed Congress. Virtual platform, 19-22 August 2021

4. Kathleen Subramoney, Nkhensani Mtileni, Ashlyn Davis, Jennifer Giandhari, Yeshnee Naidoo, Yajna Ramphal, Sureshnee Pillay, Upasana Ramphal, Houriiyah Tegally, Eduan Wilkinson, Andiswa Simane, Bhaveshan Reddy, Bonolo Mashishi, Nonhlanhla Mbenenge, Tulio de Oliveira, Zinhle Makatini, Burtram C. Fielding, Florette K. Treurnicht. Impact of intra-host immune adaptations on the evolution of SARS-CoV-2 spike protein among individuals with SARS-CoV-2 infections in South Africa, 2020 to 2022. Poster presentation - Faculty of Health Science Research Day 2022. Gauteng, South Africa, 15 September 2022

5. Kathleen Subramoney, Nkhensani Mtileni, Ashlyn Davis, Jennifer Giandhari, Yeshnee Naidoo, Yajna Ramphal, Sureshnee Pillay, Upasana Ramphal, Houriiyah Tegally, Eduan Wilkinson, Andiswa Simane, Bhaveshan Reddy, Bonolo Mashishi, Nonhlanhla Mbenenge, Tulio de Oliveira, Zinhle Makatini, Burtram C. Fielding, Florette K. Treurnicht. Impact of intra-host immune adaptations on the evolution of SARS-CoV-2 spike protein among individuals with SARS-CoV-2 infections in South Africa, 2020 to

2022. Poster presentation - OPTIONS XI. Belfast, United Kingdom, 26-29 September 2022

6. Kathleen Subramoney, Nkhensani Mtileni, Ashlyn Davis, Jennifer Giandhari, Yeshnee Naidoo, Yajna Ramphal, Sureshnee Pillay, Upasana Ramphal, Houriiyah Tegally, Eduan Wilkinson, Andiswa Simane, Bhaveshan Reddy, Bonolo Mashishi, Nonhlanhla Mbenenge, Tulio de Oliveira, Zinhle Makatini, Burtram C. Fielding, Florette K. Treurnicht. Impact of intra-host immune adaptations on the evolution of SARS-CoV-2 spike protein among individuals with SARS-CoV-2 infections in South Africa, 2020 to 2022. Oral presentation – PathRed 2023. Gauteng, South Africa, 31 August to 3 September 2023.

7. Kathleen Subramoney, Nkhensani Mtileni, Jennifer Giandhari, Yeshnee Naidoo, Yajna Ramphal, Sureshnee Pillay, Upasana Ramphal, Akhil Maharaj, Derek Tshiabuila, Houriiyah Tegally, Eduan Wilkinson, Tulio de Oliveira, Burtram C. Fielding and Florette K. Treurnicht. Molecular Epidemiology of SARS-CoV-2 during Five COVID-19 Waves and the Significance of Low-Frequency Lineages. Poster presentation – ESWI. Valencia, Spain, 17-20 September 2023. I received a young scientist scholarship from ESWI to attend the conference.

Publications directly contributing to this thesis and role of the student in these studies.

As supervised researcher and medical scientist in the Virology laboratory of the National Health Laboratory Service at Charlotte Maxeke Johannesburg Academic Hospital from 2020 to 2022, I was leading the next-generation sequencing of SARS-CoV-2 viruses in addition to my responsibilities for SARS-CoV-2 diagnostics and quality management in the laboratory. To ensure rapid response when new waves of infections started, some sequence data was generated by collaborators of the Network for National Genomics Surveillance of South Africa. I conducted genotyping assays, sequence and phylogenetic analyses for the publications. I wrote the first drafts of all manuscripts included in this thesis and addressed co-authors and reviewers' comments during the manuscripts' writing and peer review processes.

1. **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., Makatini, Z., Fielding B. C., Treurnicht, F. K., Network for Genomics Surveillance in South Africa (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), 3676-3684.

2. **Subramoney, K**., Mtileni, N., Giandhari, J., Naidoo, Y., Ramphal, Y., Pillay, S., Ramphal, U., Maharaj, A., Tshiabuila, D., Tegally, H., Wilkinson, E., de Oliveira, T., Fielding, B. C., & Treurnicht, F. K. (2023). Molecular Epidemiology of SARS-CoV-2 during Five COVID-19 Waves and the Significance of Low-Frequency Lineages. *Viruses*, *15*(5), 1194.

3. **Subramoney, K**., Mtileni, N., Davis, A., Giandhari, J., Tegally, H., Wilkinson E., Naidoo, Y., Ramphal, Y., Pillay, S., Ramphal, U., Simane, A., Reddy, B., Mashishi, B., Mbeneng, N., de Oliveira, T., Fielding, B. C., Treurnicht, F. K. (2023). SARS-CoV-2 spike protein diversity at an intra-host level, among SARS-CoV-2 infected individuals in South Africa, 2020 to 2022. *Plos One*, *18*: e0286373.

UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG

R14/49 Mrs Kathleen Subramoney

HUMAN RESEARCH ETHICS COMMITTEE (MEDICAL)

CLEARANCE CERTIFICATE NO. M210119

NAME: (Principal Investigator)	Mrs Kathleen Subramoney
DEPARTMENT:	Department of Virology National Health Laboratory Services, Charlotte Maxeke Johannesburg Academic Hospital
PROJECT TITLE:	Molecular epidemiology of severe acute respiratory syndrome coronavirus 2 whole genome and its spike glycoprotein, South Africa, 2020
DATE CONSIDERED:	29/01/2021
DECISION:	Approved unconditionally
CONDITIONS:	
SUPERVISOR:	Prof B. Fielding and Dr F.K Treurnicht
APPROVED BY:	CB Penny Dr CB Penny, Chairperson, HREC (Medical)
DATE OF APPROVAL:	09/03/2021

This clearance certificate is valid for 5 years from date of approval. Extension may be applied for.

DECLARATION OF INVESTIGATORS

To be completed in duplicate and **ONE COPY** returned to the Research Office Secretary on the Third Floor, Faculty of Health Sciences, Phillip Tobias Building, 29 Princess of Wales Terrace, Parktown, 2193, University of the Witwatersrand. I/we fully understand the conditions under which I am/we are authorized to carry out the above-mentioned research and I/we undertake to ensure compliance with these conditions.Should any departure be contemplated, from the research protocol as approved, I/we undertake to resubmit the application to the Committee. I **agree to submit a yearly progress report.** The date for annual re-certification will be one year after the date of convened meeting where the study was initially reviewed. In this case, the study was initially reviewed in **January** and will therefore be due in the month of **January** each year. Unreported changes to the application may invalidate the clearance given by the HREC (Medical).

Principal Investigator Signature — Date: 10/03/2021

PLEASE QUOTE THE PROTOCOL NUMBER IN ALL ENQUIRIES

UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG

HUMAN RESEARCH ETHICS COMMITTEE (MEDICAL)

2022/02/09

Ms K Subramoney
School of Pathology
Department of Virology
National Health Laboratory Services
Charlotte Maxeke Johannesburg Academic Hospital

Sent by e-mail to: Kathleen.Subramoney@wits.ac.za

Dear Ms Subramoney

Re: **Protocol Ref No:** M210119
Protocol Title: *Molecular epidemiology of severe acute respiratory syndrome coronavirus 2 whole genome and its spike glycoprotein, South Africa, 2020*
Principal Investigators: Ms K Subramoney

I refer to your letter of 2022/01/04.

I confirm that we have noted and approve of your proposal to:

1. extend the data collection window to include records up to and including 2022/12/31 and to conclude the study in 2024
2. genotyping to identify individuals who exhibit diversity, *i.e.* who carry both the wild type and the mutation (Aim 2)
3. design an S protein mosaic based on the study sample sequence data and design a vaccine construct (immunogen) based on this mosaic sequence (Aim 4)
4. seek AREC clearance to use a mouse model to test the immunogen

Thank you for keeping us informed.

Yours Sincerely

.....

Mr I Burns
For the Human Research Ethics Committee (Medical)

.....

Dr CB Penny, Chairperson, Human Research Ethics Committee (Medical)

ABSTRACT

The SARS-CoV-2 global pandemic has been fueled by several variants of concern (VOC) that have gained more efficient transmission or immune evasion properties over time. To better understand the diversity and evolutionary characteristics of SARS-CoV-2 lineages in South Africa we described the analysed the SARS-CoV-2 lineages and VOCs circulating during 2020 to 2022, as well the impact of the S protein and its potential to act as a candidate vaccine.

The first objective of this study was to rapidly identify emerging VOCs based on key SARS-CoV-2 S protein mutations. The second objective was to describe the impact of intra-host immune adaptations on the evolution of SARS-CoV-2 S protein genes among individuals with SARS-CoV-2 infections. Thirdly, by timing the emergence SARS-CoV-2 dominant variants we aimed to unravel the significance and abundance of low-frequency lineages that emerged during five COVID-19 waves in South Africa. The final objective was to assess if accounting for diversity among SARS-CoV-2 S protein's improved predicted epitope coverage of a derived immunogen.

Single nucleotide polymorphism (SNP) PCR-based genotyping assays targeting specific mutations were used to detect VOCs that circulated in 2021. The allele frequencies (AF) as determined by SNP PCR analysis and variant calling from FASTQ reads using galaxy.eu were performed to describe intra-host SARS-CoV-2 S protein variants. Whole genome sequencing was performed to identify and analyse SARS-CoV-2 strains circulating in South Africa from 2020 to 2022 and detect low-frequency lineages. Mosaic vaccine suite tools were used to design an optimal S protein construct from sequences generated in this study. The construct was further tested for antigenicity, toxicity, N- and O-linked glycosylation sites and CTL predictions.

A combination of P681R and L452R SNPs were detected in 73.6% (538/731) of the samples classified as Delta, while N501Y and del69/70 SNPs were detected in 3.6% (26/731) of samples classified as Alpha. The detection of the del69/70 and K417N coupled with SGTF is efficient to exclude Alpha and Beta variants and rapidly detect Omicron BA.1. SNP assays detected 5.3% of cases with Delta that displayed heterogeneity at delY144, E484Q, N501Y and P681H. However, heterogeneity was confirmed by sequencing only for the E484Q and

delY144 mutations. Variant calling from FASTQ reads identified intra-host diversity in the S protein among 9% of cases that were infected with Beta, Delta, Omicron BA.1, BA.2.15, and BA.4 lineages. Heterogeneity was primarily identified at positions 19 (1.4%) with T19IR 371 (92.3%) with S371FP, and 484 (1.9%) with E484AK, E484AQ and E484KQ.

In 2020, 24 lineages were detected, with B.1 (3%; 8/278), B.1.1 (16%; 45/278), B.1.1.348 (3%; 8/278), B.1.1.52 (5%; 13/278), C.1 (13%; 37/278) and C.2 (2%; 6/278) circulating during the first wave. Beta dominating the second wave of infection in 2020. B.1 and B.1.1 continued to circulate at low frequencies in 2021 and B.1.1 re-emerged in 2022. Beta was outcompeted by Delta in 2021, which was thereafter outcompeted by Omicron sub-lineages during the 4th and 5th waves in 2022. Several significant mutations (del69-70, delY144, E484K, N501Y and D614G) identified in VOCs were also detected in low-frequency lineages. During the 5 waves of infection, B.1 and C.1/ C.2 lineages co-circulated with a dominant VOC.

Following our findings of co-circulation of VOCs and other lineages and evidence of quasispecies we investigated if accounting for diversity of SARS-CoV-2 strains would render an improved S immunogen. The optimal mosaic S protein generated had predicted CTL epitope coverage of ~95% to 98% and was classified as an antigen based on a prediction score of 0.47. Reverse translation was used to generate the novel S gene for the expression construct SC2M2. The NTD and RBD regions were non-toxic, and the derived novel S protein comprised 10 additional N-linked glycosylation sites and 4 O-linked glycosylation sites when compared to the Wuhan Hu-1 strain.

Our study findings have shown that (i) rapid detection of emerging VOCs was possible using SNP genotyping assays, and can be used by low to middle income countries to detect Alpha, Beta, Delta and Omicron BA.1; (ii) heterogeneity within the S protein encourages escape from neutralising antibodies and the evolution of SARS-CoV-2, which may contribute to the ongoing emergence of new variants associated with continued outbreaks globally; (iii) low frequency lineages that share mutations with VOCs could lead to convergence and recombination events that result in the next novel lineages or variants that may further increase transmissibility, infectivity and escape immunity; and lastly (iv) the novel S expression construct designed, based on previous and currently circulating VOCs and lineages, could potentially be used to develop improved SARS-CoV-2 vaccines.

ACKNOWLEDGEMENTS

Thank you to my husband, Joash, for continuously encouraging and motivating me throughout my PhD and giving me the time to write.

To my son, Matthew, for sometimes telling me to do my work, and more often distracting me from writing. It would not have been as exciting and challenging without you.

To Nonhlanhla for always supporting me and always being there for me spiritually and emotionally.

Thank you to my supervisors: Florette, for always pushing me to publish manuscripts and for encouragement when things did not go as planned, without your support I would not be where I am today. Burtram, for the support and assistance during writing and always having kind words of motivation.

Most importantly thank you to my Lord and Saviour Jesus Christ for giving me the strength and guidance to get through the challenges faced during my studies.

TABLE OF CONTENTS

1.5. SARS-CoV-2 spike gene 31
1.5.1. Diversity of the SARS-CoV-2 spike glycoprotein 32
1.6. Immune responses to SARS-CoV-2 36
1.7. Justification 37
1.8. Aim and objectives 38
CHAPTER 2 39
Methodology 39
2.1. Study setting and population 39
2.2. Study samples and laboratory data 39
2.3. Sample size and selection 40
2.4. Laboratory investigations 40
2.4.1. SARS-CoV-2 diagnosis 40
2.4.2. SNP-based VOC identification and SARS-CoV-2 genome sequencing 41
a) Nucleic acid template preparation 41
b) PCR-based genotyping to detect single nucleotide polymorphisms (SNPs) associated with spike protein mutations 41
2.4.3. SARS-CoV-2 whole genome sequencing 42
2.4.4. Data analysis 44
2.4.5. Design and synthesis of a mosaic S immunogen 44
2.4.6. Epitope coverage predictions across the S glycoprotein for a candidate mosaic immunogen compared to licensed vaccines 45
2.4.7. Assessing the antigenicity and toxicity of the S immunogen 45
2.4.8. Glycosylation predictions 46
2.4.9. Synthesis of S immunogen 46
2.4.10. Cell culture and transfection 46
2.4.11. Polyacrylamide gel electrophoresis (PAGE) 47
2.5. Ethics 47
CHAPTER 3 48
Rapid identification of SARS-CoV-2 VOCs through detection of single nucleotide polymorphisms in the spike protein 48
3.1. Introduction 48
3.2. Materials and methods 49
3.2.1. Detection of VOCs using single nucleotide polymorphism (SNP) genotyping assays .49

3.2.2. Detection of del69/70 and the K417N PCR based genotyping method for rapid detection of Omicron BA.1 50
3.2.3. Genotyping by SARS-CoV-2 genome sequencing 51
3.2.4. Data analysis 51
3.3. Results 52
3.3.1. SARS-CoV-2 diagnosis among the study population 52
3.3.2. Detection of VOCs from June to October 2021 52
3.3.3. Rapid detection and confirmation of Omicron BA.1 emergence in November 2021 57
3.4. Discussion 64
3.5. Conclusion 67
CHAPTER 4 69
SARS-CoV-2 spike protein diversity at an intra-host level, among SARS-CoV-2 infected individuals in South Africa, 2020 to 2022 69
4.1. Introduction 69
4.2. Methods 70
4.2.1. Study population 70
4.2.2. Study samples and sample size 71
4.2.3. SARS-CoV-2 diagnosis 71
4.2.4. Genotyping to detect single nucleotide polymorphisms (SNPs) associated with specific amino acid mutations in the S protein 71
4.2.5. Next-generation sequencing of SARS-CoV-2 strains 71
4.2.6. Variant mutation analysis and identification of heterogeneous infections 71
4.2.7. Data analysis 72
4.3. Results 72
4.3.1. Demographic characteristics of the study population 72
4.3.2. Frequency and characteristics of heterogeneous SNPs across the S protein 75
4.3.3. Sequence-based analysis of heterogeneity in S protein 78
4.3.4. Mutations at dominant heterogeneous SNPs positions compared to global data 82
4.4. Discussion 86
4.5. Conclusion 91
CHAPTER 5 92
Molecular Epidemiology of SARS-CoV-2 during Five COVID-19 Waves and the Significance of Low-Frequency Lineages 92
5.1. Introduction 92
5.2. Materials and Methods 94
5.2.1. Study population and samples 94
5.2.2. Next-generation sequencing of SARS-CoV-2 strains 94

5.2.3. Data analysis 94
5.3. Results 94
5.3.1. Study population demographics 94
5.3.2. Epidemiology of SARS-CoV-2 lineages and VOCs, 2020 to 2022 97
5.3.3. Phylogenetic description of VOCs overtime 99
5.3.4. Genome wide diversity of low-frequency lineages 101
5.4. Discussion 108
5.5. Conclusions 113
CHAPTER 6 114
Utilising SARS-CoV-2 sequence diversity to design a potential spike (S) immunogen 114
6.1. Introduction 114
6.2. Methods 116
6.2.1. Sequence data 116
6.3. Results 116
6.3.1. Diversity of the S protein sequences 116
6.3.2. Generation of a mosaic S peptide cocktail and identifying the representative three best natural S sequence matches 119
6.3.3. Codon optimised S immunogen 121
6.3.4. Predicted antigenicity and toxicity profile of the SC2M2 immunogen 123
6.3.5. Synthesis of the mosaic S construct 126
6.3.6. Detection of expressed S protein construct 127
6.4. Discussion 128
6.5. Conclusion 130
CHAPTER 7 131
GENERAL DISCUSSION AND CONCLUSIONS 131
REFERENCES 133
Appendix A 153
Appendix B 154
Appendix C 155

LIST OF FIGURES

Figure	Title	Page
Fig.1.1	Structure of the SARS-CoV-2 S protein. (Created with BioRender). Left: the spike protein displaying the location of the S1 and S2 subunits. Right: The 3D structure of the spike protein displays the protein's different regions. HR1 and HR2 represent the heptad repeat 1 and 2 regions. The shaded rectangular block on the right represents the 2D structure of the spike protein and the amino acid change at position 685. ((1), PDB101.rcsb.org)	31
Fig. 1.2	3D illustration of SARS-CoV-2 VOCs circulating from 2020 to 2021, this template showcases the key spike protein mutations on each of the global SARS-CoV-2 variants of concern: B.1.1.7, B.1.351, P.1, B.1.526, B.1.427, B.1.429, B.1.617. (Created with BioRender).	34
Fig. 3.1	SARS-CoV-2 detected in respiratory samples tested from 1 June to November 2021. The red line graph represents the detection rate, and the bar graph represents the number of samples received for SARS-CoV-2 diagnostics during that period.	52
Fig. 3.2	Variants/lineages detected in four districts of the Gauteng Province from June to October 2021 (N=611). The percentages indicate the prevalence of the VOCs detected within a district.	54
Fig. 3.3	Variants/lineages detected in six Eastern Cape districts (N=120), from June to October 2021. The percentages indicate the prevalence of the VOCs detected within a district.	55
Fig. 3.4	Phylogenetic representation of sequences from Virology samples, NHLS Gauteng, from June to October 2021. Sequence data (circular nodes) are plotted against SARS-CoV-2 global data from	56

next clade (https://clades.nextstrain.org). The x-axis represents the number of mutations present. The purple nodes: Delta sub-lineage AY.45 (21J), light blue nodes: Delta sub-lineage AY.38 (21I), turquoise nodes: Delta sub-lineage AY.19 (21A). Figure adapted from clades nextstrain.org (accessed 31 July 2023). Data can be accessed from GISAID Identifier: EPI_SET_230801yz.

Fig. 3.5 The number of variants detected over time from June to October 2021 based on sequence data. 57

Fig. 3.6 Detection rate of SARS-CoV-2 in Gauteng, South Africa, in November 2021. The bar graph represents the number of samples that were received at Virology, NHLS for SARS-CoV-2 diagnostics. The red line graph represents the detection rate (number of positive tests/number of samples received per day). NHLS, National Health Laboratory Service; SARS-CoV-2, severe acute respiratory syndrome coronavirus 2. 59

Fig. 3.7 Frequency of SGTF, del69/70, and K417N mutations among SARS-CoV-2 positive samples collected from November 15 to 25, 2021. SGTF− represents the samples for which SGTF was absent. SGTF+ represents the samples with SGTF. From the genotyping PCR assay: The del69/70 genotype and del69/70 undetermined are samples with and without the deletion, respectively; and the K417, 417N, and K417N undetermined represent samples that had the wild-type, mutation and neither at amino acid position 417, respectively. Genotyping was not performed on positive samples collected on November 20th. 60

Fig. 3.8 Allelic discrimination plot of del69/70 (A) and K417N (B) mutations within the S gene showing florescence values (Rn). Samples within the red oval (along the x-axis) represent homozygotes for either the deletion (A) or the 417N mutation (B). 60

Samples within the green oval (along the y axis) represent those that do not have the deletion or have K417 wild type (homogeneous wild type). Samples lying outside these areas represent a negative result (does not have the mutation or wild type).

Fig. 3.9 Phylogenetic representation of sequences from samples from Virology, NHLS Gauteng. Only sequence data from November 1st to 30th, 2021 from Virology, NHLS are represented in this figure. (A) Virology, NHLS sequence data are plotted against SARS-CoV-2 global data from the next clade (https://clades.nextstrain.org). The X-axis represents the number of mutations present. The orange nodes: Omicron (B.1.1.529/21K) and its sub-lineages including BA.1 (21K) and BA.1.1 (21K), turquoise nodes: Delta sub-lineage AY.45 (21J), light blue nodes: Delta sub-lineage AY.38 (21I), dark blue nodes: Delta sub-lineage AY.19 (21A), purple node: Beta (20H) and gray node: C.1.2 (20D). Figure adapted from clades. nextstrain.org. (B) Representation of all Virology, NHLS sequence data was plotted for the period of November 1–30, 2021. From the 1st to the 14th, the number of positive SARS-CoV-2 cases was few or completely lacking and therefore sequence data is very low or absent. NHLS, National Health Laboratory Service; SARS-CoV-2, severe acute respiratory syndrome coronavirus 2. Data can be accessed from GISAID identifier: EPI_SET_230802ad. 63

Fig. 4.1 SARS-CoV-2 detection rate among respiratory samples tested from June 2020 to May 2022. A) Blue bars represent the number of samples received for SARS-CoV-2 diagnostics and the red line graph represents the detection rate over time. B) Green bars represent the number of SARS-CoV-2 samples that were tested by SNP PCR and/or sequenced and the red line graph represents the detection rate over time. 74

Fig. 4.2. Distribution of Ct-values for each SARS-CoV-2 mutations tested. The x-axis displays the Ct-values from SARS-CoV-2 diagnostic testing compared to the genotyping results for each mutation and/or wild type detected for each SNP assay. Samples that were undetermined for a genotyping assay are associated with the assays inability to detect the mutant and/or wild type. 75

Fig. 4.3 Allelic discrimination plots of mutants, wild type and heterogeneous SNPs (mutant/wild type). Black block indicates the samples that were negative for the mutant and/or wild type for each SNP (<0.50). Red ovals indicate the samples that were heterogeneous (both wild type and mutant present). X-axis represents the allelic distribution of mutants and y-axis represents the allelic distribution of the wild types. 76

Fig. 4.4 Frequency of the heterogeneous SNPs identified from June to October 2021 (N=50). X-axis represents the date during which the heterogeneous SNP was identified. Y-axis: the line graph represents the number of samples that were observed with a particular heterogeneous SNP, and the bar graphs represent the total number of samples (N) in which the SNP was observed on each day. 78

Fig. 4.5 Box and whisker plot showing the mutant allele frequencies (AF) at amino acid positions within the S protein that have more than 1 amino acid change. Circles represent the outliers, line within the box represents the mean AF. Mutations (T19K, P681S D950, D950K, and D950R) that only have a single AF value were excluded from the figure. 79

Fig. 4.6 Frequency of the heterogeneous mutations (A) and VOC/lineages (B) identified from sequence data over time. 4a) X-axis represents 80

the date during which the heterogeneous SNP was identified, and the y-axis represents the percentage of samples that were observed with a particular heterogeneous SNP. 4b) X-axis represents the period during which the VOC/lineages were identified, and the y-axis represents the percentage of samples with a specific VOC/lineage. Omicron BA.1, BA.1.1, BA.1.10, BA.1.13, BA.1.14, BA.1.17, BA.1.17.2, BA.1.18, BA.1.19 and BA.1.21 were grouped as Omicron BA.1.

Fig. 4.7 Prevalence of samples with S371FP over time. The black line graph represents the total number of samples with S371FP heterogeneity each month. The bar graphs represent the SARS-CoV-2 VOCs and lineages observed with the S371FP heterogeneous mutation from 2021 to 2022. 81

Fig. 4.8 Frequency of mutations occurring at aa positions 19, 371, and 484 in the S protein. Custom NextStrain time tree (left) displays the sequence data from this study from June 2020 to May 2022, with specific aa changes at the heterogeneous positions. The phylogenetic tree (middle) represents the global SARS-CoV-2 sequence data from January 2020 to May 2022, with specific aa changes at the heterogeneous positions (https://nextstrain.org/ncov/gisaid/global/all-time, accessed 22 June 2022). Clades displayed in phylogenetic trees represent the following SARS-CoV-2 pangolin lineages or VOCs (https://covariants.org/variants, accessed 22 June 2022): 20A and 20C include B.1 lineages, 20B includes B.1.1 and its sub-lineages; 20D includes B.1.1.1 lineage, Delta 21A, and Delta 21J includes B.1.617.2 and AY lineages; Alpha 20I includes B.1.1.7; Omicron 21M includes B.1.1.529 lineage; Omicron 21K includes BA.1 lineage; Omicron 21L includes BA.2 lineages and branches out to Omicron 22A (BA.4) and Omicron 22B (BA.5). Graphs on the right show the frequency of the aa changes over time 84

(https://nextstrain.org/ncov/gisaid/global, accessed 22 June 2022). Tables show the frequency of wild-type aa and mutations present globally from GISAID (https://www.gisaid.org/epiflu-applications/covsurver-mutations-app/, accessed 22 June 2022).

Fig. 5.1 Prevalence of SARS-CoV-2 lineages over time from 2020 to 2022 by epiweek. The bar graph represents the SAR-CoV-2 lineages and VOCs identified in our study cohort. The black line graph represents the total number of samples that were sequenced during each epiweek. (A) Distribution of SARS-CoV-2 lineages in 2020, (B) Distribution of SARS-CoV-2 lineages in 2021, and (C) Distribution of SARS-CoV-2 lineages in 2022. 98

Fig. 5.2 Phylogenetic analysis of SARS-CoV-2 lineages and VOCs detected in South Africa from 2020 to 2022. (A) Radial phylogenetic tree displays the lineages identified in South Africa. Tree rooted with the Wuhan-Hu1 strain. Our dataset was compared to the global database embedded in NextStrain. (B) Nextclade custom time-tree displaying the increase in the number of mutations observed for each SARS-CoV-2 clade observed as time progressed. 100

Fig. 5.3 SARS-CoV-2 low-frequency lineages identified in this study. The blue-shaded circles represent the samples designated as low-frequency lineages from our study cohort. The horizontal clusters in the background represent the additional lineages and VOCs identified globally. 101

Fig. 5.4 SARS-CoV-2 low-frequency lineages identified over time. (A) The bars represent the SAR-CoV-2 lineages, and the line graph represents the total number of samples that were successfully sequenced during each month from 2020 to 2022. (B) The shaded dots represent the samples designated as low-frequency lineages. 102

The x-axis represents the total number of mutations identified in each lineage across the whole genome. The y-axis represents the date during which the lineages were observed.

Fig 6.1 Workflow described to design and test the spike protein construct 116

Fig. 6.2 Phylogenetic tree displaying our sequence data from the wild type to the Omicron waves, March 2020 to April 2022. The data from this study is represented as coloured circles under its respective clusters compared to global data. The additional lines without circles represent the global sequence data. 117

Fig. 6.3: Entropy plot of amino acid sequence coverage compared to the Wuhan-Hu1 reference strain. Data includes the spike protein of all South African sequences from 2020 to 2022, from our study. The black shaded areas represent the amino acid similarities at an amino acid position. The grey area represents missing sequence data. 118

Fig 6.4 Spike protein sequence data aligned against the Wuhan Hu-1 reference (first row) sequence. The amino acid positions are represented on the horizontal axis and the horizontal axis represents each spike protein sequence. Each colour indicates a different amino acid. White represents gaps present in the sequence data. 118

Fig. 6.5 The unique 9-mer coverage of the spike peptides with mosaic alignments representing 4 possible scenarios for the 3 best natural sequences (214, 316, 545) identified as representative of the mosaic peptide cocktail. Best 1 represents seq 214 or hCoV-19/SouthAfrica/CERI-KRISP-K027930/2021, Best 2 represents seq 316 or hCoV-19/SouthAfrica/CERI-KRISP-K033221/2021, Best 3 represents seq 545 or hCoV-19/SouthAfrica/KRISP- 119

K018784/2021). The dark areas indicate poor CTL epitope coverage associated with the amino acid mutations in the spike protein of SARS-CoV-2. Identical sequences were removed, errors, gaps and non-specific nucleotides were removed.

Fig. 6.6 Epitope coverage of spike peptide mosaic. Graphs a to f display the coverage of potential 9-mer T cell epitopes of different instances of peptide cocktails. The coverage is described as the average number of potential epitopes per natural strain that are matched per natural strain. 120

Fig. 6.7 Graphical display of predicted positional 9-mer epitope coverage for the three best natural sequences that will cover CTL epitopes present in the derived mosaic protein. The matched 9-mers are displayed according to the nucleic acid alignment positions (x-axis). The sequence coverage of the matched 9-mers are displayed in the y-axis. 120

Fig. 6.8 Protein sequence of the optimised S protein mosaic (SC2M2) obtained from the 3 best natural protein sequences. 121

Fig 6.9 Codon usage and coverage of the SC2M2 before and after codon optimisation. 122

Fig. 6.10 Codon optimised SC2M2 nucleotide sequence. 123

Fig. 6.11 N-linked glycosylation site predications. A) The red text represents the N-glycosylated across the SC2M2 construct. B) Displays the summary of theN-glycosylation predications based on a threshold of >0.5. C) Graphical representation of the N-glycosylation sites present in the Wuhan Hu-1 S protein. D) Graphical representation of the N-glycosylation sites present in the SC2M2 construct. 125

Fig. 6.12 Map of the selected optimal spike protein construct (SC2M2) cloned in pcDNA3.1(-) myc-His A. Diagram created by snapgene. 127

Fig. 6.13 SC2M2 base calls (A) and fluorescent intensity (B) of each nucleotide base. 127

Fig. 6.14 SDS-PAGE showing the protein separation of the SC2M2 construct. The seeblue plus 2 ladder was used for the marker on the right of the gel. The black rectangle indicates the spike protein band present in SC2M2 supernatant (well 1 and 2). 128

LIST OF TABLES

Table	Title	Page
Table 1.1	Primary variants of concern identified globally since 2020	33
Table 1.2	Summary of vaccines efficacy against SARS-CoV-2 VOCs	35
Table 2.1	Variants and associated mutations (SNPs)	41
Table 3.1	Interpretation of results for VOCs based on the presence or absence of specific mutations	51
Table 3.2	The VOCs detected from SNP genotyping assays for variant- specific S protein mutations	53
Table 3.3	Prevalence of S gene target failure (SGTF) among patients who tested positive for SARS-CoV-2 (N=175)	58
Table 3.4	Comparison of Ct values from TaqPath™ COVID-19 N gene with del69/70 and K417N genotypes for cases where genotyping could not determine the variant type	61
Table 4.1	Demographics of individuals with SARS-CoV-2 included in this study (N=2982)	73
Table 4.2	Summary of SNP combinations with heterogeneous amino acids (n=50)	77
Table 4.3	Patient status for cases where heterogeneity in SARS-CoV-2 S protein was observed (N=210)	82

Table 4.4	Age groups for cases for which heterogeneity in S protein was identified	82
Table 4.5	Allele frequencies and prevalence of homogeneous mutations identified among all individuals in the study population	85
Table 5.1	Demographic characteristics of successfully sequenced SARS-CoV-2 samples from 30 March 2020 to 13 October 2022 (N=2547)	96
Table 5.2	Mutations detected across the whole genome of B.1 lineage from 2020 to 2021	103
Table 5.3	Mutations detected across the whole genome of B.1.1 lineage from 2020 to 2022	103
Table 5.4	Demographic data for participants from whom B.1 and B.1.1 lineages were detected in 2020 to 2022	104
Table 5.5	Mutations detected across the whole genome of B.1.1.348 lineage in 2020	105
Table 5.6	Mutations detected across the whole genome of B.1.1.52 lineage from 2020	105
Table 5.7	Demographic data for participants from whom B.1.1.438 and B.1.1.52 lineages were identified in 2020 to 2022	105
Table 5.8	Demographic data for participants for whom C.1 and C.1.2 lineages were detected in2020 to 2022	106

Table 5.9	Demographic data for participants for whom C.2 lineages were detected in 2020 to 2022	107
Table 5.10	Mutations detected across the whole genome of C.1 lineage from 2020 to 2021	107
Table 5.11	Mutations detected across the whole genome of C.1.2 lineage from 2020 to 2021	107
Table 5.12	Mutations detected across the whole genome of C.2 lineage from 2020	108
Table 6.1	Toxic peptides detected from the optimised S protein immunogen	124
Table 6.2	O-linked glycosylation predications for SC2M2 S protein construct	126
Table 6.3	O-linked glycosylation predications for Wuhan Hu-1	126
Table 6.4	Quality control checks performed on the SC2M2 construct	126

LIST OF ABBREVIATIONS

%	percent
<	less than
>	greater than
≥	greater than equal to
°C	Degrees Celsius
µl	microlitres
3CLpro	3C-like protease
ACE2	angiotensin-converting enzyme 2
Amps	ampoules
CAI	codon adaptive index
CDC	Centers for Disease Control and Prevention
CDW	Central Data Warehouse
CMJAH	Charlotte Maxeke Johannesburg Academic Hospital
CO_2	Carbon dioxide
COVID-19	coronavirus disease
CST	community screen and test
Ct	cycle-threshold
CTLs	cytotoxic T cells
DMEM	Dulbecco's Modified Eagle Medium
DNA	Deoxyribonucleic acid
E	envelope
EDTA	Ethylene diamine tetra-acetic acid
FBS	foetal bovine serum
GISAID	Global Influenza Sequence Database
hCoV	Human coronaviruses
HEK	Human Embryonic Kidney
HEPES	4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid
HIV	human immunodeficiency virus infection
HLA	Human leukocyte antigens
iSNV	intra-host single nucleotide variants
kDa	kilodalton
KRISP	KwaZulu-Natal Research Innovation and Sequencing Platform
M	membrane
mAbs	Monoclonal antibodies
MEM	Modified Eagle Medium
MERS-CoV	Middle East respiratory syndrome coronavirus
MHC	Major Histocompatibility Complex
MIS	multisystem inflammatory syndrome
ml	millilitres
N	nucleocapsid
NGS	next generation sequencing

NHLS	National Health Laboratory Service
NICD	National Institute for Communicable Diseases
NP	nasal and nasopharyngeal swabs
nsp	non-structural protein
NTD	N-terminal domain
OP	oropharyngeal swabs
ORF	open-reading frame
PBS	phosphate buffered saline
PCR	polymerase chain reaction
PLpro	papain-like protease
QC	quality control
RBD	receptor binding domain
RdRp	RNA-dependent RNA-polymerase
RNA	Ribonucleic acid
rRT-PCR	real-time reverse transcription-based polymerase chain reaction
S protein	Spike glycoprotein
SA	South Africa
SARS-CoV-2	severe acute respiratory syndrome coronavirus 2
SDS-PAGE	Sodium dodecyl-sulfate polyacrylamide gel electrophoresis
SGTF	S gene target failure
SNP	single nucleotide polymorphism
SVM	support vector machine
TBE	Tris/Borate/EDTA
TNA	Total nucleic acids
US	United States
UV	ultraviolet
VOC	variants of concern
WHO	World Health Organization

CHAPTER 1

Literature review

1.1. History of human coronaviruses

Human coronaviruses (hCoVs) including hCoV-NL63, -HKU1, -OC43, and -229E, have been circulating globally, and based on molecular clock inferences, for at least a few hundred years. These viruses are responsible for acute upper respiratory tract infections of a mild nature. The first cases of human CoV-229E and hCoV-OC43 were identified in the 1960s, while hCoV-NL63 and hCoV-HKU1 were detected in 2004 and 2005, respectively (2). hCoVs infections are consistently detected globally and infections typically peak during winter and spring (3). These viruses cause mild symptoms, but in rare instances may cause severe disease among infants, adults >65 years of age, and immunocompromised individuals due to naïve or weakened immune systems (2,4,5).

Human CoVs that have led to outbreaks of global concern include the severe acute respiratory syndrome coronavirus (SARS-CoV) and Middle East respiratory syndrome coronavirus (MERS-CoV). SARS-CoV is the causative pathogen of an atypical pneumonia which was initially reported in China in late 2002 (6–8). The virus eventually spread to 29 countries, with close to 8000 laboratory-confirmed cases, and a case-mortality rate of about 10% (3). MERS-CoV, on the other hand, was identified in 2012 in Saudi Arabia as the cause of severe acute respiratory illness associated with high mortality rates (6–8). MERS-CoV was predominant in the Middle East and North African countries and transmitted to other parts of the world through travel (3). The World Health Organization (WHO) reported 2134 laboratory-confirmed MERS-CoV cases across 27 countries, of which ~35% lead to deaths (9). For MERS-CoV, which continues to circulate in the Middle East, seasonality was observed in spring from March to May based on observations made from 2012 to 2014 (10).

1.2. The epidemiology of SAR-CoV-2

In December 2019, the first case of SARS-CoV-2 infection and its associated respiratory syndrome called coronavirus disease (COVID-19) was identified in Wuhan, China. Ever since the virus has rapidly disseminated globally. SARS-CoV-2 was declared the “sixth public health emergency of international concern” in 2020 by the WHO. As of 20 March 2023, ~761 million

SARS-CoV-2 cases were reported globally, of which ~9.5 million cases and 175 313 deaths were observed in Africa. South Africa was accountable for 0.5% (4 068 224) of global cases, with a 1.5 % (102 595/6 873 477) mortality rate (11).

Globally, people of all ages, races, and gender are susceptible to SARS-CoV-2 infection, with the highest rate of infection observed among 25-64 years and the highest mortality rates observed among adults >65 years, for the duration of the pandemic (12–15). The virus is transmitted from an infected person through contact with nasal, oral, and eye mucosal secretions and through droplet inhalation from sneezing or coughing (16,17). The incubation period for SARS-CoV-2 is 3 to 14 days, depending on the person's immunological conditions (18). Patients with COVID-19 commonly present with fever, dry cough, and fatigue (16,19,20). Severe symptoms include shortness of breath, loss of appetite, confusion, persistent pain or pressure in the chest, and high temperature (>38 °C) (19). Critical disease features include chronic obstructive pulmonary disease, multi-organ failure, septic shock, and respiratory failure (20).

1.3. Co-morbidities and clinical syndromes associated with SARS-CoV-2

The most common co-morbidities identified among patients with COVID-19 that have resulted in increased mortality include hypertension, obesity, diabetes mellitus, and chronic kidney disease (21,22). Patients with cancer are one of the highest at-risk groups due to weakened immune systems as a result of tumor growth and anti-cancer treatment (23). About 50% of French kidney transplant patients had sustained shedding of SARS-CoV-2 up to 30 days post-infection and some patients up to 60 days post-infection compared to immunocompetent patients who shed for a median of 18 days to 25 days (24). In South Africa, human immunodeficiency virus infection (HIV) and acquired immunodeficiency syndrome (AIDS) are very important comorbidities that may contribute to prolonged SARS-CoV-2 shedding. Increased diversity was observed in spike gene sequences of SARS-CoV-2 strains from immunocompromised patients with prolonged viral shedding or resurgence following short periods of suppression (25–27).

After recovery from COVID-19, many people experience post-COVID conditions (28). This comprises of new, returning, or ongoing health conditions, which can persist for up to 4 weeks after initial infection. Post-COVID conditions were also reported in people that initially had

asymptomatic SARS-CoV-2 infections. As described by the United States (US) Centers for Disease Control and Prevention (CDC), "Long COVID" or "Long-Haul COVID" (post-COVID) conditions occur 4-6 weeks or months after initial illness and symptoms include fatigue, difficulty concentrating, headaches, loss of smell/taste, dizziness when standing, heart palpitations, joint or muscle pain and depression or anxiety (29), to name a few. Severe outcomes are also experienced with post-COVID conditions which include multi-organ disease that affects the heart, lungs, kidneys, skin, and brain, as well as multisystem inflammatory syndrome (MIS) and autoimmune conditions (29). MIS has been reported among children and adolescents in Europe and the US. Although extremely rare, children with MIS predominantly present with shock, cardiac dysfunction, abdominal pain, and increased inflammatory markers (30). The CDC has also reported several adults with MIS which included cardiovascular, gastrointestinal, dermatologic, and neurological symptoms (29,30).

1.4. SARS-CoV-2 vaccines approved during the initial rollout in South Africa

Since the rollout of the SARS-CoV-2 vaccines in South Africa, 6 vaccines have been approved since December 2022. However, only Comirnaty (Pfizer, Inc., and BioNTech), Jcovden/ Ad26.S (Janssen Pharmaceuticals Companies of Johnson & Johnson, and Covishield (Oxford/ AstraZeneca) were placed in use in the country (https://covid19.trackvaccines.org/country/south-africa/, accessed 01 June 2024). Presently licensed SARS-CoV-2 vaccines are mainly based on the viral spike (S) glycoprotein as vaccine immunogen in the following vaccine design strategies: (i) Comirnaty (Pfizer, Inc., and BioNTech) is a nucleoside-modified RNA vaccine that expresses the complete S glycoprotein (31). In May 2021, Comirnaty was licensed for use in South Africa and vaccinations were in progress. (ii) Jcovden/ Ad26.S (Adenoviral vector-based: JNJ-78436735) is a recombinant vector vaccine that utilises adenovirus serotype 26 as a vehicle to deliver the SARS-CoV-2 S glycoprotein (32,33). In South Africa, most healthcare workers received the Jcovden/ Ad26.S vaccine as part of an extended phase 3 clinical trial. The adenovirus vector was modified to prevent its replication and adenoviral disease in humans. Vaccine availability is still challenging for the National Department of Health with 2 million doses of the Jcovden/ Ad26.S vaccine being discarded (14 June 2021) due to failures in good manufacturing processes. About 8 million (~14%) South Africans had been vaccinated (06 August 2021) with a vision to double this number by 20 August 2021 with an ultimate target to immunise 60% of the 58.6 million

people. Covishield is a non-replicating viral vector designed from chimpanzee adenovirus with similar genetic characteristics to the SARS-CoV-2 S protein (34). The vaccine did not prove effective against mild to moderate illness caused by Beta (10.4% vaccine effectiveness) in the South African population (34). By August 2022, approximately 37 million South Africans wre vaccinated against COVID-19, majority of whom where immunised with Jcovden and Comirnaty (35).

1.5. SARS-CoV-2 spike gene

The S protein is the main viral surface glycoprotein (36,37) encoded by the S gene open reading frame of ~3000 base pairs. The S protein is a class I protein that enables the successful transmission of SARS-CoV-2 between humans through furin-mediated cleavage (38). Cleavage of the S1 and S2 fusion peptides, by the furin-like proteases, results in trimer conformation of the protein structure which allows infectivity by progeny virions (39). The 1273 amino acids long S protein (Fig. 1.1) comprises S1 and S2 subunits. The furin cleavage site at the S1-S2 subunits allow for attachment to the human angiotensin converting enzyme-2 (ACE-2) receptor with its receptor binding domain (RBD) (40,41). The S2 subunit fuses with the host cell membrane during the entry (41).

The S glycoprotein is the primary target for inducing neutralising antibodies and specifically the RBD and N-terminal domain (NTD) (37). It has high mutation rates and mutations primarily occur in the RBD and NTD (42–45). During the first wave of the pandemic, mostly in the B.1 lineage: 21 substitutions were observed in the RBD which included novel mutations N282D, P491H, G496W, P499A, S514C, Q675H, and N1098D which are involved in host changes, antigenic drift, binding to the hosts surface receptor and antibody recognition; and in the NTD the delY144 was a novel deletion detected (44).

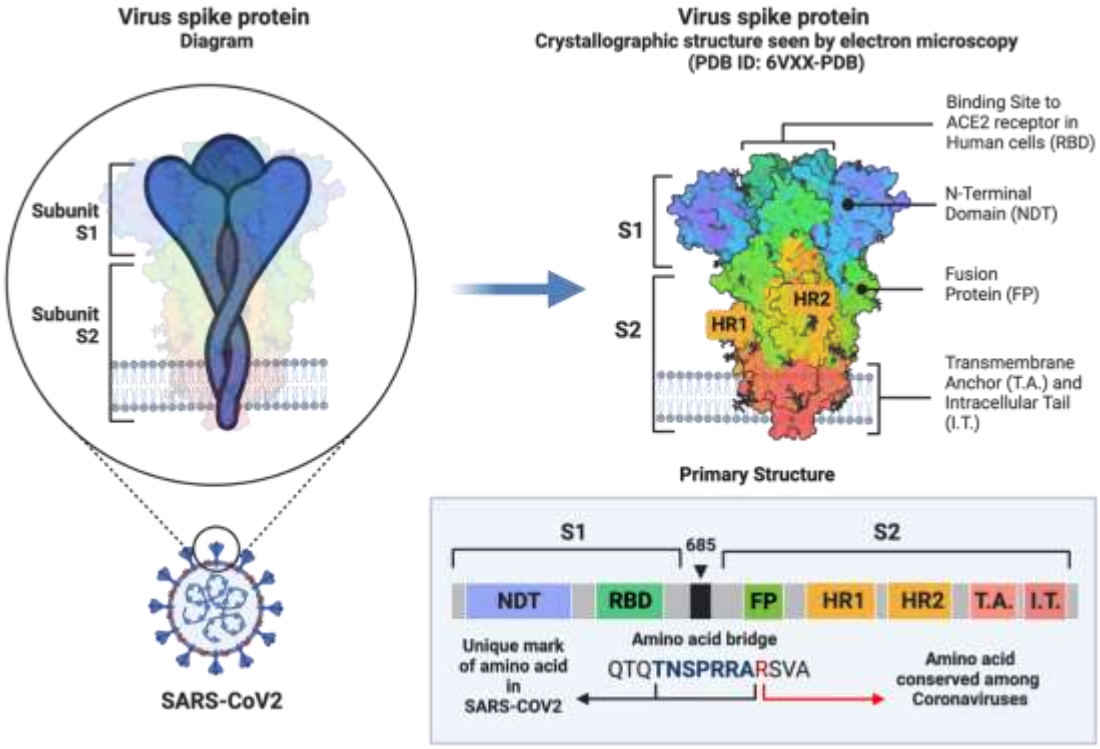

Fig 1.1: Structure of the SARS-CoV-2 S protein (Created with BioRender). Left: the spike protein displaying the location of the S1 and S2 subunits. Right: The 3D structure of the spike protein displays the protein's different regions. HR1 and HR2 represent the heptad repeat 1 and 2 regions. The shaded rectangular block on the right represents the 2D structure of the spike protein and the amino acid change at position 685. ((1), PDB101.rcsb.org)

1.5.1. Diversity of the SARS-CoV-2 spike glycoprotein

a) Emerging mutations with the emergence of VOCs

As SARS-CoV-2 emerged over time, it was characterised by gradual evolutionary adaptation with a single notable mutation (D614G) in the S protein of B.1 lineage, providing a 20% growth advantage to the virus in comparison to previous lineages (46). Tegally *et al* (2021) identified at least 16 novel variants among national genomic data from South Africa, representing the first wave in 2020 of which 3 SARS-CoV-2 lineages (B.1.1.54, B.1.1.56, and C1) dominated and were local variants of concern (VOC) (47). VOCs demonstrate increased transmissibility, cause an increase in disease severity, have reduced neutralising antibody against previous infection or vaccination, reduced vaccine effectiveness or challenges during diagnostic detection, all of which warrant further public health response (48).

These 3 lineages accounted for 42% of the total infections across South Africa, but this dataset was dominated by genomic data from KwaZulu-Natal. The C1 lineage has 16 additional

mutations compared to the Wuhan-Hu1 strain, including the D614G in the S protein. C.1 became the dominating lineage in Gauteng, South Africa, from July 2020 (47). This was the most common lineage detected and remained prevalent until the end of August 2020. In December 2020, the Beta VOC was identified for the first time with primary S protein mutations including K417N, E484K, and N501Y (Fig.1.2). The variant led to a substantial rise in the rates of transmission by 40-70%, across the South African population. In South Africa, rapid evolutionary changes were evident as new VOCs emerged including Alpha, Beta, Delta, and Omicron lineages with several significant mutations present.

VOCs produced large numbers of non-synonymous mutations in the S protein, and additional phenotypic properties including increased transmissibility, antigenicity, and infectivity (49,50). Since the beginning of the pandemic, Alpha, Beta, Gamma, Delta, Omicron and their sub-lineages (Table 1.1) were recognised as public health concerns and were closely monitored due to their changing transmissibility and immune escape properties (40). Mutations in the S proteins NTD and RBD resulted in the emergence of several VOCs (Table 1.1 and Fig.1.2). Alpha was initially identified in the United Kingdom, Gamma was identified in Brazil, while Beta, Delta and Omicron initially detected in South Africa and eventually spread globally (40,42). The N501Y mutation in the RBD was reported to be present in all variants, except Delta (51). The Alpha variant is highly transmissible and is characterised by the del69-70 and del144 in the NTD, N501Y in RBD, and P681H near the furin cleavage site of the S glycoprotein (52). Beta, on the other hand, has three primary signature amino acid substitutions namely K417N, E484K, and N501Y, in the RBD. These mutations are associated with escape from neutralising antibody responses directed at the RBD; the rest of the neutralising antibody responses are directed toward the NTD (43,53). Delta, and its sub-lineages, are recognized by the T478K mutations in the S protein, which is reported to result in enhanced infectivity and transmissibility (54–56). Alpha and Delta comprise of mutations that promotes binding of the virus to the hosts ACE-2 receptor and provides optimal furin cleavage which improves transmissibility by up 65% (40). Omicron and its sub-lineages, initially identified in Southern Africa, is known for its ability to modify entry phenotype in conjunction with immune escape mutations (49,57). Globally, all VOCs evolved from the previous lineages, but since the emergence of the Omicron (B.1.1.529 lineage), all consecutive waves have been a result of Omicron sub-lineages including BA.1 to BA.5, BQ.1, and several other Omicron recombinants. Important mutations that emerged with the introduction of Omicron BA.4 and BA.5 includes

del69/70, L452R, F486V, and Q493R in the S protein. Mutations at positions 452, 486 and 493 may encourage viral attachment to the human ACE-2 receptor and reduction in antibody binding sites. Additional mutations detected in Omicron and its sub-lineages includes K856R, A2710T, L2084I, P3395H, T3255I, 13758V, del2083, del3674-3676 in ORF1b; P314L, I1566V, del27-29 in ORF1b2; P10S in ORF9b; T91I in the E protein; and D3G, Q19E and A63T in the M protein (58). There is still uncertainty regarding the function of these mutations and the role they play in neutralising antibody activity (58).

Table 1.1: Primary variants of concern identified globally since 2020

WHO classification	Pango lineage (ancestral lineage)	Nextstrain clade
Alpha	B.1.1.7	20I (V1)
Beta	B.1.351	20H (V2)
Gamma	P.1	20J (V3)
Epsilon	B.1.427/ B.1.429	20C
Delta	B.1.617.2	21A
Lota	B.1.526	20C
Omicron	B.1.1.529	21M

Accessed 02/05/2023: https://www.cdc.gov/coronavirus/2019-ncov/variants/variant-classifications.html 1633452601089 and https://covariants.org/variants

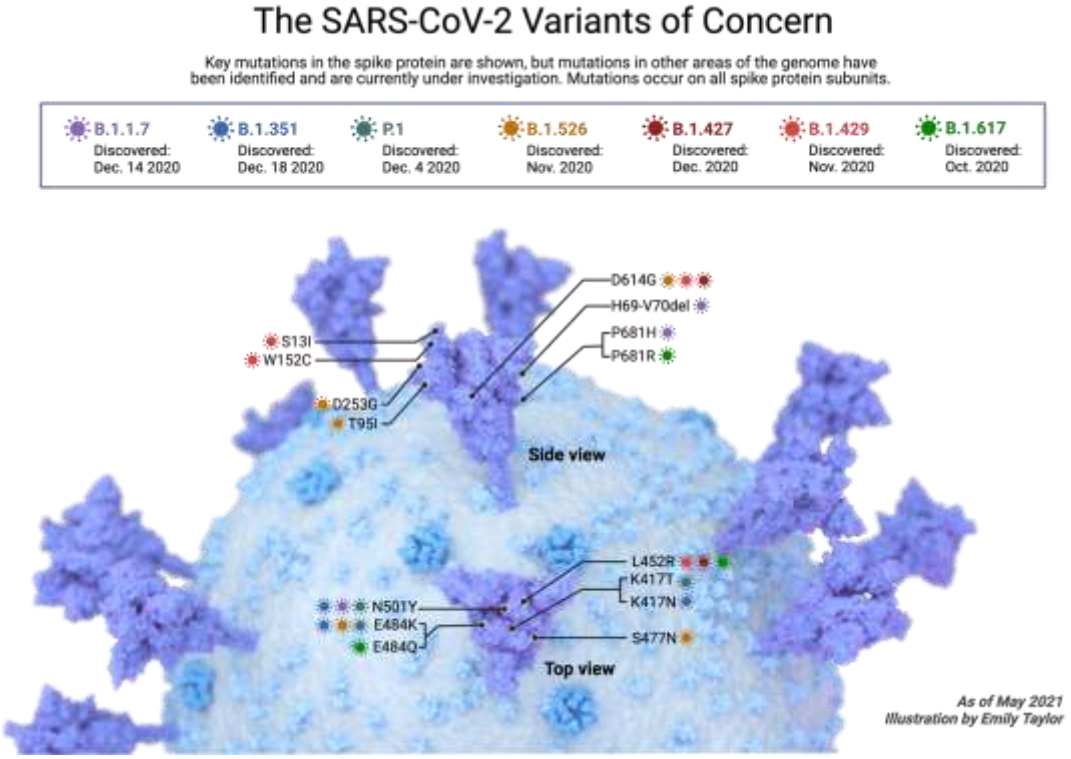

Figure 1.2: 3D illustration of SARS-CoV-2 VOCs circulating from 2020 to 2021, this template showcases the key spike protein mutations on each of the global SARS-CoV-2 variants of concern: B.1.1.7, B.1.351, P.1, B.1.526, B.1.427, B.1.429, B.1.617. (Created with BioRender)

Since January 2021, the SARS-CoV-2 Beta variant, originating in South Africa, was initially transmitted to several other African countries including Botswana, Gambia, and Zambia, followed by increasing prevalence in Europe and the US (59). Compared to other variants of concern, the Delta variant does not have mutations at positions 484 and 501 of the spike protein. The Delta variant that eventually emerged in India was first detected in South Africa in early 2021, but the extent of community spread is still unclear. In South Africa, variants associated with mutations that are currently presenting at lower frequency should not be ignored as they may seed future variants of concern due to continued vaccine-driven or natural immune selection pressures.

b) Impact of spike diversity and VOCs on vaccine effectiveness

The emergence of the Beta variant in South Africa resulted in the reduced efficacy of the ChAdOx1 nCoV-19 vaccine (Oxford/ AstraZeneca) in a multi-country clinical trial in which South Africa participated (Table 1.2). This reduced vaccine efficacy resulted in the suspension of the planned rollout of this vaccine in South Africa. At that time, the Novavax NVX-CoV2373 vaccine efficacy against the Beta variant was reported to be 43% (Table 1.2). Moreover, the Ad26.COV2-S vaccine had a greater efficacy (Table 1.2) against Beta - which showed resistance to neutralising antibodies - in preventing severe to critical disease. Therefore, this vaccine was still acceptable and was used in South Africa despite the dominance of the Beta variant. The reduction in vaccine effectiveness was less pronounced for the Alpha and Delta variants, with Delta dominating the 3rd wave of infections in South Africa (60,61). When testing Ad26.COV2.S and BNT162b2 vaccines against Omicron lineages among the South African population, the vaccine effectiveness was up to ~70% for 2 doses of either vaccine (62).

Table 1.2: Summary of vaccines efficacy against SARS-CoV-2 VOCs

Vaccine	Efficacy	References
NVX-CoV2373 (Novavax)	43% against Beta (HIV negative and positive individuals)	(53,63,64)
ChAdOx1 nCoV-19 (AstraZeneca)	10.4% (mild-to-moderate disease)	(53,65,66)

Janssen/ Ad26.COV2-S (Johnson & Johnson)	52% against Beta (moderate to severe disease) 73% against Beta (severe to critical disease) 72% against Omicron (severe disease)	(32,33,62)
Comirnaty/ BNT162b2 (Pfizer-BioNTech)	100% against Beta (severe disease) 71% against Omicron (severe disease)	(31,62)

1.6. Immune responses to SARS-CoV-2

The CD8+ cytotoxic T cells (CTLs) and CD4+ helper T cells play important roles in host immune defences against viral infections. CTLs destroy virus infected cells that present viral protein fragments/ epitopes bound on their surface with the assistance of the Major Histocompatibility Complex (MHC) class 1 molecules (67,68). MHC is located on the surface of nucleated cells that use antigenic activity to destroy foreign proteins when the virus infects the cells. The protein fragments/ epitopes become accessible by MHC class 1 which activates CTLs to destroy the infected cells by cytotoxicity. CD4+ T cells utilise antigens bound to MHC2 receptors to assist with viral clearance (67,68). MHC class 2 is expressed on antigen-presenting cells including macrophages, dendritic cells, monocytes and B cells. CD4+ T cells identify MHC class 2 that bind antigens but lack cytotoxic activity and require cytokines release to encourage the function of macrophages and CTLs activity.

SARS-CoV-2 specific CD8+ (70%) and CD4+ (100%) T cells were also identified among COVID-19 patients, in a separate study (69). Investigating the T cell responses within infected and uninfected individuals can provide insight into protective immunity against SARS-CoV-2. A study that investigated patients that recovered from infection with SARS-CoV-2 compared to unexposed patients, showed that recovered patients consistently developed CD4+ T cell responses against SARS-CoV-2 (70). Approximately 50% of the T cell responses were directed towards the spike protein. In addition, the study also showed that spike-specific CD4+ T cell responses were mainly associated with the level of anti-RBD IgG titers (70). CD4+ T cell response to the S protein was associated with the high levels of anti-SARS-CoV-2 IgG and IgA titers. SARS-CoV-2 reactive CD4+ T cells were also detected in 40-60% of unexposed individuals, which may suggest cross-reactive T cell response between previously circulating hCoVs and SARS-CoV-2 or transient SARS-CoV-2 infection that the participants were unaware of (69). Most of the SARS-CoV-2 variants house mutations that escape neutralization through B cell responses, similar to that observed in the Wuhan-Hu1 strain (58).

Since the S protein is the focal target of current vaccines, understanding the T cell responses to SARS-CoV-2 can assist in the development of improved SARS-CoV-2 vaccines. The inclusion of the conserved regions (lower variability) of SARS-CoV-2 specific antigens and known T cell epitopes for vaccine development could ensure that neutralising antibodies coupled with T cell responses reduce morbidity and mortality among vaccinated persons (71).

1.7. Justification

It has been approximately three years since the first case was reported in South Africa and the availability of appropriate vaccines that target emerging VOCs are continually being developed. The emergence of what is now termed SARS-CoV-2 VOCs, due to the presence of mutations in the spike glycoprotein that makes these strains more resistant to antibody-mediated neutralisation, has highlighted the need for ongoing characterisation of SARS-CoV-2 strains. Current variants of concern like the Alpha variant, the Beta variant that was first detected in South Africa, and the Delta variant, had shown escape from neutralising antibodies induced by the dominant vaccine immune epitopes on the S glycoprotein. Circulation of these variants also resulted in reduced efficacy of licensed SARS-CoV-2 vaccines rolled out to stem the ongoing pandemic. In addition, despite all the vaccines being based on the S glycoprotein from the original Wuhan strain and the circulation of variants, reported vaccine efficacies are quite variable and range from less than 50% to over 90%. This highlights the importance of focusing on genomics surveillance and research on S glycoprotein.

To date, genome data are available for over 16 million SARS-CoV-2 viruses globally, of which data from South Africa now represents just over 55 000 genomes as of 09 November 2023 (72). Much of the SARS-CoV-2 genome data available also lack completeness of the S glycoprotein sequence data which is used to inform continued viral evolution and its impact on vaccine effectiveness. Global genomic data also contribute to decision-making on how to mitigate the effects of viral evolution in future vaccine designs for improved efficacy. Currently, viral genomic data from South Africa and Africa is completely underrepresented among the global dataset and therefore any updates to current vaccine constructs will be based on SARS-CoV-2 strains that dominate in Europe, the Americas and Asia.

Although the SARS-CoV-2 variants that dominated the five waves of the pandemic in South Africa is described, the diversity within each variant population is still being defined and our

understanding of the factors that contribute to the emergence of dominant variants followed by their persistence or eventual replacement is slowly progressing. We therefore aim to assess the molecular epidemiology and immune adaptations across the SARS-CoV-2 S glycoprotein to understand the virological characteristics that defined the SARS-CoV-2 pandemic in Gauteng from 2020 to 2022. In addition, we also aimed to assess if accounting for diversity (immune adaptations) through a derived mosaic S protein candidate immunogen will improve predicted epitope coverage and design an antigenic candidate SARS-CoV-2 vaccine that targets the primary circulating VOCs and lineages observed in this study.

1.8. Aim and objectives

Aim: To describe the molecular epidemiology and immune adaptations across the SARS-CoV-2 S protein and its impact on immune escape and the emergence of viral variants in Gauteng, South Africa, from 2020 to 2022

Objective 1: Rapid identification of emerging variants based on primary mutations in SARS-CoV-2 spike glycoproteins in Gauteng, circulating in 2021.

Objective 2: To describe the molecular epidemiology of SARS-CoV-2 S genes over time from 2020 to 2022 in South Africa, to assess the timing of the emergence of dominant variants, and to assess the significance of low-frequency lineages (potential to influence emergence of new variants) that emerged during the COVID-19 waves.

Objective 3: To describe the impact of intra-host immune adaptations on the evolution of SARS-CoV-2 S protein genes among individuals with SARS-CoV-2 infections, in Gauteng from 2020 to 2021. This will provide a clearer description of the immune response mechanisms that drive evolution as the pool of vaccinated or naturally exposed persons increase over time.

Objective 4: To assess if accounting for SARS-CoV-2 diversity (immune adaptations) through a derived mosaic S protein immunogen will improve predicted epitope coverage when compared to licensed vaccines.

CHAPTER 2
Methodology

2.1. Study setting and population

This study was conducted in the Department of Virology of the National Health Laboratory Service (NHLS) at Charlotte Maxeke Johannesburg Academic Hospital (CMJAH) and the University of the Witwatersrand during 2020 to 2022. The study population included persons of all ages from whom respiratory tract specimens were submitted for SARS-CoV-2/ COVID-19 diagnosis. This included in-patients and out-patients from hospitals in Gauteng Province as well as specimens from community screen and testing programs facilitated by the NHLS and the Gauteng Department of Health. In addition, referred test request samples were received from the Eastern Cape, Free-State, KwaZulu-Natal, Limpopo, and Western Cape provinces of South Africa.

2.2. Study samples and laboratory data

Upper and lower respiratory tract specimens such as nasal and nasopharyngeal swabs (NP), throat swabs and oropharyngeal swabs- (OP), sputum, tracheal and lung aspirates, bronchoalveolar aspirates or lavages were received for SARS-CoV-2 diagnosis. NP/OPs were either collected in viral or inactivation transport medium (Wuxi NEST Biotechnology, Wuxi, China), or as dry swabs which were reconstituted in the laboratory in 1 ml phosphate buffered saline (PBS) or viral transport medium.

Residual respiratory specimens were stored at -20 °C until retrieved for SARS-CoV-2 genome sequencing. These specimens were anonymised by allocating unique research (eg. NCV1, where N= NHLS, C= CMJAH, V= Virology, and the number (1) = unique sample number) numbers to replace the patient's laboratory number which was used to track the corresponding specimens for testing.

For the positive samples included in this study the associated epidemiologic and clinical information was obtained from the NHLS Central Data Warehouse (CDW). After data cleaning the CDW dataset was anonymised for presentation in aggregated form.

2.3. Sample size and selection

Convenient sampling based on available residual respiratory specimens that tested positive for SARS-CoV-2 from March 2020 through December 2022 were done. In addition, due to the huge volumes received for testing during the pandemic, SARS-CoV-2 positive specimens with Ct-values <31 were listed on an Excel spreadsheet and randomised by year, month, and site were selected for sequencing. A total of 50 to 80 positive samples were selected per week during peak and off-peak periods however, when <50 positives were detected, all samples were selected for inclusion. For genotyping by PCR, all positive samples from the study period were selected regardless of the Ct-value.

2.4. Laboratory investigations

2.4.1. SARS-CoV-2 diagnosis

Total nucleic acid (TNA) were extracted from the respiratory samples using any of the three extraction platforms and compatible extractions kits: (i) the MagNA Pure 96 (Roche Diagnostics, Germany) with the MagNA Pure 96 small volume DNA and viral NA extraction kit (Roche), (ii) the microlab® NIMBUS® (Hamilton, United States) systems with the Viral DNA/RNA 200 C extraction kit (Hamilton), or (iii) the KingFisher Flex (Thermo Fisher Scientific, Massachusetts, United States) semi-automated platform with the MagMAX viral/pathogen nucleic acid isolation kit (Thermo Fisher Scientific), according to the manufacturers' instructions.

Multiplex real-time reverse transcription-based polymerase chain reaction (rRT-PCR) assays were performed on TNA according to the manufacturers' instructions for: (i) Allplex™ 2019-nCoV Assay (Seegene Inc., Korea) on the CFX96 real-time platform (Bio-Rad, California, United States); (ii) TaqPath™ COVID-19 assay (Thermo Fisher Scientific) on the QuantStudio™ 5 (Thermo Fisher Scientific) real-time platform, or (iii) the cobas® SARS-CoV-2 assay (Roche Diagnostics) on the high-throughput automated the cobas® 8800 platform (Roche Diagnostics). In addition, urgent SARS-CoV-2 diagnostic test requests were done on a point of care all-in-one extraction and real-time PCR system called the Biofire FilmArray RP2.1 assay (BioFire Diagnostic, United States) and the BioFire Torch (BioFire Diagnostic) platform.

2.4.2. SNP-based VOC identification and SARS-CoV-2 genome sequencing

a) Nucleic acid template preparation

Total nucleic acids (TNA) were extracted from the selected samples using the semi-automated KingFisher Flex purification system (Thermo Fisher Scientific) with the MagMAX™ Viral/Pathogen II nucleic acid extraction kit (Thermo Fisher Scientific), as per the manufacturer's instructions. The TNA eluate was used for the SNP PCR-based assays and for amplification in preparation for sequencing.

b) PCR-based genotyping to detect single nucleotide polymorphisms (SNPs) associated with spike protein mutations

The TaqMan SARS-CoV-2 spike (S) protein singleplex mutation panels (Thermo Fisher Scientific) were used to identify 11 mutations (T20N, delH69/70, delY144, K417N, L452R, E484K, E484Q, N501Y, D614G, P681H, and P681R), as per manufacturer's instructions. Combinations of mutations allowed for the detection of Alpha, Beta, Delta, Gamma and Kappa VOCs (Table 2.1). The mutation panels were used in conjunction with the TaqPath™ 1-Step qPCR Master Mix, CG (Thermo Fisher Scientific), and run on the QuantStudio™ 5 real-time PCR instrument, according to the manufacturer's protocol.

Table 2.1: Variants and associated mutations (SNPs)

Variants	Lineage	Primary mutations detected
Alpha	B.1.1.7	del69/70, delY144, N501Y, P681H
Beta	B.1.351	K417N, E484K, N501Y
Delta	B.1.617.2	L452R, P681R
Gamma	P.1	T20N, E484K, N501Y
Kappa	B.1.617.1	L452R, E484Q, P681R
Unable to classify	Unknown	Single mutations

Analysis was performed using the QuantStudio 5™ design and analysis and TaqMan® Genotyper software. Alleles that clustered along the x-axis (allele 1 VIC) were represented as homogeneous wild type with AF≥ 0.5, alleles that clustered along the y-axis (allele 2 FAM) represented homogeneous mutants with AF≥ 0.5, and alleles that clustered between the x-axis

and y-axis represented heterogeneous wild type/mutant. In summary, this assay thus allowed us to determine the presence of homogeneous and/ or heterogeneous (i.e., mixed mutations associated with more than 1 genotype) virus populations within an individual.

2.4.3. SARS-CoV-2 whole genome sequencing

a) cDNA synthesis and complete genome amplification

The cDNA was prepared with 5.5 µl extracted TNA and 1µl of 5x LunaScript® RT SuperMix (New England Biolabs, USA). The reaction was initiated at 25 °C for 2 min (annealing), reverse-transcription was activated at 55 °C for 10 min, and heat inactivation was at 95 °C for 1 min. The whole genome was amplified using the tiling PCR method which was performed using the ARTIC V3, V4, and V4.1 primers (Inqaba Biotec, Gauteng, South Africa) provided as 2 pools to generate non-overlapping genome fragments (43,73,74). Two sets of reactions (pools 1 and 2) were set up for each primer pool. Amplification was performed with the Q5® Hot Start High-Fidelity 2X Master Mix kit (New England Biolabs, Massachusetts, United States). The total reaction volume was 12.5 µl which comprised 2.5 µl of 5X Q5 Reaction Buffer, 0.25 µl of 10mM dNTPs, 0.125 µl of Q5 Hot Start DNA Polymerase, 1.8 µl of 10 µM primer pool, 5.4 µl of nuclease-free water and 2.5 µl cDNA. The PCR was initiated with heat activation at 98 °C for 30 sec followed by 35 cycles of denaturation at 98 °C for 15 sec and annealing and extension at 65 °C for 5 min.

b) Library prep and next-generation sequencing

Equal volumes of each of the amplicons from pools 1 and 2 were combined for each sample. The pooled amplicons were quantified using the Qubit 4 fluorometer with the Qubit™ 1X dsDNA High Sensitivity kit according to the manufacturer's instructions (Thermofisher Scientific, Massachusetts, United States). One hundred nanograms of each sample was required for library preparation. The pooled amplicons sequenced in-house at the NHLS/CMJAH Virology Laboratory was performed using paired-end (2x150 paired-end reads) sequencing on the Illumina MiSeq platform (Illumina, San Diego, CA, USA) after Nextera DNA library preparation (Illumina, San Diego, USA). Illumina sequencing was performed following the manufacturer's instructions with the following modifications to the library preparation: an input of 7.5 µl of a final concentration of 100ng DNA was used, with quarter reactions applied to the mastermixes prepared for tagmentation, enrichment, and clean up. A final concentration

of 8 μM with 1% PhiX control was prepared and loaded into the flow-cell which was sequenced on the MiSeq platform with the MiSeq V2 300 cycle kit (Illumina, San Diego, USA).

Samples submitted to the KwaZulu-Natal Research Innovation and Sequencing Platform (KRISP) for sequencing were sequenced on the Illumina MiSeq platform (Illumina, San Diego, CA, USA) as described and Nanopore sequencing with the MinION (Oxford Nanopore Technologies) following the Nanopore Midnight protocol using the Rapid Barcoding kit (Oxford Nanopore Technologies, Oxford, United Kingdom) (75). The primers used in the Midnight protocol (76,77) were similar to the ARTIC primers described and were updated from ARTIC V3 to V4.1 primers (43,73,74) as VOCs emerged to prevent primer mismatch. Each of the amplified pools (similar to that described for the ARTIC primers) generated 1200bp overlapping amplicons. Each pool was combined for each sample and used directly for library preparation using the Rapid Barcoding kit as per the manufacturers' instructions. The samples were barcoded and then pooled together in a single tube, which was then purified using AMPure XP bead clean-up method (provided with the kit). The libraries were then loaded into the R9.4.1 flow-cell which was sequenced using the GridION X5 or the MinION platforms.

c) Sequence assemblies and analysis

Genome assembly was performed using Genome Detective (78) or Exatype (79) online tools to generate the consensus sequences. The Wuhan-Hu-1 strain (NC_045512.2; https://www.ncbi.nlm.nih.gov/nuccore/1798174254/, accessed 3 April 2023) was used as the reference sequence throughout the assembly and analysis steps. The consensus sequences were aligned against the Wuhan-Hu-1 strain using AliView, and thereafter uploaded to Global Influenza Sequence Database (GISAID; https://www.gisaid.org/) for curation.

Curated sequence data from this study were downloaded from GISAID for analysis. The sequences were uploaded to Nextclade (https://clades.nextstrain.org/, (80)) and the Pango lineage tool (https://pangolin.cog-uk.io/) to determine the clade, lineage, and VOC as applicable. The CoVServer program on GISAID and Stanford Coronavirus Antiviral and Resistance Database (covdb.stanford.edu/sierra/sars2) were used to identify mutations and their frequencies.

Phylogenetic trees were constructed to describe the evolutionary relationship between this study sequences and other South African and global sequences. The phylogenetic trees and time-trees were designed using Nextstrain (https://clades.nextstrain.org) and Nextstrain custom build embedded with pangolearn. Pangolearn (https://cov-lineages.org/resources/pangolin/pangolearn) comprises an algorithm to assign lineages rapidly. The model utilised sequences uploaded into GISAID where lineages are assigned manually during curation of a maximum-likelihood tree which provides accurate lineage or VOC calling. These tools are embedded with SARS-CoV-2 whole genome sequences from across the globe, which cover all lineages and VOCs detected from 2019 to date.

2.4.4. Data analysis

The prevalence and variation of SARS-CoV-2 strains and its associated mutations, from the first to the fifth COVID-19 wave, were analysed. The analysis was performed across different age groups (<5, 5-14, 15-24, 25-44, 45-65, >65 years), by collection site, period (epiweek, month, or year), and patient outcomes. Analyses were performed using Stata® 14 (StataCorp, College Station, Texas, USA) and Excel.

2.4.5. Design and synthesis of a mosaic S immunogen

a) S gene sequence alignment and translation

S genes that were sequenced together with all other South African sequences available on GISAID were downloaded. GISAID's data access agreement policy (https://www.gisaid.org/about-us/acknowledgements/sharing-policy/) was adhered to. Nextclade S gene sequences were excised from aligned genomes. Further analysis of aligned sequences was done in AliView and translated into amino acids and the S protein region was trimmed.

b) Mosaic design

The mosaic vaccine design tool (https://www.hiv.lanl.gov/content/sequence/MOSAIC/) allows the generation of mosaic protein sequences by selecting and combining optimal cytotoxic T cell epitopes through a computational algorithm from the multiple sequence alignment to generate recombinant protein sequences that resemble the natural viral proteins. An optimum mosaic protein is assembled from optimal natural epitopes in each sequence present in the multiple sequence alignment which are assembled to generate a single (or multiple)

recombinant "natural" protein sequence(s). Fischer *et al.* (81) reported that through the use of this mosaic method, the predicted immune coverage of HIV diversity by mosaics was greatly increased when compared to coverage by natural-sequence-based vaccines. MOSAIC designer was run and 3 best natural sequences representing the mosaic protein were identified. An optimal S protein immunogen was generated using the translated amino acid sequences as input for the mosaic vaccine design tool.

2.4.6. Epitope coverage predictions across the S glycoprotein for a candidate mosaic immunogen compared to licensed vaccines

Predicting T cell epitopes is essential for use *in silico* for the development of appropriate vaccines. The derived mosaic S immunogen epitope and positional CTL epitope coverage was assessed and compared to licensed S vaccines using the online tools available on the HIV Sequence Database (https://www.hiv.lanl.gov/content/sequence/MOSAIC/posicover.html).

a) CTL epitope predictions

The mosaic vaccine tool suite was used to determine the positional CTL epitope coverage assessment for SARS-CoV-2 S protein sequence to show how much of the respective protein's length are covered by the derived mosaic peptide cocktails (https://www.hiv.lanl.gov/content/sequence/MOSAIC/). The images of the 9-mer coverage by position and unique 9-mer counts images were downloaded. For these tools identical sequences were removed, errors, gaps and non-specific nucleotides were removed, and sequence names shortened to acceptable lengths.

2.4.7. Assessing the antigenicity and toxicity of the S immunogen

Toxicity was predicted using the ToxinPred tool (https://webs.iiitd.edu.in/raghava/toxinpred/algo.php). Predicted toxins were defined as having a support vector machine (SVM) threshold value of ≥ 0.00 and an expected (E) value cut-off of 10. The ToxinPred2 tool (https://webs.iiitd.edu.in/raghava/toxinpred2/) was used to assess toxin properties for the whole S proteins. The tool was used to highlight the regions in the protein sequence that are toxic and should be excluded from the vaccine construct.

The antigenicity of the epitopes was determined using VaxiJen v.20 with default settings which was set to virus and sequence output (http://www.ddg-pharmfac.net/vaxijen/VaxiJen.html). The threshold of probable antigen was defined as a score ≥ 0.4 with 70% accuracy.

2.4.8. Glycosylation predictions

N-glycosylation sites were predicted by submitting the amino acid sequences to the NetNGlyc (https://services.healthtech.dtu.dk/services/NetNGlyc-1.0/) (82). The amino acid sequence was submitted to NetOGlyc (https://services.healthtech.dtu.dk/services/NetOGlyc-4.0/) to predict the O-linked glycosylation sites (83). A threshold of ≥ 0.5 for an amino acid was predicted as positive N- and/or O-glycosylation site. Analysis of the N- and O- glycosylation sites in the S protein of SARS-CoV-2 is necessary for identifying immunogens for improved vaccine design since it influences the efficacy of vaccine candidates.

2.4.9. Synthesis of S immunogen

An optimal mosaic protein was reconstituted from the 3 best natural sequences and reverse translated before it was codon-optimised for enhanced expression in mammalian cells. The optimised mosaic S gene sequence including the signal peptide and terminal stop codon was submitted for cloning and synthesis (BioMatik, Ontario, Canada) in a pCDNA3.1(-) myc-HisA plasmid vector. Cloning sites included NheI(GCTAGC) and EcoRV(GATATC). The synthesised construct was received in lyophilised form which was centrifuged briefly and reconstituted in 100 µl of Tris-EDTA buffer to make up a final concentration of 100ng/µl. The reconstituted form of the DNA construct was used for transfection.

2.4.10. Cell culture and transfection

pcDNA clones with the S protein inserts were transfected into HEK-293T cells for optimal protein expression following transfection. The 293T cells were resuscitated in 10% growth medium which was prepared by adding 50 ml of FBS to 435 ml of Dulbecco's Modified Eagle Medium (DMEM), 2.5 ml of gentamicin and 12.5 ml HEPES. The cells were maintained with 10% growth medium. Transfection was then performed with Lipofectamine™ 3000 transfection reagent (Invitrogen, United States) to obtain high transfection efficiency and was used according to the manufacturer's instructions, with 6 well plates. Two milliliters of 10% growth medium was added to each well after the transfection process. The plates were incubated in 5% CO_2 at 37°C for 48h. The cells and culture supernatant fluids were harvested separately after 48 h post transfection to recover the expressed S proteins. Approximately 0.5

× 10^6 of 293Tcells were added to each well. Lipofectamine™ 3000 Reagent (7.5 µl) was diluted in 125 µl Opti-MEM™ Medium (Reduced-Serum Medium) and vortexed briefly. Five microliters of the S protein construct were diluted (for 0.6 µg of construct) in 250 µl Opti-MEM™ Medium, and 1.2 µl P3000™ Reagent was added thereafter. The DNA-lipid complex was added to each seeded well in the 6 well flat bottom plate and topped up with 200 µl of DMEM with 10% FBS (excluding antibiotics). The plates were incubated for 48 h in 5% CO_2 at 37 °C and checked for impact on subsequent immunity against emerging variants using an inverted microscope every 24 h. Cells were harvested once CPE was evident. The negative control used was untransfected 293T cells and the positive control was the unmodified S protein plasmid.

2.4.11. Polyacrylamide gel electrophoresis (PAGE)

Identification of expressed and purified mosaic S immunogen and controls were determined by their expected size of the protein expressed in the supernatant fluids. Laemmli sample buffer (1x) was prepared by combining 2.5 µl Tris-EDTA, 5ml of 10% SDS, 20ml glycerol, 10ml β-Mercaptoethanol, and 1 ml tryphan blue. For each sample, 50 µl of the supernatant was mixed with 150 µl of 1x laemmli buffer. Denaturation was undertaken by boiling the samples with buffer at 95 °C for 5min. Gel electrophoresis was performed using the Eco-min Biometra (1.5mm thick) SDS-PAGE vertical electrophoresis system with a mini-PROTEAN TGX Stain-Free gel. The tank was filled with 1x Tris/Glycine/SDS buffer. Twenty microliters of Blueye prestained protein ladder (Sigma Aldrich, United States) was loaded, followed by 20 µl of each sample, and run for 1 h at 100 Amps. The prestained gel was viewed under the UV light in the gel doc system. The expected band size of ~180 kDa for the S polyprotein construct will be used to verify protein expression. The apsize of the S1 and S2 regions are approximately 76.4 kDa and 56.3 kDa.

2.5. Ethics

This study was approved by the University of Witwatersrand Human Research Ethics Committee (M210119).

CHAPTER 3

Rapid identification of SARS-CoV-2 VOCs through detection of single nucleotide polymorphisms in the spike protein

Reference (Appendix A): 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., Makatini, Z., Fielding B. C., Treurnicht, F. K., Network for Genomics Surveillance in South Africa (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), 3676-3684.

3.1. Introduction

The rapid evolution of SARS-CoV-2 has led to the emergence of several VOCs, which have been classified as variants with key characteristic features bearing significant epidemiological and clinical consequences (48). In 2020, the Wuhan lineage was replaced by B.1 and B.1.1 lineages, after which VOCs emerged independently starting with Beta (B.1.351 lineage), Delta (B.1.617.2 lineage) and then Omicron (B.1.1.529 lineage) (84). In South Africa, the initial circulating Wuhan strain, in addition to several other SARS-CoV-2, was replaced by Beta, marking the first emergence of a VOC in the country. The Beta variant was first detected in South Africa in late 2020 and dominated from October 2020 through May 2021 during the second wave of infections; Delta dominated the third wave from June through November 2021; while the C.1.2 lineage (variant under monitoring) was detected at very low frequencies from July through December 2021 (43,85–87). In South Africa, recently emerged Omicron (B.1.1.529 lineage) was initially identified in a patient declared COVID-19 positive on 9 November 2021 and since then Omicron has dominated the fourth wave (88). Since the initial detection of Omicron, the number of cases increased more rapidly compared to previous waves (88,89).

Omicron has over 30 mutations in the spike (S) protein, some of which overlap with the Alpha (del69/70, P681H), Beta (K417N, N501Y) and Delta (G142D and T478K) VOCs

(75,88,90,91). A number of these mutations, including the del69/70, are predicted or known to have an impact on immune escape or transmissibility (75). Since the discovery of Omicron (B.1.1.529 lineage), this VOC has been split into four sub-lineages namely BA.1, BA.1.1, BA.2 and BA.3. Omicron BA.1 was the dominant variant in South Africa from November 2021, followed by Omicron BA.2 that circulated from December 2021. The deletion at amino acid position 69/70 was a distinguishing feature between the Omicron lineages: the del69/70 is present in Omicron BA.1 and B.A.3, but is absent in Omicron BA.2 (92). The del69/70 affects the amplification of the S gene target during polymerase chain reaction (PCR) resulting in the S gene target failure (SGTF) detected by TaqMan SARS-CoV-2 S protein SNP assays (19,93). The latter was found to be significant for the detection and reporting of Omicron (19,93).

Although next-generation sequencing (NGS) remains the ideal tool for surveillance and detection of novel SARS-CoV-2 VOCs, many low to middle income countries are unable to effectively implement this tool due to lack of resources, facilities or expertise. Inconsistencies in testing and time delays in generating and releasing sequencing data were reported to hinder surveillance initiatives in African countries (86). Although still posing many challenges, the implementation of molecular diagnostic testing is far more achievable when coordinated efforts are made in comparison to the implementation of NGS.

Here, we investigated the use of the Thermo Fisher TaqPath™ COVID-19 assay and the SARS-CoV-2 genotyping assays to rapidly identify infections caused by emerging VOCs during the Delta and Omicron BA.1 waves in South Africa.

3.2. Materials and methods

3.2.1. Detection of VOCs using single nucleotide polymorphism (SNP) genotyping assays

a) Study population

The study population (as described in 2.1) included individuals from the Gauteng and Eastern Cape Provinces, from 01 June 2021 to 31 October 2021. This included samples collected from in-patients, out-patients and through community surveillance. The majority of respiratory samples received included NP and or OPs in viral or inactivation transport medium (Wuxi NEST Biotechnology, Wuxi, China), or dry swabs reconstituted in 1 ml PBS or viral transport medium.

b) *Total nucleic acid extractions*

Extraction was performed as described in 2.4.2.

c) SARS-CoV-2 diagnostics

SARS-CoV-2 was diagnosed using testing methods described in 2.4.1.

d) SNP genotyping

SNP genotyping was performed as described in 2.4.3.

3.2.2. Detection of del69/70 and the K417N PCR based genotyping method for rapid detection of Omicron BA.1

a) Study population and sample selection

Samples that tested positive for SARS-CoV-2, from the diagnostic SARS-CoV-2 PCR performed from 01 November to 30 November 2021, were randomly selected by collection date and site, regardless the Ct-value. This period was selected since the detection rate (number of positive tests/number of samples received per day) of SARS-CoV-2 began to steadily increase which was an indication of the beginning of a potential new wave.

b) Total nucleic acid extraction

TNA was extracted using the KingFisher Flex purification system as described in 2.4.2. Samples that were initially tested on the CFX and BioFire platforms were re-tested using the TaqPath™ COVID-19 assay to detect S gene target failure (SGTF). The SGFT was defined as any sample for which the S gene did not amplify, but the N and/or ORF1ab genes amplified with Ct values <38.

c) SNP genotyping by PCR

SNP genotyping was performed, as described in 2.4.3, to detect the del69/70 and K417N mutations characteristic of Omicron BA.1. These assays differentiate between wild type (H69V70 and K417) and mutant (del69/70 and 417N) in each sample. Analysis was performed as described in 2.4.3.

3.2.3. Genotyping by SARS-CoV-2 genome sequencing

Samples with Ct-values <31 were randomly selected across collection date and site on a weekly basis for SARS-CoV-2 whole genome sequencing. As part of the Network for Genomics Surveillance in South Africa (NGS-SA), samples were submitted to the KwaZulu-Natal Research Innovation and Sequencing Platform (KRISP), the National Institute for Communicable Diseases (NICD), or sequencing was performed in-house. Samples were amplified using the ARCTIC V4 primers (94) and libraries were prepared using the Nextera XT DNA Library Prep Kit (Illumina) which were sequenced on the MiSeq platform (Illumina, California, Unites States), or COVIDSeq (Illumina) library kits were used with the NextSeq platform (Illumina) (75,95). Sequence reads were assembled using the Genome Detective SARS-CoV-2 online tool (https://www.genomedetective.com/app/typingtool/virus/) by KRISP, the Exatype pipeline (https://sars-cov-2.exatype.com/) by the NICD and the Galaxy SARS-CoV-2 pipeline (https://usegalaxy.eu/) for in-house sequence reads. The consensus sequences were uploaded onto the GISAID (https://www.gisaid.org/) for curation. Sequences were then downloaded from GISAID for further analysis using the Nextstrain (https://clades.nextstrain.org) online tool for the construction of the phylogenetic trees and Pangolin lineage assigner (https://pangolin.cog-uk.io) was used to confirm the lineages.

3.2.4. Data analysis

Alpha, Beta, Delta, Gamma and Kappa were identified based on a combination of primary mutations as indicated in 2.4.2, Table 2.1.

For the detection of Omicron BA.1, since the del69/70 and K417N were also present in other VOCs, genotyping by PCR was used to identify VOCs using the algorithm for results interpretation was followed as per table 3.1 during the study period.

Table 3.1: Interpretation of results for VOCs based on the presence or absence of specific mutations

Variant	del69/70	K417N
Alpha	√	x
Beta	x	√
Delta	x	x
Omicron BA.1	√	√

The prevalence and trends of SARS-CoV-2 variants and their associated mutations were analysed across different age groups (<1, 1-4, 5-17, 18-24, 25-44, 45-60, >60), gender, period (based on specimen collection day/month) and patient status (patients attending ARV clinic, community screen and testing, deceased, in- and out-patients, healthcare workers etc), in November 2021.

3.3. Results

3.3.1. SARS-CoV-2 diagnosis among the study population

The overall detection rate was 22.4% (37 783/168 533) from 1 June to 30 November 2021 (Fig. 3.1). The peak detection rate (>40%) was observed during the 3rd COVID-19 wave, from 29 June to 17 July 2021. From 8 September to 21 November 2021 the detection rate decreased below 10% but increased from 20 November marking the beginning of the 4th COVID-19 wave.

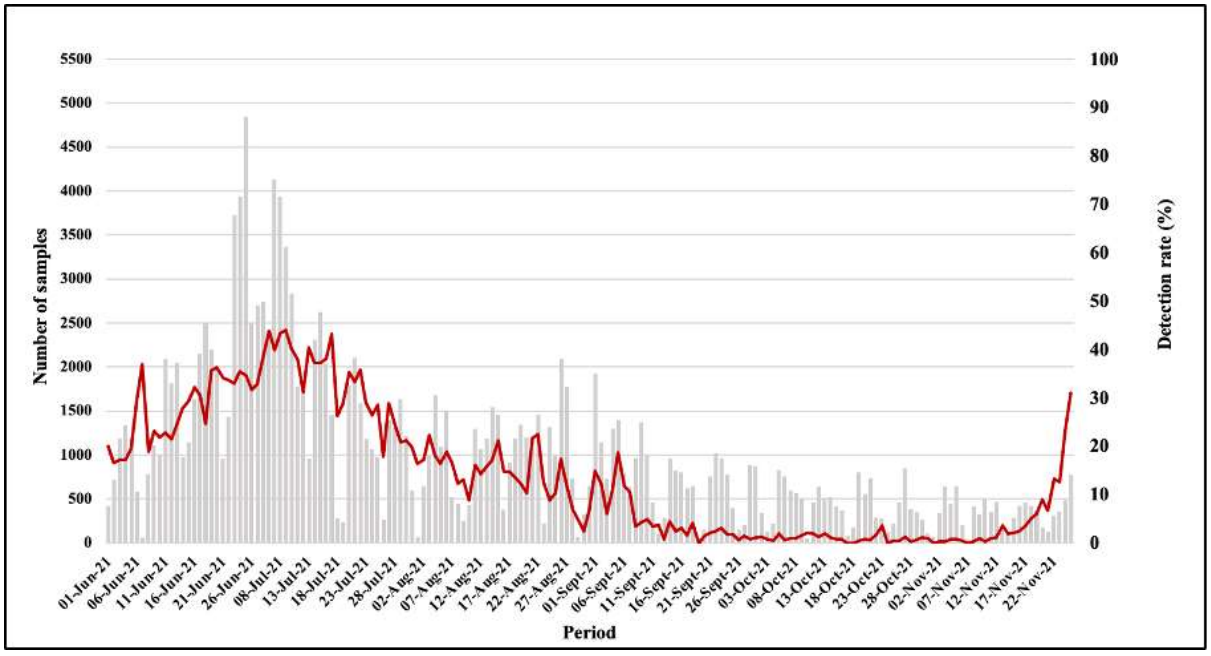

Fig. 3.1: SARS-CoV-2 detected in respiratory samples tested from 1 June to November 2021. The red line graph represents the detection rate, and the bar graph represents the number of samples received for SARS-CoV-2 diagnostics during that period.

3.3.2. Detection of VOCs from June to October 2021

a) Description of VOCs detected from SNP PCR genotyping assays

A combination of P681R and L452R SNPs were detected in majority (73.6%; 538/731) of the samples which were classified as Delta, followed by the N501Y and del69/70 SNPs (3.6%;

26/731) which were classified as Alpha (Table 3.2). Of the 731 participants, 611 resided in Gauteng and the remaining 120 resided in the Eastern Cape. In the Gauteng, Delta dominated at 71.8% (439/611), followed by Alpha (2.1%; 13/611), while Beta and Kappa were only detected in <1% of the samples (Fig 3.2). Similarly, in the Eastern Cape, Delta was the dominant VOC (82.5%; 99/120) (Fig. 3.3). We were unable to determine the VOCs (unclassified) in 25.5% (156/611) of the samples from Gauteng and 17.5% (21/120) from Eastern Cape. In Gauteng, the City of Johannesburg district (61.5%; 376/611) was home to the majority of individuals tested (Fig 3.2), followed by Sedibeng district (0.1%; 109/611). Alpha was only detected in the City of Johannesburg and Sedibeng, while Kappa was only detected in the City of Tshwane district. In the Eastern Cape, most of the individuals tested were from Nelson Mandela Bay Metro (27.5%; 33/120), Sarah Baartman (24.2%; 29/120) and Joe Gqabi (21.7%; 26/120) districts. Only Delta was detected in the Eastern Cape with a prevalence of >80% across all districts except Joe Gqabi where only 54% (14/26) were confirmed as Delta.

Table 3.2: The VOCs detected from SNP genotyping assays for variant-specific S protein mutations

Variants	Lineage	Primary mutations detected	Detection, n/N (%)
Alpha	B.1.1.7	N501Y, del69/70	26/731 (3.6)
Beta	B.1.351	N501Y, E484K, K417N	4/731 (0.6)
Delta	B.1.617.2	P681R, L452R	538/731 (73.6)
Kappa	B.1.617.1	P681R, L452R, E484Q	2/731 (0.3)
Unclassified	Unknown	Single mutations	177/731 (24.2)

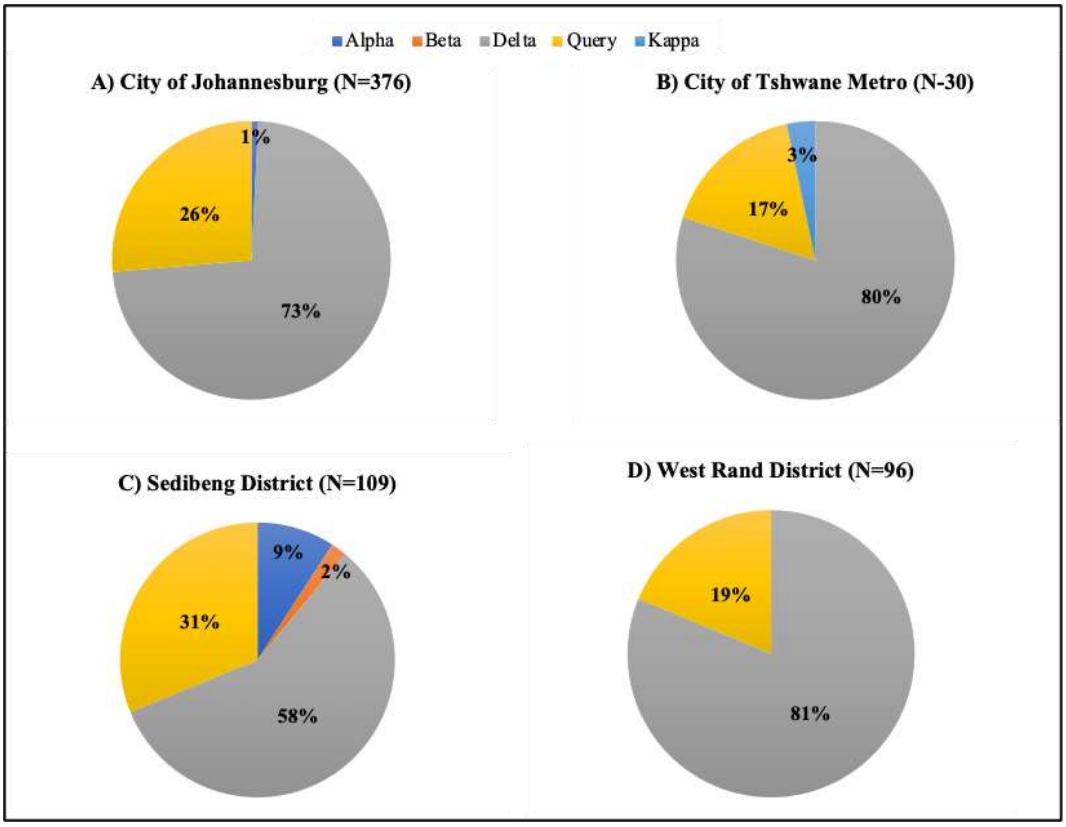

Fig. 3.2: Variants/lineages detected in four districts of the Gauteng Province from June to October 2021 (N=611). The percentages indicate the prevalence of the VOCs detected within a district.

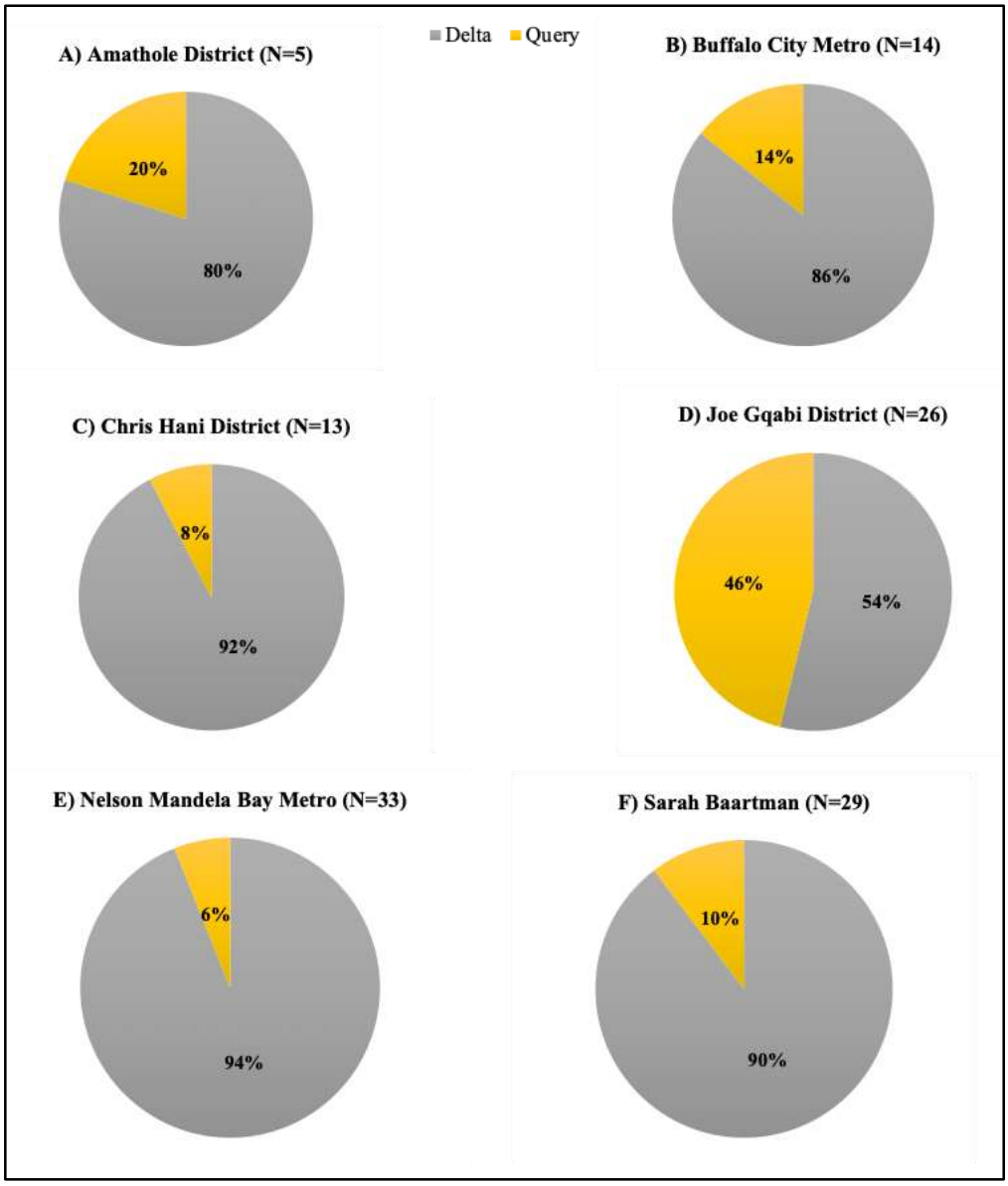

Fig. 3.3: Variants/lineages detected in six Eastern Cape districts (N=120), from June to October 2021. The percentages indicate the prevalence of the VOCs detected within a district. Only Delta and query lineages are represented in this figure since these were only detected in the Eastern Cape.

b) Genotypes confirmed by sequencing

Of the 731 samples genotyped by SNP PCR, 236 samples were randomly sequenced. The majority of the samples sequenced were classified as Delta (87.0%; 214/236), however only 174/236 were also classified as Delta from SNP PCR with the remaining detected as Beta, Kappa, wild type or unclassified. One hundred percent (1/1) of samples classified as Alpha from SNP PCR was confirmed by sequencing. The remaining 8.1% (20/246) samples were incorrectly classified (Beta, Kappa, wild type or unclassified) by the SNP PCR assay and were

confirmed as C.1.2 or Delta from sequencing. All Delta samples confirmed from sequencing, clustered with Delta sub-lineages AY.45, AY.38 and AY.19, all of which had both the P681R and L452R mutations. Other lineages and VOCs were not represented on the tree due to low sequence quality across regions of the genome other than the S gene (Fig. 3.4). Delta was detected from June to October 2021, peaking (46.3%; 99/214) in September (Fig. 3.5). C.1.2 was detected at lower frequencies (10%; 2/20 to 45%; 9/20) from July to October 2021. Alpha was only observed once in June and B.1 once in July.

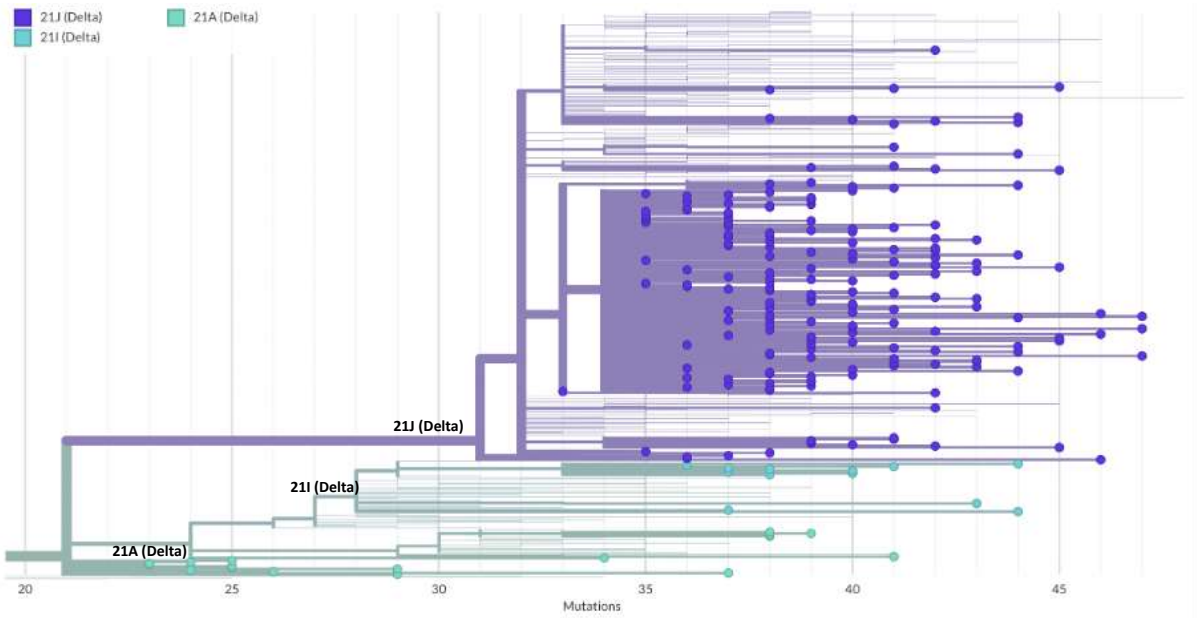

Fig. 3.4: Phylogenetic representation of sequences from Virology samples, NHLS Gauteng, from June to October 2021. Sequence data (circular nodes) are plotted against SARS-CoV-2 global data from next clade (https://clades.nextstrain.org). The *x*-axis represents the number of mutations present. The purple nodes: Delta sub-lineage AY.45 (21J), light blue nodes: Delta sub-lineage AY.38 (21I), turquoise nodes: Delta sub-lineage AY.19 (21A). Figure adapted from clades nextstrain.org (accessed 31 July 2023). Data can be accessed from GISAID Identifier: EPI_SET_230801yz.

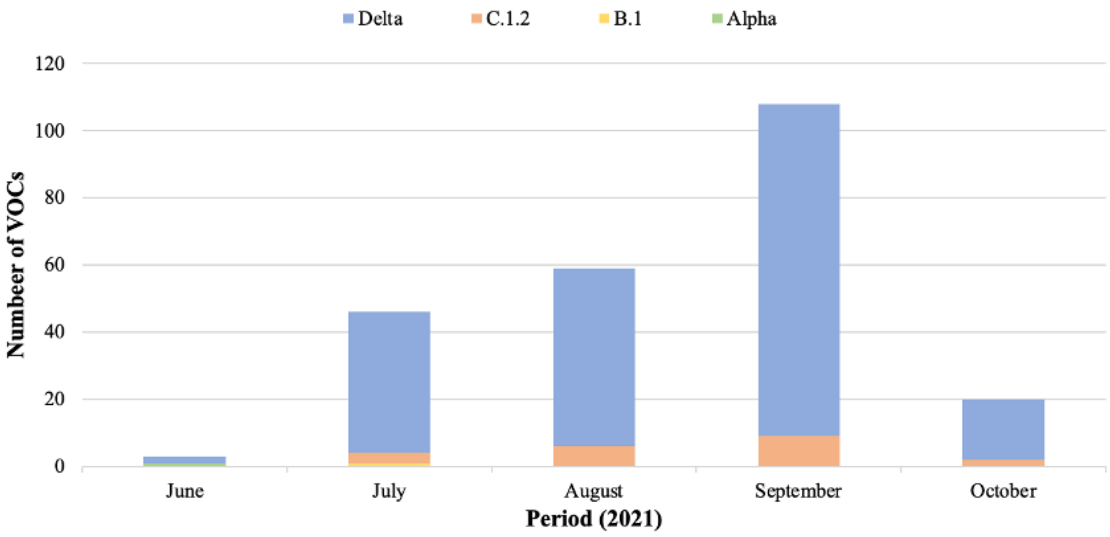

Fig. 3.5: The number of variants detected over time from June to October 2021 based on sequence data.

3.3.3. Rapid detection and confirmation of Omicron BA.1 emergence in November 2021

a) Demographics of individuals and prevalence of SGTF

For the period 1 to 30 November 2021, a total of 11 549 diagnostic tests for SARS-CoV-2 were performed, of which 1 589 (13.8%) samples tested positive. The number of samples received and tested were less from 6-7th, 13-14th, 20-21st, 27-28th November, which were indicative of weekends (Fig 3.6). The detection rate of SARS-CoV-2 was below 6.7% from 1st to 21st November 2021, and by the last week of November (22nd to 30th), the detection rate increased steadily up to 47.5% (Fig 3.6). From 15 to 25 November, 32.6% (175/536) positive samples were randomly selected to assess SGTF and genotyping by PCR. Among these, 57.7% (101/175) were females. Community surveillance samples represented 69.7% (122/175), in-patient samples 9.1% (16/175) and out-patients 13.1% (23/175) (Table 3.3). Among both males (43/71; 60.6%) and females (54/101; 53.5%), most participants were 25-44 years of age (Table 3.3).

Of the 175 samples that were genotyped, 55% (97/175) were initially tested with the Allplex 2019-nCoV and BioFire RP2.1, and only 2.1% (2/97) of these did not have SGTF (Table 3.2) when re-tested with the TaqPath™ COVID-19 assay. The remaining 78 samples, which were initially tested with the TaqPath™ COVID-19 assay, all displayed SGTF. Thus overall, 98.9% (173/175) of SARS-CoV-2 positive study samples received from 15 to 25 November 2021 exhibited SGTF (Fig 3.7).

Table 3.3: Prevalence of SGTF among patients who tested positive for SARS-CoV-2 (N=175)

Gender	Age (y)	SGTF present, n/N(%)			
		Community screening	In-patient	Out-patient	Unknown
Female (101/175; 57.7%)	<5 (n=1)	1/70 (1.4)	-	-	-
	5-24 (n=26)	18/70 (25.7)	3/13 (23.1)	2/14 (14.2)	2/3 (66.7)
	25-44 (n=54)	39/70 (55.7)	6/13 (46.2)	9/14 (64.3)	-
	45-60 (n=16)	10/70 (14.3)	2/13 (15.4)	3/14 (21.4)	1/3 (33.3)
	>60 (n=1)	1/70 (1.4)	-		-
	Unknown (n=3)	1/70 (1.4)	2/13 (15.4)	-	-
Male (71/175; 40.6%)	<5 (n=0)	-	-		-
	5-24 (n=19)	14/49 (28.6)	1/3 (33.3)	1/9 (11.1)	3/9 (33.3)
	25-44 (n=43)	30/49 (61.2)	1/3 (33.3)	6/9 (66.7)	5/9 (55.6)
	45-60 (n=7)	4/49 (8.2)	-	2/9 (22.2)	1/9 (11.1)
	>60 (n=2)	1/49 (2.0)	1/3 (33.3)	-	-
Unknown (3/175; 1.7%)	<5 (n=0)	-	-	-	-
	5-24 (n=1)	1/3 (33.3)	-	-	-
	25-44 (n=2)	2/3 (66.7)	-	-	-
	45-60 (n=0)	-	-	-	-
	>60 (n=0)	-	-	-	-
***Total**		122/175 (69.7)	16/175 (9.1)	23/175 (13.1)	12/175 (6.9)

Abbreviation: SARS-CoV-2, severe acute respiratory syndrome coronavirus 2. *no SGTF= 2

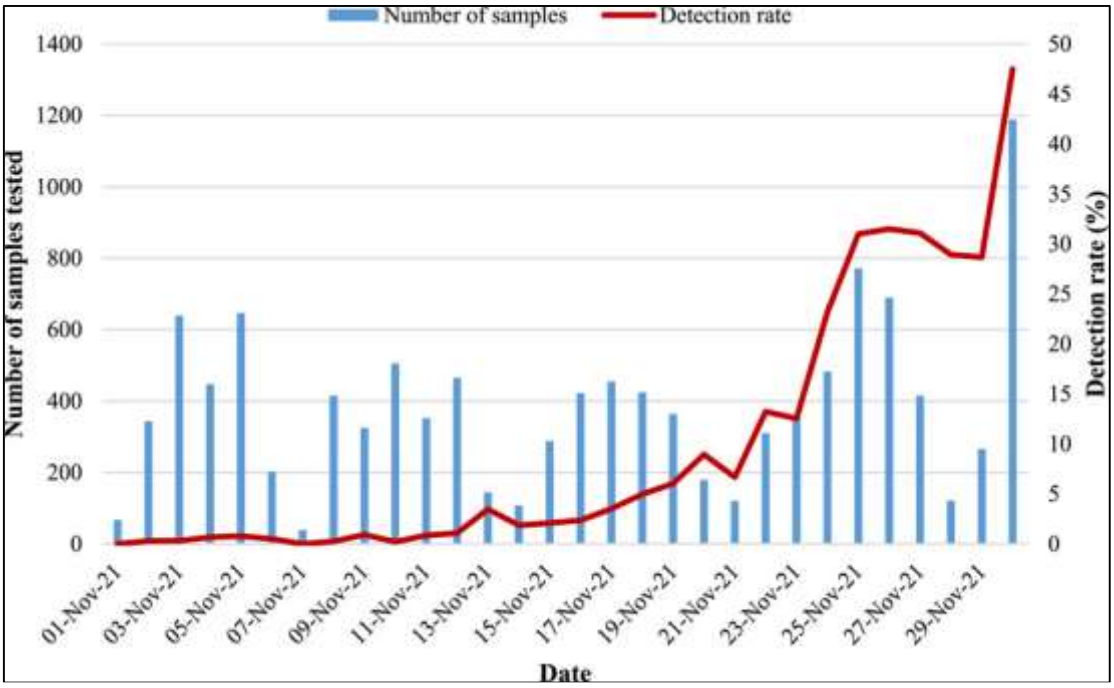

Fig. 3.6: **Detection rate of SARS-CoV-2 in Gauteng, South Africa, in November 2021.** The bar graph represents the number of samples that were received at Virology, NHLS for SARS-CoV-2 diagnostics. The red line graph represents the detection rate (number of positive tests/number of samples received per day). NHLS, National Health Laboratory Service; SARS-CoV-2, severe acute respiratory syndrome coronavirus 2.

b) Detection of del*69/70 and K417N by PCR-based genotyping*

The del69/70 and K417N mutations, characteristic of the Omicron BA.1 sub-lineage, were observed in 88.0% (154/175) of samples (Fig 3.7), of which 87/154 (56.5%) were among SARS-CoV-2 positives initially diagnosed with the Allplex 2019-nCoV and BioFire RP2.1 assays. All mutations were homogeneous (Fig 3.8). Among the samples that did not display SGTF (2/175; 1.1%) one did not have the del69/70 but the 417N mutation was present, and one had both the 417N and the del69/70 (both were confirmed by sequencing as BA.1) (Fig. 3.7). Here, genotyping identified an additional 2 samples that had the wild type genotype at amino acid positions 69/70 and 417. We were unable to determine the presence of del69/70 for 8.0% (14/175) of samples with SGTF as a result of processing an error when the genotyping assays were performed. These samples were SARS-CoV-2 positive with N gene Ct-values of 15.4 to 34.8 for the TaqPath™ COVID-19 assay (Table 3.4).

The initial Ct values generated during the TaqPath™ COVID-19 diagnostic assay were examined to determine if they could possibly have an impact on genotyping results. The lower the Ct value the higher the viral load and vice versa, and therefore we expected to successfully

genotype all samples with a lower Ct value. The results listed in Table 3.4 show evidence that this was not always the case. Only 28.6% (4/14) of samples were sequenced and were confirmed Omicron BA.1, of which 100% (4/4) had the del69/70 mutation, 50% (2/4) had the K417N mutation and the remaining 2 had a K417KN mixed mutation.

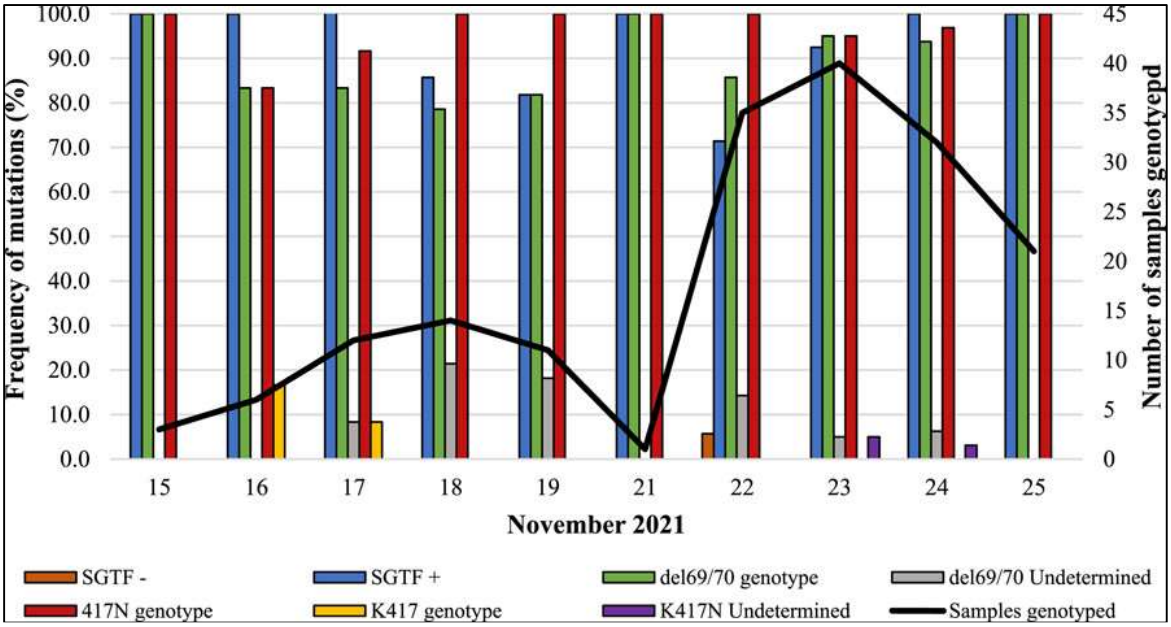

Fig. 3.7: **Frequency of SGTF, del69/70, and K417N mutations among SARS-CoV-2 positive samples collected from November 15 to 25, 2021.** SGTF− represents the samples for which SGTF was absent. SGTF+ represents the samples with SGTF. From the genotyping PCR assay: The del69/70 genotype and del69/70 undetermined are samples with and without the deletion, respectively; and the K417, 417N, and K417N undetermined represent samples that had the wild-type, mutation and neither at amino acid position 417, respectively. Genotyping was not performed on positive samples collected on November 20th.

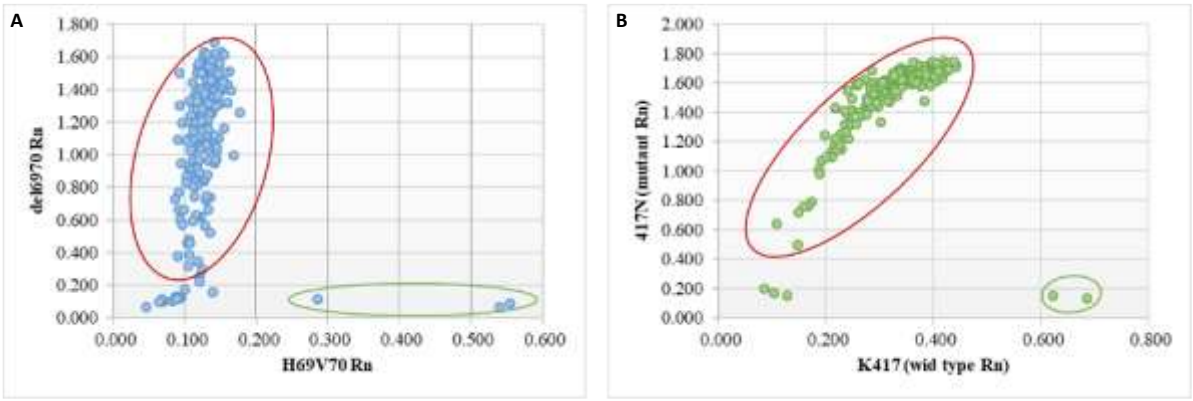

Fig. 3.8: Allelic discrimination plot of del69/70 (A) and K417N (B) mutations within the S gene showing florescence values (Rn). Samples within the red oval (along the x-axis) represent homozygotes for either the deletion (A) or the 417N mutation (B). Samples within the green oval (along the y axis) represent those that do

not have the deletion or have K417 wild type (homogeneous wild type). Samples lying outside these areas represent a negative result (does not have the mutation or wild type).

Table 3.4: Comparison of Ct values from TaqPath™ COVID-19 N gene with del69/70 and K417N genotypes for cases where genotyping could not determine the variant type.

Sample ID	N gene (TaqPath™ COVID-19)*	del69/70 Call	del69/70 mutant Ct value	K417N call	K417N mutant Ct value
Sample 8	19,2	Homogeneous M/M	24,5	Undetermined	Undetermined
Sample 11	18,6	Homogeneous M/M	23,7	Undetermined	Undetermined
Sample 15	Undetermined	Undetermined	Undetermined	Undetermined	Undetermined
Sample 33	34,8	Undetermined	Undetermined	Homogeneous M/M	37,7
Sample 116	20,5	Undetermined	Undetermined	Homogeneous M/M	21,8
Sample 117	22,8	Undetermined	Undetermined	Homogeneous M/M	24,3
Sample 118	32,3	Undetermined	Undetermined	Homogeneous M/M	34,9
Sample 119	31,2	Undetermined	Undetermined	Homogeneous M/M	31,4
Sample 127	15,4	Undetermined	Undetermined	Homogeneous M/M	25,5
Sample 128	21,8	Undetermined	Undetermined	Homogeneous M/M	18,4
Sample 129	35,3	Undetermined	Undetermined	Homogeneous M/M	23,9
Sample 131	30,3	Undetermined	Undetermined	Homogeneous M/M	28,6
Sample 145	33,9	Undetermined	Undetermined	Homogeneous M/M	23,7
Sample 153	26,7	Undetermined	Undetermined	Homogeneous M/M	24,5
Sample 162	28,5	Undetermined	Undetermined	Homogeneous M/M	30,6
Sample 163	28,5	Undetermined	Undetermined	Homogeneous M/M	18,1

*Initial Ct values generated during the TaqPath™ COVID-19 diagnostic assay.

c) Confirmation of genotypes by sequencing

A total number of 242 samples were sequenced from 1 to 30 November 2021, of which 77.3% (187/242) were successful. Among successful sequences, Omicron BA.1 sub-lineage (170/187; 90.9%) was dominant, followed by Omicron BA.1.1 sub-lineage (2.1%; 4/187), B.1.1.529 sub-lineage (0.5%; 1/187), Delta (4.8%; 9/187), Beta (0.5%; 1/187) and C.1.2 (0.5%; 1/187) (Fig.3.9B). Among successful genome sequences 43% (80/187) were first genotyped by PCR and the remaining 57.2% (107/187) were only sequenced.

Of the samples genotyped by PCR, sequencing confirmed that 97.5% (78/80) were Omicron BA.1 sub-lineage (21K clade), while the other 2 belonged to the B.1.1.529 lineage and Delta sub-lineage AY.19 (Fig.3.9A and 3.9B). In our study population among the identified Omicron variants, we confirmed the characteristic genotype-specific mutation K417N in 63.8% (51/80) with 20.0% (16/80) heterogeneous at position 417 (K417KN). Among the remaining 13/80

(16.3%) samples for position amino acid 417, 11/80 (13.8%) only had wild type and 2/80 (2.5%) had missing sequence data. At amino acid positions 69-70 in the spike protein 91.3% (73/80) had the del69/70 mutation, 2.5% (2/80) did not have the deletion, 5.0% (4/80) had missing sequence data and 1.25% (1/80) had missing sequence data at amino acid position 69/70 but had A67ADGV and I68IM mutations.

Thirty-two additional mutations were observed in the S protein sequences including A67V (91.3%; 73/80), T95I (92.5%; 74/80), G142D (91.3%; 73/80), del143-145 (91.3%; 73/80), N211I (76.3%; 61/80), del212 (76.3%; 61/80), R214REPE (75%; 60/80), G339D (83.8%; 67/80), S371L (66.3%; 53/80), S373P (66.3%; 53/80), S375F (66.3%; 53/80), N440K (81.3%; 65/80), G446S (66.3%; 66/80), S477N (83.8%; 67/80), T478K (95%; 76/80), E484A (66.3%; 66/80), Q493R (66.3%; 66/80), G496S (81.3%; 65/80), Q498R (81.3%; 65/80), N501Y (81.3%; 65/80), Y505H (81.3%; 65/80), T547K (92.5%; 74/80), D614G (100%; 80/80), H655Y (100%; 80/80), N679K (96.3%; 77/80), P681H (97.5%; 78/80), N764K (91.3%; 73/80), D796Y (92.5%; 74/80), N856K (91.3%; 73/80), Q954H (93.8%; 75/80), N969K (95%; 76/80) and L981F (96.3%; 77/80).

One hundred and seven samples were genotyped by sequencing only, for which we did not have residual sample to perform the genotyping. The sequence data confirmed that the dominant variant circulating was Omicron BA.1 (90.6%; 92/107) which had similar mutation profiles to the sequencing results from the samples that were genotyped by PCR. The remaining 15 belonged to Delta sub-lineages AY.45 and AY.38 (8.4%; 9/107), Beta (1.0%; 1/107), Omicron BA.1.1 (3.7%; 4/107), and C.1.2 lineage (1.0%; 1/107) (Fig.3.8A). Delta sub-lineages were detected from the 2nd to 17th November, Beta was detected 3rd November, and C.1.2 was detected 18th November (Fig.3.9B). Omicron BA.1.1 was detected from the 17th to 24th November 2021 which coincides with peak detection period of Omicron BA.1.

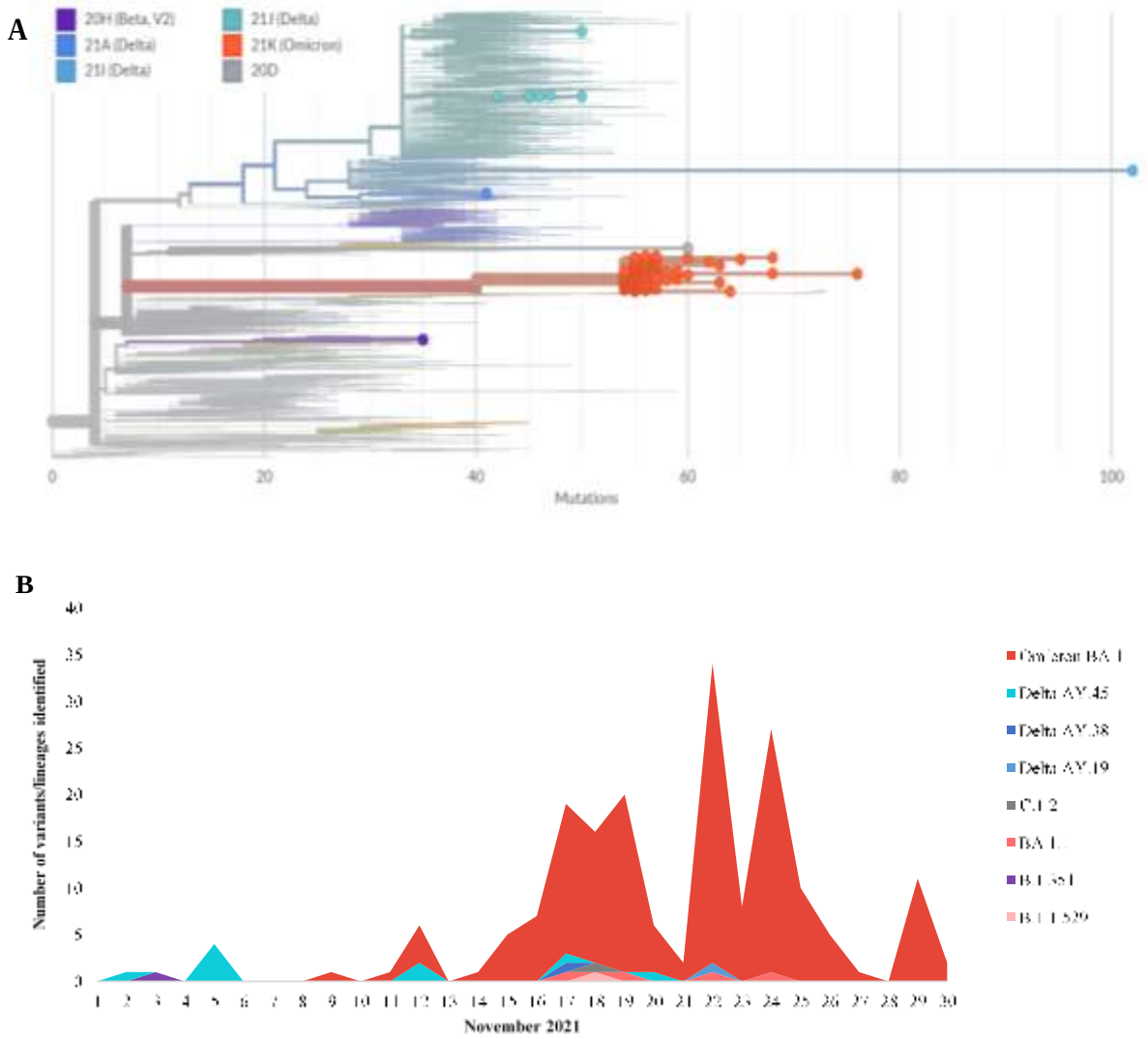

Fig. 3.9: Phylogenetic representation of sequences from samples from Virology, NHLS Gauteng. Only sequence data from November 1st to 30th, 2021 from Virology, NHLS are represented in this figure. (A) Virology, NHLS sequence data are plotted against SARS-CoV-2 global data using nextclade (https://clades.nextstrain.org). The *X*-axis represents the number of mutations present. The orange nodes: Omicron (B.1.1.529/21K) and its sub-lineages including BA.1 (21K) and BA.1.1 (21K), turquoise nodes: Delta sub-lineage AY.45 (21J), light blue nodes: Delta sub-lineage AY.38 (21I), dark blue nodes: Delta sub-lineage AY.19 (21A), purple node: Beta (20H) and gray node: C.1.2 (20D). (B) Representation of all Virology, NHLS sequence data was plotted for the period of November 1–30, 2021. From the 1st to the 14th, the number of positive SARS-CoV-2 cases was few or completely lacking and therefore sequence data is very low or absent. NHLS, National Health Laboratory Service; SARS-CoV-2, severe acute respiratory syndrome coronavirus 2. Data can be accessed from GISAID identifier: EPI_SET_230802ad.

3.4. Discussion

Alpha was initially detected in the United Kingdom in September 2020, followed by Delta in March 2021 in India (96). In our study Alpha was only observed in June 2021 at very low frequencies in the Gauteng Province. Alpha was detected based on the N501Y and del69/70 SNPs. Our study showed that Delta dominated the third wave in South Africa from June to October 2021 peaking in September 2021. Similarly, Delta replaced other circulating VOCs globally, by September 2021 (96).

The SNP PCR assays correctly classified the majority of the Delta variants in this study by targeting the L452R and P681R regions. Similarly, these SNPs were previously described as being suitable for the rapid detection of Delta and Kappa (including E484Q), where 100% of samples were confirmed as Delta/Kappa by sequencing (97). The C.1.2 lineage, a variant of interest (VOI), was observed in South Africa from June to October 2021, during the Delta wave, which was also observed in a previous South African study (87). VOI include SARS-CoV-2 lineages that have reduced neutralising antibody activity against previous infection or immunization, reduced vaccine efficacy or challenges during diagnostic testing, and a predicted increase in transmissibility or severity of disease (48). The SNP PCR could not be used to detect C.1.2 as it was designed for the detection of VOCs and misclassification of SARS-CoV-2 samples was imminent since the C.1.2 lineage harbors several mutations (N501Y, E484K, delY144, del243-244) that are also detected in Alpha, Beta and Delta variants (87). In addition, misclassification by the SNP assays may have also been a result of the low sensitivity of the E484Q and P681H assays. Omicron was first observed among the samples received in our laboratory on 9th November 2021. In November 2021, the number of positive SARS-CoV-2 cases doubled on a daily basis, with the majority of cases being attributed to Omicron BA.1 (89). Our data showed that at the beginning of the 4th wave, males and females in 25-44 year age group belonged to the highest percentage that tested positive of SARS-CoV-2, of which were predominantly infected with Omicron. Although the number of Delta cases decreased as Omicron emerged, our data shows that Omicron was highly transmissible compared to Delta which was also evident in a previous study (96).

Several SARS-CoV-2 mutations within the RBD have been previously reported to play a role in neutralising antibody escape. These mutations include K417N, G446S, E484A, Q493R, N440K, S371L and S375F, and have been identified in Omicron, which dominated South

Africa's 4th wave (98,99). When compared to previously dominant VOCs, Omicron displays the greatest diversity, with over 30 mutations including the S protein mutations del69/70, T95I, G142D, del143-145, K417N, T487K, N501Y, N655Y, N679K and P681H (88,90,91), which are also observed in the Alpha, Beta and Delta variants (75,90,92).

This study describes how the temporal use of the TaqPath™ COVID-19 real-time PCR assay, characterised by the property to identify SGTF when the S protein sequence contains the deletions of codons 69 and 70, together with PCR-based genotyping assays enhanced our ability to rapidly identify and enrich for Omicron BA.1 during its emergence and seeding of a new wave of infections in November 2021. The TaqPath™ COVID-19 assay targets the S gene of the virus, which does not amplify in the presence of the deletion at amino acid position 69/70, resulting in SGTF as described by the WHO (19). From the initial published sequences, it was evident that Omicron BA.1 contains the 69/70 deletion which is associated with SGTF (75,88,91). A recent South African study used SGTF as a proxy for Omicron BA.1 by selecting samples with SGTF and Ct-values ≤30 for the N and/or ORF1ab genes (100). However, in the latter study the SGTF was only confirmed if the testing laboratories used the TaqPath™ COVID-19 assay. In our study, the majority of samples that had SGTF, including Ct-values <38 for N and ORF1ab genes, also had the del69/70 and were confirmed Omicron BA.1 with sequencing.

Of note is the shared del69/70 in Omicron BA.1 and Alpha. In both VOCs, this deletion results in SGTF when detecting SARS-CoV-2 using the TaqPath™ COVID-19 assay. As a result, it has been suggested that the use of del69/70 and SGTF for identifying Omicron BA.1 is not possible since these characteristics will not uniquely identify Omicron BA.1 (93). However, including the delY144 SNP PCR assay we could differentiate Alpha from Omicron BA.1 (51). While the Alpha variant was sparsely identified, especially in Gauteng compared to the Western Cape province, at the beginning of South Africa's 3rd wave, it was quickly outcompeted by the dominant Delta VOC (101,102). A previous study showed that targeting the S gene could be informative for the detection of variants circulating within a region (101). However, from November through December 2021 Alpha was still circulating in other regions globally, including the United Kingdom, Canada, United States, Germany, Belgium, France, India, Georgia, Israel, and Italy, (data accessed 28 January 2022, https://microreact.org/project/1WCmB1aqTRTkfCa3S1F9wT-global-sars-cov-2-2021-10-

082021-12-29?dfc=lineage&dfo=equals&dfv=B.1.1.7&cbc=lineage). To accurately distinguish between Alpha and Omicron BA.1, we combined del69/70 genotyping with K417N (not present in Alpha) genotyping. The K417N mutation is also present in Beta, but del69/70 is not, and therefore, SGTF will not be observed. This combination of genotypes/mutations facilitated the development of a genotyping assay that can distinguish between these VOCs, despite shared mutations. We could therefore conclude that combined use of the del69/70 and K417N genotyping assays could be useful for the detection of Omicron BA.1 in laboratories that do not use the TaqPath™ COVID-19 assay.

For the samples that were genotyped but did not amplify for either the mutant or the wild type, we speculate that this may be due to poor sample quality or inappropriate storage conditions which affected downstream processing since the genotyping assay was performed on samples with Ct-values <38. In addition, mutations observed near the amino acid positions being scrutinised, such as A67ADGV and I68IM, may be present in the primer binding sites which may inadvertently have a negative impact on the genotyping assay for mutations at position 69/70. Samples for which the VOC could not be classified due to low quality and sensitivity, could have been identified as low frequency lineages with mutations present other than those found in VOCs. Genotyping data was not collected on 20 November since the number of samples that tested positive for SARS-CoV-2 was relatively low on the day and residual samples were not available for testing. However, these samples were selected for sequencing, from which we were able to confirm that the Omicron BA.1 was present. Importantly, the genotyping assay successfully identified mixed infections at some amino acid positions and sequencing also highlighted mixed amino acids at positions close to the SNPs of interest.

The first limitation of this study was the limited availability or lack of suitable controls (wildtype and mutant) at the time to assess the sensitivity of the genotyping assays to detect each genotype. We, therefore, utilised known positives as controls in addition to the commercially available controls in cases where commercial controls were of poor quality. Secondly, the results for the E484K/Q and P681H/R SNPs from the PCR assay were challenging to interpret due to interference from viral sequence with an adjacent mutation (103) and therefore resulted in misclassification of VOCs. In such instances, the introduction of additional SNPs specific to each VOC would be beneficial for the detection of VOCs. Thirdly, the sample size was very small in comparison to the number of positive SARS-CoV-2 cases

detected during the period selected in November 2021. Therefore, this may have skewed the data distribution (age, gender, patient status) of individuals for which samples had SGTF. Fourth, the short period we focused on covered Omicron BA.1, however from December 2021 Omicron BA.1.1, BA.2, and BA.3 started circulating in South Africa (92). Omicron BA.2 does not have the del69/70 while BA.1.1 and BA.3 do. Consequently, this implies that we will not be able to differentiate between Omicron BA.1, BA.1.1, and BA.3 as they share both the del69/70 and K417N mutations; while Omicron BA.2 has the K417N which may not be easily distinguishable from Beta. Among our study sample, BA.3 was not identified, and overall only 1.0% (15/1579) of BA.3 was reported from South Africa, in November 2021 (92). In a study from Japan, their genotyping assay targeted the G339D mutation which is present in all 3 Omicron sub-lineages and the T547K mutation which easily differentiates Omicron BA.1 from BA.2 and BA.3 (104). These mutations are also present among our study samples and their inclusion may substantially improve one's ability to differentiate between VOCs in countries, apart from South Africa. Another study described specific mutations to discriminate between BA.1.1, BA.2 and BA.3 (105). BA.1.1 can be uniquely identified with the R346K mutation; BA.2 with the T19I, del24-26, A27S, V213G, T376A, and R408S; and BA.3 with the del216, none of which were identified among samples from this study (105). The del69/70 seems to occur in specific lineages, during unfavourable conditions, to assist with viral functionality such as changing the conformation of the NTD loop, optimising spike cleavage and enhancing infectivity (50). Toggling between the presence or absence of the del69/70 mutation in SARS-CoV-2 lineages, as they emerge in succession, may indicate that these 2 amino acid positions may be important for complete functionality of the S protein and therefore it is lost due to immune pressure in some emergent lineages when conditions are favorable. Thus, it may be important to consider its impact as part of a SARS-CoV-2 S vaccine immunogen to target these specific lineages.

3.5. Conclusion

Sequencing remains the gold standard for the detection of novel VOCs. However, in low to middle-income countries, which may lack the expertise, facilities or access to resources to perform NGS, the use of genotyping assays coupled with routine surveillance may assist with the detection of Alpha, Beta, Delta and Omicron BA.1. However, additional assays are required to differentiate between lineages that harbor mutations like those found in VOCs, and to differentiate between Omicron BA.1, BA.2, and BA.3 sub-lineages. We have demonstrated

that the identification of SGTF and/or the del69/70 mutation together with the K417N mutation will provide a rapid, cost-effective, and reliable method for the detection of the Omicron BA.1. This study showed that a combination of key SNPs allowed the rapid identification of an emerging variant at the time and similar approaches could be use in future to confirm or refute re-emergence of these VOCs. However, we acknowledge the need for additional assays to target specific unique mutations that will differentiate one lineage from another based on the identification of unique mutations, without the use of sequencing. Identification of the SGTF property of the TaqPath™ COVID-19 assays resulted in the development of genotyping assays focused on mutations that rapidly emerge in the S gene to identify variants. However, selection of more conserved gene regions that are still able to correctly identify major VOCs or VOIs rapidly may overcome the temporality of current assays. The latter is necessary to prevent laboratories from continuously changing the assays used due to targeted genes not annealing appropriately because of emerging mutations detected in highly variable regions. Targeting conserved regions may be of greater benefit to characterise mutation profiles of current and future variants or lineages of interest as they emerge. Ultimately, we observed that genotyping as well as sequencing identified some mixed infections. Therefore, we subsequently aimed to investigate the extent of mixed SARS-CoV-2 infections or viral quasispecies in study participants.

CHAPTER 4

SARS-CoV-2 spike protein diversity at an intra-host level, among SARS-CoV-2 infected individuals in South Africa, 2020 to 2022

Reference (Appendix B): Subramoney, K., Mtileni, N., Davis, A., Giandhari, J., Tegally, H., Wilkinson E., Naidoo, Y., Ramphal, Y., Pillay, S., Ramphal, U., Simane, A., Reddy, B., Mashishi, B., Mbenenge, N., de Oliveira, T., Fielding, B. C., Treurnicht, F. K. (2023). SARS-CoV-2 spike protein diversity at an intra-host level, among SARS-CoV-2 infected individuals in South Africa, 2020 to 2022. *Plos One*, *18*: e0286373. https://doi: 10.1371/journal.pone.0286373

4.1. Introduction

The primary function of the coronavirus spike glycoprotein (S protein) is to facilitate virus attachment and infection of host cells (106). Variations occurring within its amino acid (aa) sequence could influence the interaction with the host receptor, viral pathogenesis, transmissibility, and infectivity (106). Similarly, since the emergence of SARS-CoV-2 in 2019, the S protein has evolved rapidly, resulting in diverse lineages of which some manifested as variants of concern (VOC) (107). Globally, waves of infections were mainly driven by VOCs of the Alpha, Beta, Delta, Gamma, and Omicron lineages. In South Africa, the SARS-CoV-2 Wuhan-Hu1 wild-type strain dominated the first coronavirus disease of 2019 (COVID-19) wave, the second and third waves were dominated by Beta and Delta VOCs respectively, and the fourth and slowly emerging fifth waves were dominated by Omicron BA.1, BA.2, BA.4 and BA.5 lineages (43,47,75,100,108)). The continuous emergence of new SARS-CoV-2 lineages over time and the reversion of mutations together with recombination within the SARS-CoV-2 genome, predominantly in the S protein, could result in strains becoming increasingly resistant to antibody-mediated neutralisation. This highlights the need for characterisation of intra-host SARS-CoV-2 S protein diversity.

Several studies have shown that studying intra-host SARS-CoV-2 S protein diversity leads to further insight and understanding of SARS-CoV-2 transmission dynamics and could be

important for the improvement of SARS-CoV-2 vaccines (109–112)). The Alpha, Beta, and Delta variants showed escape from neutralising antibodies induced by the dominant vaccine immune epitopes in the S protein. Circulation of these variants also resulted in varying efficacy (50-90%) of licensed SARS-CoV-2 vaccines, which comprised the Wuhan-Hu1 derived S protein. In addition, spatiotemporal analysis of intra-host single nucleotide variants (iSNVs) showed increased genetic diversity following symptomatic infections. In South Africa, iSNVs analysis coupled with bottleneck transmission, was successfully used to determine the transmission dynamics of SARS-CoV-2 in nosocomial outbreaks (112). The study identified patients with similar iSNV profiles together with the evolution of SARS-CoV-2 infections. A limitation of the study was that intra-host variants were analysed at a nucleotide level with limited focus being placed on the impact of variants on possible immune escape based on aa variations.

Although SARS-CoV-2 variants that dominated the waves from 2020 to 2022 in South Africa have been described, the frequency of heterogeneous infections and heterogeneity within each defined lineage and VOC lineage is not well defined. Therefore, the factors that contribute to the emergence of dominant variants, including factors that could affect variant persistence or eventual replacement is still unclear. Intra-host diversity studies will assist with identifying and characterising the frequency of mutations and possible heterogeneity within the SARS-CoV-2 S protein. This will broaden our understanding of the virus-host adaptations, virulence factors, and pathogenicity and potentially provide information for more effective vaccine design. In this study we aimed to determine the diversity of SARS-CoV-2 S protein among SARS-CoV-2 infected individuals and if heterogeneity has an impact on intra-host adaptations.

4.2. Methods

4.2.1. Study population

Study population is described in 3.1. and included samples from June 2020 to May 2022. CMJAH received SARS-CoV-2 samples within the Gauteng province, as well as referrals from the Eastern Cape, Free-State KwaZulu-Natal, and Western Cape provinces of South Africa. This study contributed towards the national surveillance for SARS-CoV-2 in South Africa for which formal patient consent was not required. The study was approved by the University of Witwatersrand Human Research Ethics Committee (M210119) and all participants' data was anonymised and presented in an aggregated form.

4.2.2. Study samples and sample size

Study samples were included as described in 3.2. Residual samples were stored for further characterisation of the infecting strains. Approximately 50-80 SARS-CoV-2 positive samples were randomly selected per week for genomics surveillance from 2020-2022. We strived to include at least 10 samples from each district from which samples were collected per day, for detection of single nucleotide polymorphisms (SNP) in the S protein.

4.2.3. SARS-CoV-2 diagnosis

Routine diagnostic SARS-CoV-2 extraction and real-time PCR was performed following the method described in 3.4.1.

4.2.4. Genotyping to detect single nucleotide polymorphisms (SNPs) associated with specific amino acid mutations in the S protein

A PCR-based genotyping method which targets SNPs across the S protein of several mutations associated with several VOCs (refer to method 2.4.3)

4.2.5. Next-generation sequencing of SARS-CoV-2 strains

Samples were sequenced as described in 3.4.4. Genome assembly was done using genome detective (https://www.genomedetective.com/app/typingtool/virus/) by KRISP and Exatype (https://sars-cov-2.exatype.com/) online tools. Consensus sequences were analysed using Nextstrain (https://clades.nextstrain.org) and Pangolin (https://pangolin.cog-uk.io/) online tools for variant assignment.

4.2.6. Variant mutation analysis and identification of heterogeneous infections

FASTQ reads generated from sequencing were analysed using the COVID-19 workflows with galaxy.eu (15). To generate allele frequencies for major and minor variants at the aa level across the S protein: for Illumina pair-end reads we ran the “COVID-19: variation analysis on ARTIC PE data (V0.5)” together with the “COVID-19: variation analysis reporting” workflow; and the “COVID-19 variation analysis of ARTIC ONT (V0.3.1)” workflow was used for MinION single reads. The final output from galaxy.eu indicated if the sequence data passed the quality checks (indicated as “PASS”). Only the sequences that passed were included in this analysis. Heterogeneous mutations were identified if the allele frequency ranged from 0.1 to 1.0 for more than 1 mutation.

4.2.7. Data analysis

Descriptive analysis was done across different age groups (<5, 5-14, 15-24, 25-44, 45-60, and >60 years), gender, province, and specimen collection site (individuals that sought community screen and test services, healthcare workers staff clinic, in-patient wards, out-patient wards, mortuary). Intra-host sequences and genotyping data were analysed to determine the proportion of individuals infected with homogeneous and/or heterogeneous variant populations. A box-whisker analysis of heterogeneous aa mutations was used to display the intra-host frequency of heterogeneous SNPs across the S protein. A time-tree was generated by running our sequence dataset through a Nextstrain custom build (https://github.com/nextstrain/ncov) and the most prevalent heterogeneous mutation positions were selected for analysis. The positions with heterogeneous mutations were compared to the global prevalence of mutations at the same positions from 2020 through to 2022, using GISAID CoVServer (https://www.gisaid.org/epiflu-applications/covsurver-mutations-app/; (16)).

4.3. Results

4.3.1. Demographic characteristics of the study population

Over the study period, a total of 667 269 respiratory samples were tested, of which 107 495 (16.1%) tested positive for SARS-CoV-2. A total of 2982/107 495 (2.8%) SARS-CoV-2 positive samples were analysed to assess the proportion with mixed infections. Of these, 20.3% (601/2982) were analysed by SNP PCR only, 68.2% (2034/2982) were sequenced and 11.6% (347/2982) were both were analysed by SNP PCR and sequenced.

Just under 50% of SARS-CoV-2 positive samples belonged to adults in the age group 25-44 years (1351/2982; 45.3%), followed by older adults aged 45-60 years (594/2982; 20.0%) (Table 4.1). The proportion of females (57.9%; 1726/2982) was greater than males (39.6%; 1181/2982). Most samples were from individuals residing in Gauteng province (2830/2982; 95.0%). Individuals that were out-patients and in-patients represented 11.5% (344/2982) and 8.1% (242/2982) respectively, while 77.9% (2324/2982) of individuals were from community screen and test (CST) (Table 4.1). ARV clinic and cardiac clinic attendees represented 25.0% (86/344) and 0.6% (2/344) of outpatients respectively.

Table 4.1. Demographics of individuals with SARS-CoV-2 included in this study (N=2982)

Demographics	***Genotyped (n/N, (%))***
Age group (years)	
<5	35/2982 (1.2)
5-14	243/2982 (8.1)
15-24	444/2982 (14.9)
25-44	1358/2982 (45.5)
45-60	594/2982 (20.0)
>60	220/2982 (7.4)
Unknown	86/2982 (2.9)
Gender	
Female	1726/2982 (57.9)
Male	1181/2982 (39.6)
Unknown	75/2982 (2.5)
Province/ District	
Eastern Cape	147/2982 (4.9)
Free-State	1/2982 (0.0)
Gauteng	2830/2982 (95.0)
KwaZulu-Natal	3/2982 (0.1)
Western Cape	1/2982 (0.0)
Specimen collection site	
Community screen and test	2324/2982 (77.9)
Healthcare workers staff clinic	59/2982 (2.0)
In-patient ward	242/2982 (8.1)
Out-patient ward	344/2982 (11.5)
Mortuary	13/2982 (0.4)

Five distinct SARS-CoV-2 waves of infection were observed during the study period from 01 June 2020 through 31 May 2022, which were indicative of the 1st (SARS-CoV-2 Wuhan-lineage dominant), 2nd (Beta dominant), 3rd (Delta dominant), 4th (Omicron BA.1 and BA.2 dominant) and 5th (Omicron BA.4 and BA.5 dominant) SARS-CoV-2 waves of infection (Figure 4.1). The SARS-CoV-2 detection rate (Figure 4.1a) ranged from 10.0-34.4% during the 1st wave (28 June to 14 August 2020), 11.0-38.8% in the 2nd wave (14 December 2020 to 24 January 2021), 10.2-44.0% in 3rd wave (31 May to 27 August 2021), 12.1-57.7% during 4th wave (27 November 2021 to 07 January 2022) and in the 5th wave (16 April to 21 May 2022)

it ranged from 11.1-23.8%. Figure 4.1b highlights (green bars) the epiweeks where samples were genotyped by SNP PCR and or sequencing.

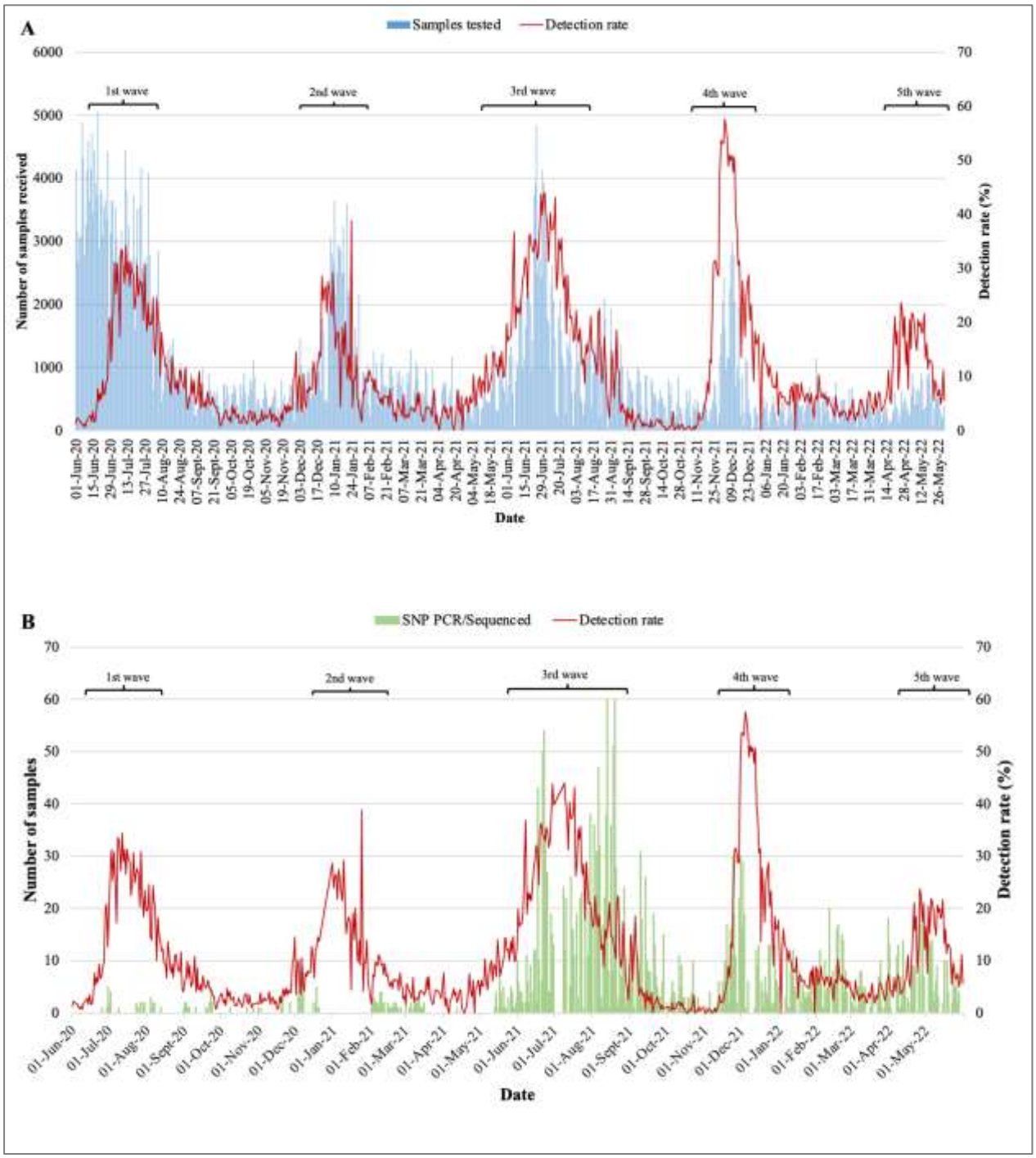

Fig. 4.1: SARS-CoV-2 detection rate among respiratory samples tested from June 2020 to May 2022. A) Blue bars represent the number of samples received for SARS-CoV-2 diagnostics and the red line graph represents the detection rate over time.

4.3.2. Frequency and characteristics of heterogeneous SNPs across the S protein

Figure 4.2 highlights all the SNPs in the SARS-CoV-2 S protein included in this analysis with allelic discrimination plots for SNPs shown in Figure 4.3. Mutations were primarily observed at positions L452 (69.6%; 652/937) [Figure 5.2e and 5.3G], N501 (5.4%; 51/937) [Fig 5.2h and 5.3H], D614G (90.1%; 299/332) [Fig 5.2i and 5.3I], and P681 (82.3%; 771/937) [Fig 5.2k and 5.3K].

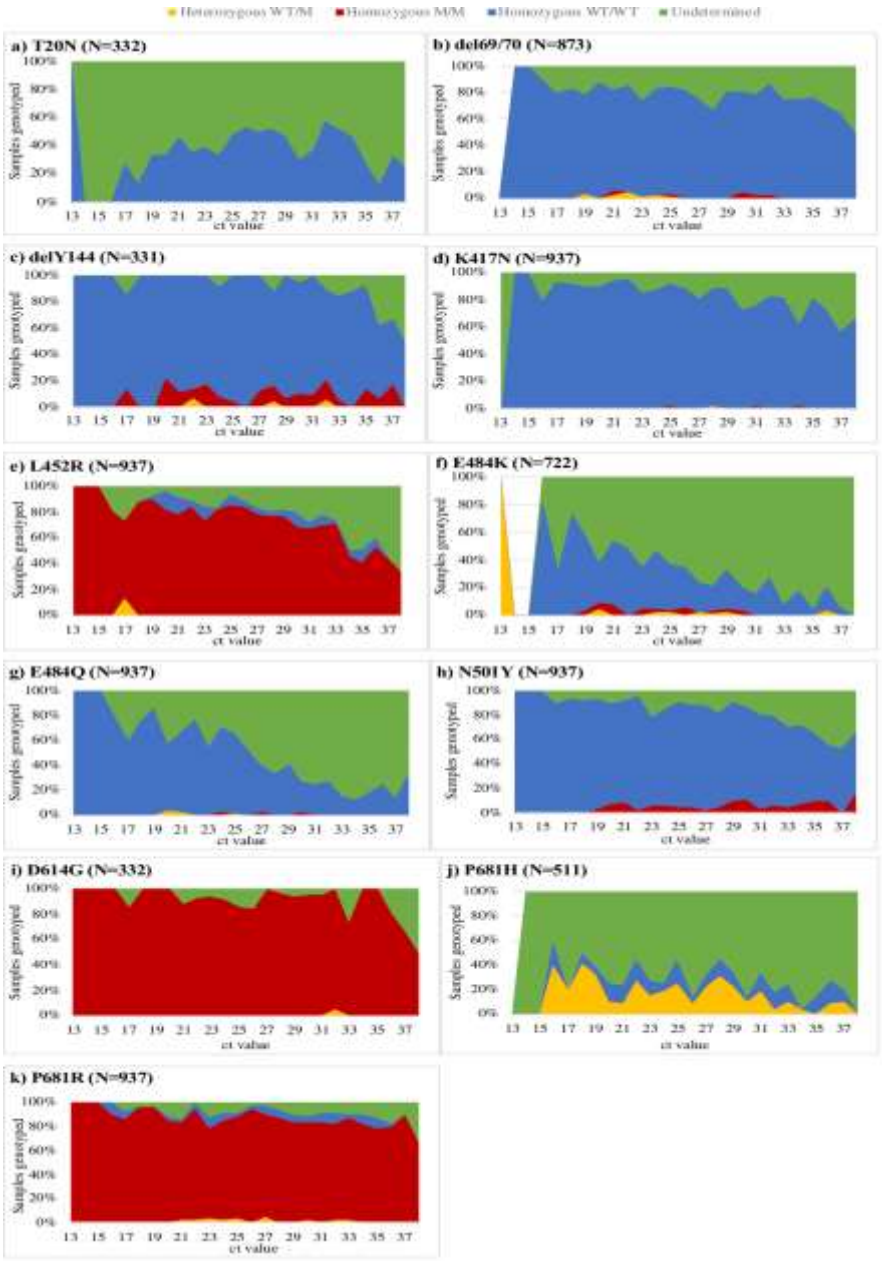

Fig. 4.2: Distribution of Ct-values for each SARS-CoV-2 mutations tested. The x-axis displays the Ct-values from SARS-CoV-2 diagnostic testing compared to the genotyping results for each mutation and/or wild type detected for each SNP assay. Samples that were undetermined for a genotyping assay are associated with the assays inability to detect the mutant and/or wild type.

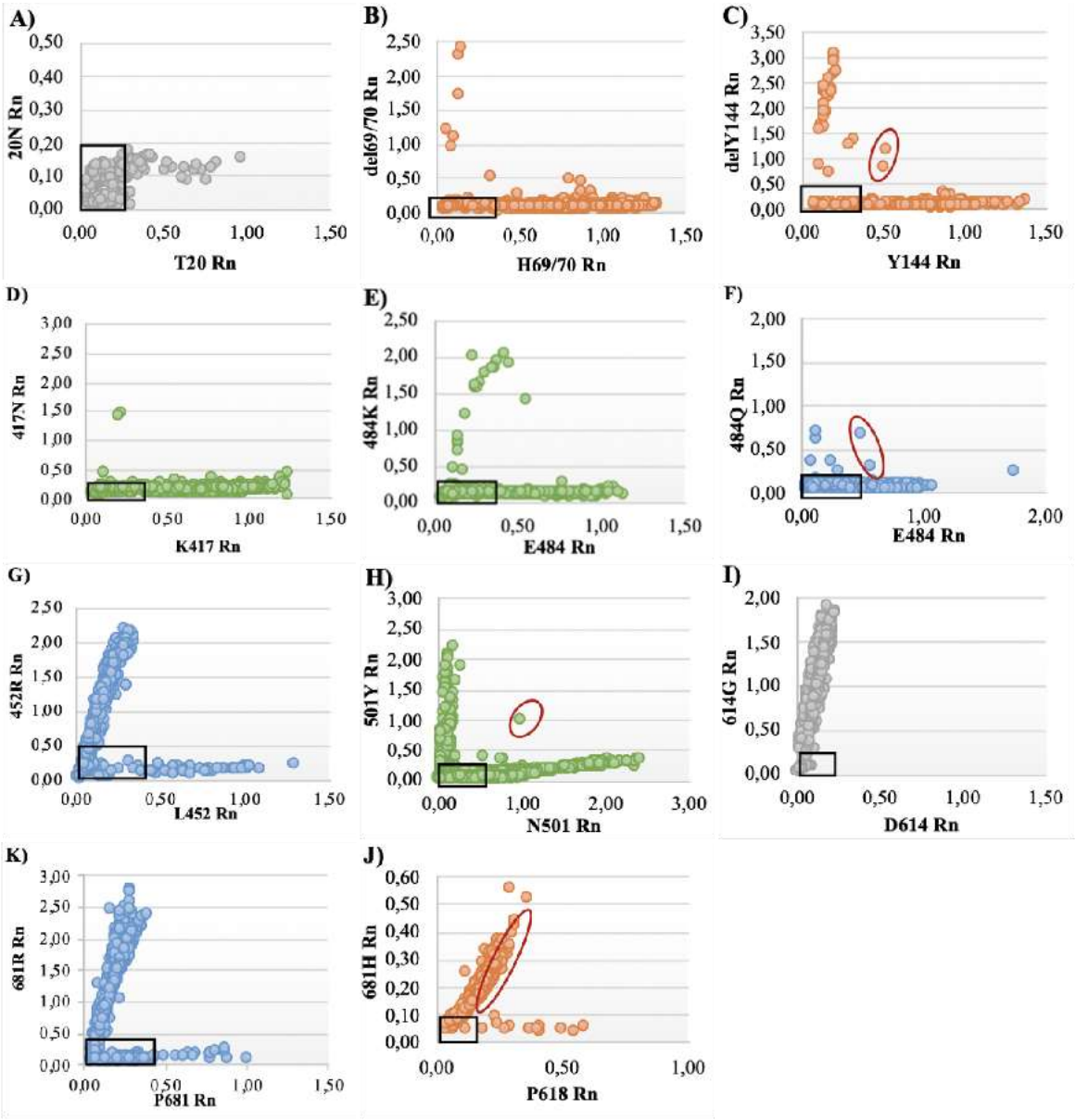

Fig. 4.3: Allelic discrimination plots of mutants, wild type and heterogeneous SNPs (mutant/wild type). Black block indicates the samples that were negative for the mutant and/or wild type for each SNP (<0.50). Red ovals indicate the samples that were heterogeneous (both wild type and mutant present). X-axis represents the allelic distribution of mutants and y-axis represents the allelic distribution of the wild types.

Heterogeneity (mix of wild type and mutant alleles at SNP positions indicated by *) was observed for 50/948 (5.3%) cases for only 4 of the 11 SNPs (Table 4.2). These included the E484Q (2/948; 0.2%), N501Y (1/948; 0.1%), P681H (45/511; 8.8%) and delY144 (2/331; 0.6%). P681R/L452R/P681H* profile of mutations was primarily identified from 17 July to 20 August 2021 and 13 September 2021, whereas the P681R/L452R/P681H*/D614G mutation

profile was observed from 9 to 17 September 2021 (Fig. 4.4). Samples with heterogeneous SNPs that were sequenced confirmed heterogeneity at positions 144 (YY144Y, AF=0.2) and 484 (E484Q, AF=0.4). Sequencing was performed for P681H and N501Y, however we could not confirm the presence of the SNPs due to low quality sequence data in these regions.

Table 4.2: Summary of SNP combinations with heterogeneous amino acids (n=50).

Combination of SNPs (N=50)	No. of samples n (%)	Heterogeneous positions	Possible VOC/VOI	Sequencing confirmed variant/lineage	Sequencing confirmed heterogeneous amino acid
L452R/**delY144***/D614G	1 (2)	del144/Y144	Delta	Poor quality	No[1]
P681R/L452R/**delY144***/D614G	1 (2)	del144/Y144	Delta	Delta	del144/Y144
P681R/L452R/E484Q/**P681H***/D614G	1 (2)	P681H	Delta	Delta	No
P681R/L452R/**E484Q***	1 (2)	E484Q	Delta	Delta	E484Q
P681R/L452R/**E484Q***/D614G	1 (2)	E484Q	Delta	Delta	E484Q
P681R/L452R/**N501Y***	1 (2)	N501Y	Delta	Poor quality	No[1]
P681R/L452R/**P681H***	20 (40)	P681H	Delta	Delta	No
P681R/L452R/**P681H***/D614G	20 (40)	P681H	Delta	Delta	No
P681R/L452R/**P681H***/N501Y	2 (4)	P681H	Delta	Poor quality	No[1]
P681R/L452R/**P681H***/del6970	1 (2)	P681H	Alpha/Delta	Delta	No
P681R/**P681H***	1 (2)	P681H	Alpha/Delta	Poor quality	No[1]

*Heterogeneous amino acids. Poor quality: a VOC or lineage could not be determined due to poor sequence. [1]Sequence data absent at the amino acid position of interest.

We were unable to determine the SNP for 2.0% (21/948) of samples on any of the SNP assays. In addition, undetermined results (negative or no amplification for a SNP) were obtained for 67.5% (345/511) of samples analysed with the P681H SNP assay, 66.3% (479/722) for E484K, 59.0% (196/332) for T20N, 53.2% (505/948) for E484Q, 20.7% (196/948) for L452R, 20.5% (179/873) for del69/70, 17.2% (163/948) for N501Y, 16.0% (152/948) for K417N, 11.2% (27/332) for D614G, 8.1% (77/948) for P681R, and 7.3% (24/331) for delY144. This is despite the PCR diagnostic results having an average Ct value of 25.

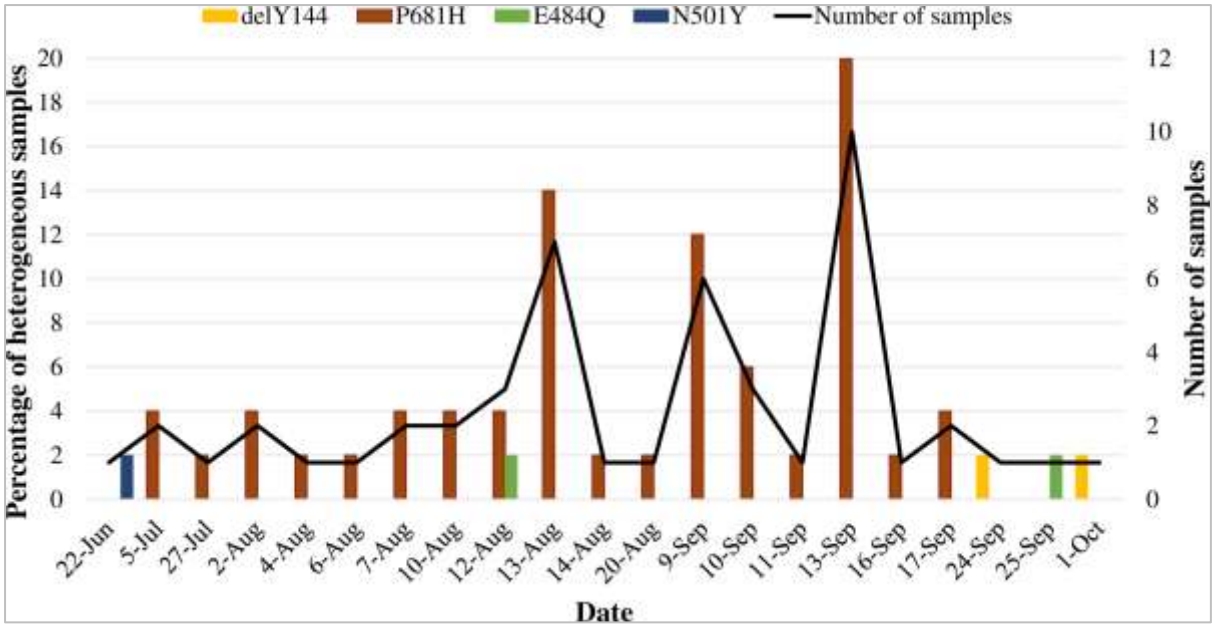

Fig. 4.4: Frequency of the heterogeneous SNPs identified from June to October 2021 (N=50). X-axis represents the date during which the heterogeneous SNP was identified. Y-axis: the line graph represents the total number of samples (N) that were observed with a heterogeneous SNP, and the bar graphs represent the percentage of samples in which the SNP was observed on each day.

4.3.3. Sequence-based analysis of heterogeneity in S protein

Close to 9% (210/2381; 8.8%) of South African SARS-CoV-2 cases sequenced displayed heterogeneity at one or more SNP positions. A total 7 aa positions (19, 80, 191, 371, 484, 681, 950) across the S protein, with AF scores ranging from 0.1 to 1.0, were identified with multiple heterogeneous mutations (Fig 4.5). At position 19, T19IR mutations (3/210; 1.4%) had an AF of 0.2-0.7 for 19I (C>T, 21618) and 19R (C>G, 21618), whereas T19KR (1/210; 0.5%) had an AF of 0.4 for 19K (C>A, 21618) and 0.5 for 19R. At position 371, the majority (92.9%; 195/210) of cases with heterogeneous mutations had a high AF >0.8-0.9 for 371F (C>T; 22674) and 371P (T>C, 22763) (Fig 4.5). Lastly, at position 484 among 1.9% (4/210) cases, E484KQ was present with an AF of 0.1 for 484K and 0.4 for 484Q, E484AK had an AF of 0.2 for 484A (A>C, 23013) and 0.7 for 484K (G>A, 23012), and E484AQ had an AF of 0.5 for 484A and 0.4 for 484Q (G>C, 23012).

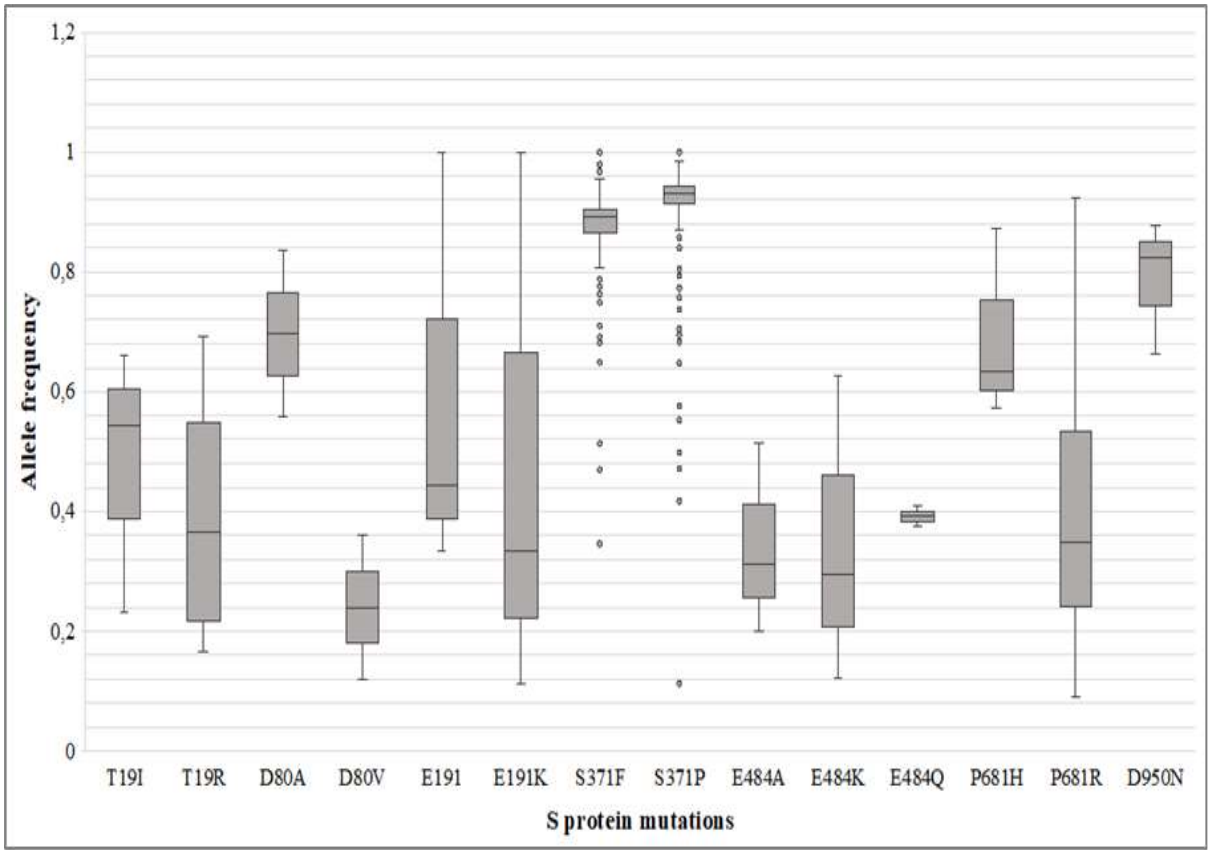

Fig. 4.5: Box and whisker plot showing the mutant allele frequencies (AF) at amino acid positions within the S protein that have more than 1 amino acid change, based on sequencing. Circles represent the outliers, line within the box represents the mean AF. Mutations (T19K, P681S D950, D950K, and D950R) that only have a single AF value were excluded from the figure.

Of the 3 samples with T19IR heterogeneous mutations, each sample was classified as Delta (June 2021), Omicron BA.21.5 (March 2022), and unassigned (August 2022) (Fig 4.6A and B). The sample with T19KR was identified in October 2021 but could not be classified (unassigned) due to lower sequence quality at other regions of the genome. S371FP heterogeneous mutations were identified from June 2021 through to May 2022 of which 1/195 (0.5%) was classified as Delta. The S371FP heterogeneous mutation rapidly increased from the Delta (June to August 2021) to the Omicron waves and dominated during the Omicron BA.1/BA.2 wave from

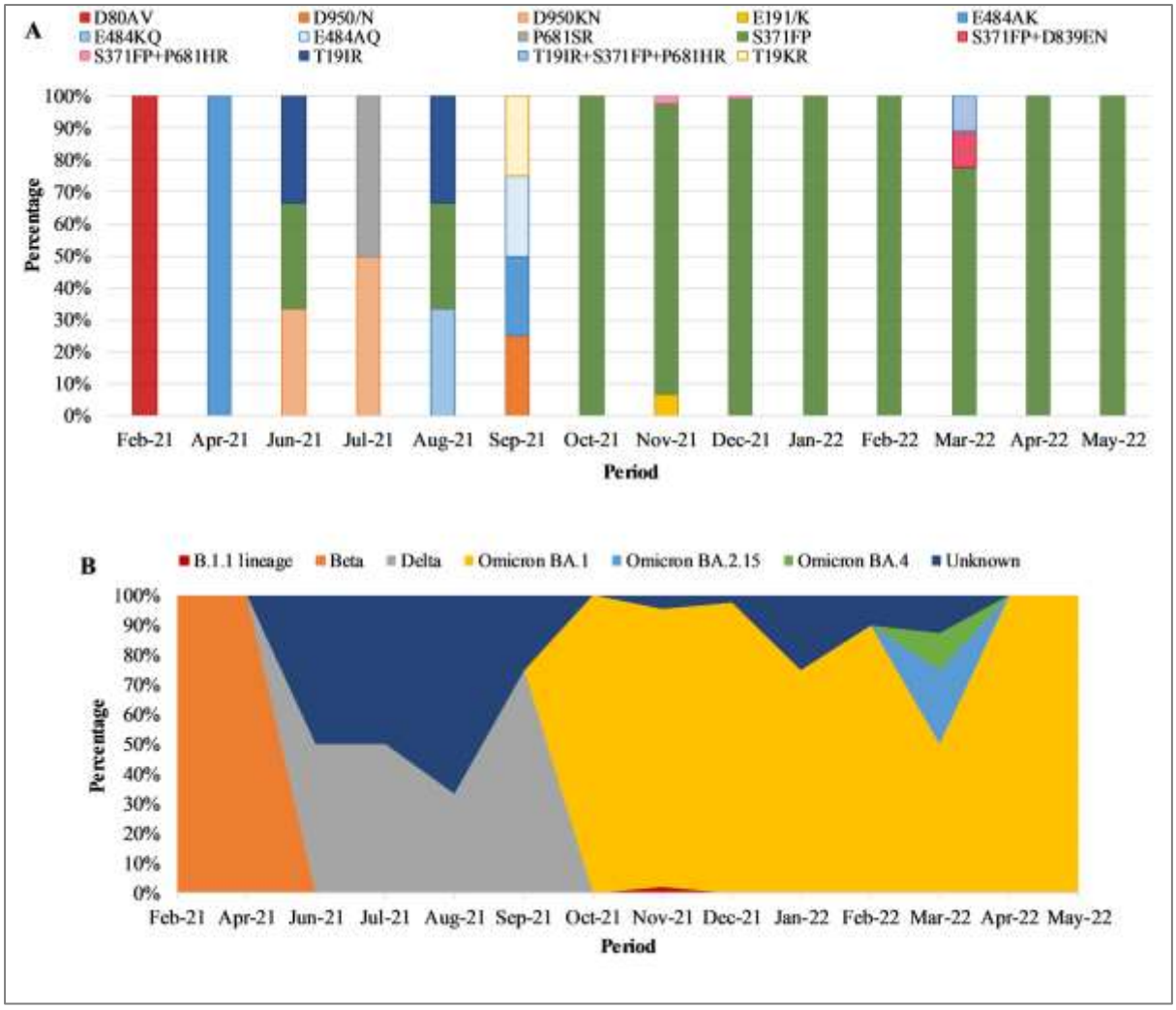

Fig. 4.6: Frequency of the heterogeneous mutations (A) and VOC/lineages (B) identified from sequence data over time. 4a) X-axis represents the date during which the heterogeneous SNP was identified, and the y-axis represents the percentage of samples that were observed with a heterogeneous SNP. 4b) X-axis represents the period during which the VOC/lineages were identified, and the y-axis represents the percentage of samples with a specific VOC/lineage. Omicron BA.1, BA.1.1, BA.1.10, BA.1.13, BA.1.14, BA.1.17, BA.1.17.2, BA.1.18, BA.1.19 and BA.1.21 were grouped as Omicron BA.1.

November 2021 (41/195; 21.0%) to December 2021 (122/195; 62.6%) (Fig 4.7). From January 2022 to May 2022 the S371FP mutation was observed but at lower frequencies from 2.1% (4/195) to 5.1% (10/195) from the Omicron BA.1/BA.2 wave to the Omicron BA.4/BA.5 wave. The majority of the S371FP was identified among 92.8% (181/195) of cases classified as Omicron BA.1 lineage strains of which 50.3% (98/195) were BA.1, 11.8% (23/195) were BA.1.1, 11.3% (22/195) were BA.1.21, and the remaining 19.5% (38/195) were BA.1.10, BA.1.13, BA.1.14, BA.1.17, BA.1.17.2, BA.1.18, and BA.1.19. S371FP was also detected in 1.0% (2/195) Omicron BA.2.15, 0.5% (1/195) Omicron BA.4, and 0.5% (1/195) B.1.1 lineages.

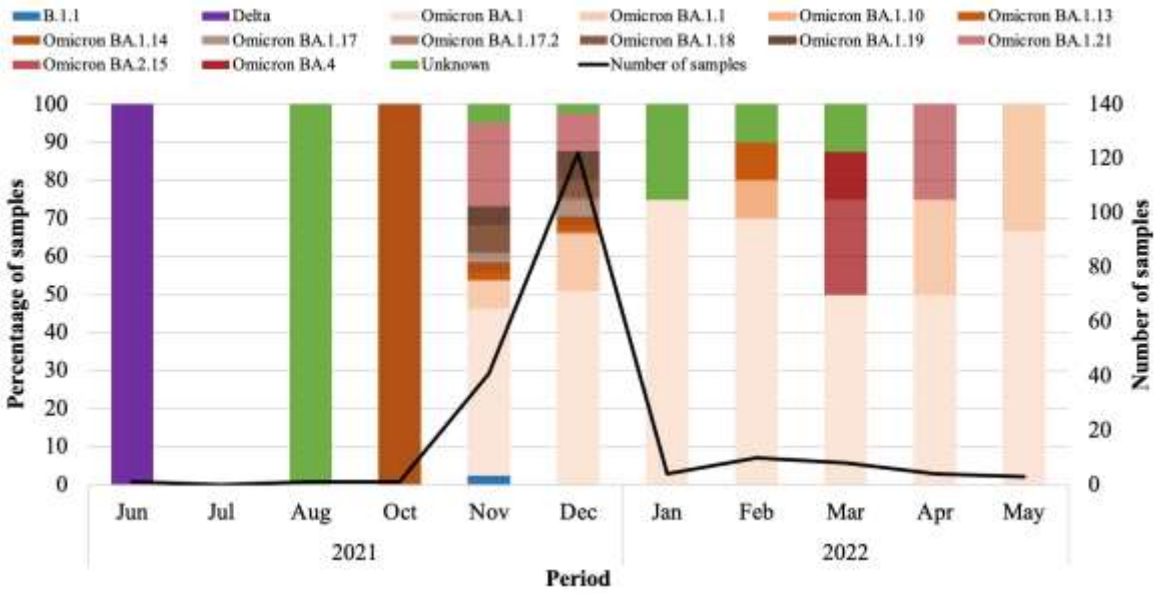

Fig. 4.7: Prevalence of samples with S371FP over time. The black line graph represents the total number of samples with S371FP heterogeneity each month. The bar graphs represent the SARS-CoV-2 VOCs and lineages observed with the S371FP heterogeneous mutation from 2021 to 2022.

The lineage or VOC could not be determined for 4.6% (9/195) of the sequences due to low sequence data quality in some regions of the S gene (Fig 4.6B). The majority (146/195; 74.9%) of cases with the S371FP were from individuals that sought CST, while the remaining cases (25.2%) were from deceased individuals (1/195; 0.5%), HCW (4/195; 2.1%), in-patients (24/195; 12.3%) or out-patients (20/195; 10.3%) [Table 4.3]. Individuals aged 5-14 (43/195; 22.1%), 15-24 (48/195; 24.6%), and 25-44 (58/195; 29.7%) were the predominant age groups with S371FP (Table 4.4). Three percent (5/195) of out-patients visited ARV clinics. Two cases with multiple heterogeneous mutations which included S371FP/P681HR were classified as B.1.1 and Omicron BA.1 lineages; detected in November 2021 and December 2021 from CST, respectively. Whereas T19IR/S371FP/P681HR and S371FP/D389EN mutation profiles were identified in March 2022 among sequences from 2 cases classified respectively as Omicron BA.2.15 and Omicron BA.4 from CST (Fig 4.6 and Table 4.3). E484KQ (1/4; 25%) was observed in an out-patient from an ARV clinic in August 2022 and classified as Delta (Fig 4.6 and Table 4.3). E484AK (2/4; 50%) and E484AQ (1/4; 25%) were identified among individuals that sought CST. E484AK was observed among samples classified as Beta and Delta in April 2021 and October 2021, respectively. While E484AQ (1/4; 25%) was identified in October 2021, and classified as Delta.

Table 4.3: Patient status for cases where heterogeneity in SARS-CoV-2 S protein was observed (N=210)

Amino acid positions	Heterogeneous mutations	Patient status												
		ARV clinic (%)	Community screen and test (%)	Deceased Post-mortem (%)	HCW (%)	In-patient (%)	In-patient gastro (%)	In-patient labour ward (%)	In-patient neurology and dermatology (%)	In-patient oncology (%)	In-patient paediatric (%)	In-patient paediatric cardio (%)	Out-patient (%)
D80 (N=2)	D80AV		1 (50)										1 (50)
D950 (N=3)	D950/N		1 (33.3)										
	D950KN		2 (66.7)										
E191 (N=3)	E191/K		3 (100)										
E484 (N=4)	E484AK		1 (25.0)										
	E484KA		1 (25.0)										
	E484KQ	1 (25)											
	E484QA		1 (25.0)										
P681 (N=1)	P681SR		1 (100)										
S371 (N=195)	S371FP	5 (2.6)	146 (74.9)	1 (0.5)	4 (2.1)	8 (0.4)	1 (0.5)	1 (0.5)	1 (0.5)	2 (1.0)	6 (3.1)	1 (0.5)	15 (7.7)
	S371FP+D839EN		1 (0.5)										
	S371FP+P681HR		1 (0.5)	1 (0.5)									
	T19IR+S371FP		1 (25.0)										
T19 (N=4)	T19IR+S371FP+P681HR		1 (25.0)										
	T19IR		2 (50.0)										
	T19KR		1 (25.0)										

Table 4.4: Age groups for cases for which heterogeneity in S protein was identified

Amino acid positions	Heterogeneous mutations	Age group (years)						
		<5 (%)	5-14 (%)	15-24 (%)	25-44 (%)	45-60 (%)	>60 (%)	Unknown (%)
D80 (N=2)	D80AV				2 (100)			
D950 (N=3)	D950/N				1 (33.3)			
	D950KN			1 (33.3)	1 (33.3)			
E191 (N=3)	E191/K				2 (66.7)	1 (33.3)		
E484 (N=4)	E484AK				1 (25)			
	E484KA				1 (25)			
	E484KQ				1 (25)			
	E484QA					1 (25)		
P681 (N=1)	P681SR		1 (100)					
S371 (N=194)	S371FP	2 (1.0)	43 (22.2)	48 (24.7)	58 (29.9)	23 (11.9)	13 (6.7)	3 (1.5)
	S371FP+D839EN					1 (7.1)		
	S371FP+P681HR		1 (7.1)					1 (7.1)
T19 (N=4)	T19IR		1 (25.0)		1 (25.0)			
	T19IR+S371FP+P681HR					1 (25.0)		
	T19KR				1 (25.0)			

4.3.4. Mutations at dominant heterogeneous SNPs positions compared to global data

The consensus sequence data was used to compare our data to global data. Globally, from January 2020 through to May 2022, multiple aa changes were seen at positions 19, 371, and 484 among 11 140 058 SARS-CoV-2 genome sequences uploaded on GISAID (accessed 22 June 2022). At aa position 19, the T19I mutation was detected in 15.8% of global sequences (January 2022 through May 2022) in Omicron BA.2 (21L clade) and T19R was identified among 38.3% of Delta lineages from March 2021 through March 2022, similarly identified in our study sequences (Figure 4.8A). At position 371, the S371F mutation was globally dominant (15.21%) and identified from December 2021 to May 2022 among Omicron BA.2 lineages, while our study identified S371F primarily among Omicron BA.4 (22A clade) which clustered out from Omicron BA.2 from June 2021 to May 2022. S371P is not yet reflected since it was only observed in June 2022 in 396 global genomes and was not observed from our data due to

S371F being the dominant mutation from the consensus sequences (Figure 4.8B). S371L was identified among our consensus sequences but was not observed among the heterogeneous mutations. At position 484, E484A was identified in the majority (34.1%) of global sequences from November 2021 to May 2022 and classified as Omicron BA.1 and BA.2 lineages, similar to our findings but also identified among Omicron BA. 4 and BA.5 (22B clade). Whereas E484K was identified in 2.3% of global sequences belonging to all lineages except Omicron lineages. E484Q was observed among 0.2% of sequences from Beta and Delta lineages detected from May 2021 to December 2022 (Figure 4.8C).

Fifty homogeneous SNPs associated with the primary VOCs detected in South Africa were identified (Table 4.5). SNPs identified in Beta included D80A, D215G, K417N, E484K, N501Y, D614G, and A701V with an AF ranging from 0.6-1.0. Individuals infected with Delta had T19R, G142D, delFR157-158, L452R, and T478K (AF ranging from 0.5-1.0) mutations. Forty-one mutations (Table 4.5) with an AF of 0.6-1.0, were detected among Omicron-infected individuals. Mutations reported at positions 346 and 444 in Omicron sub-lineages were not detected.

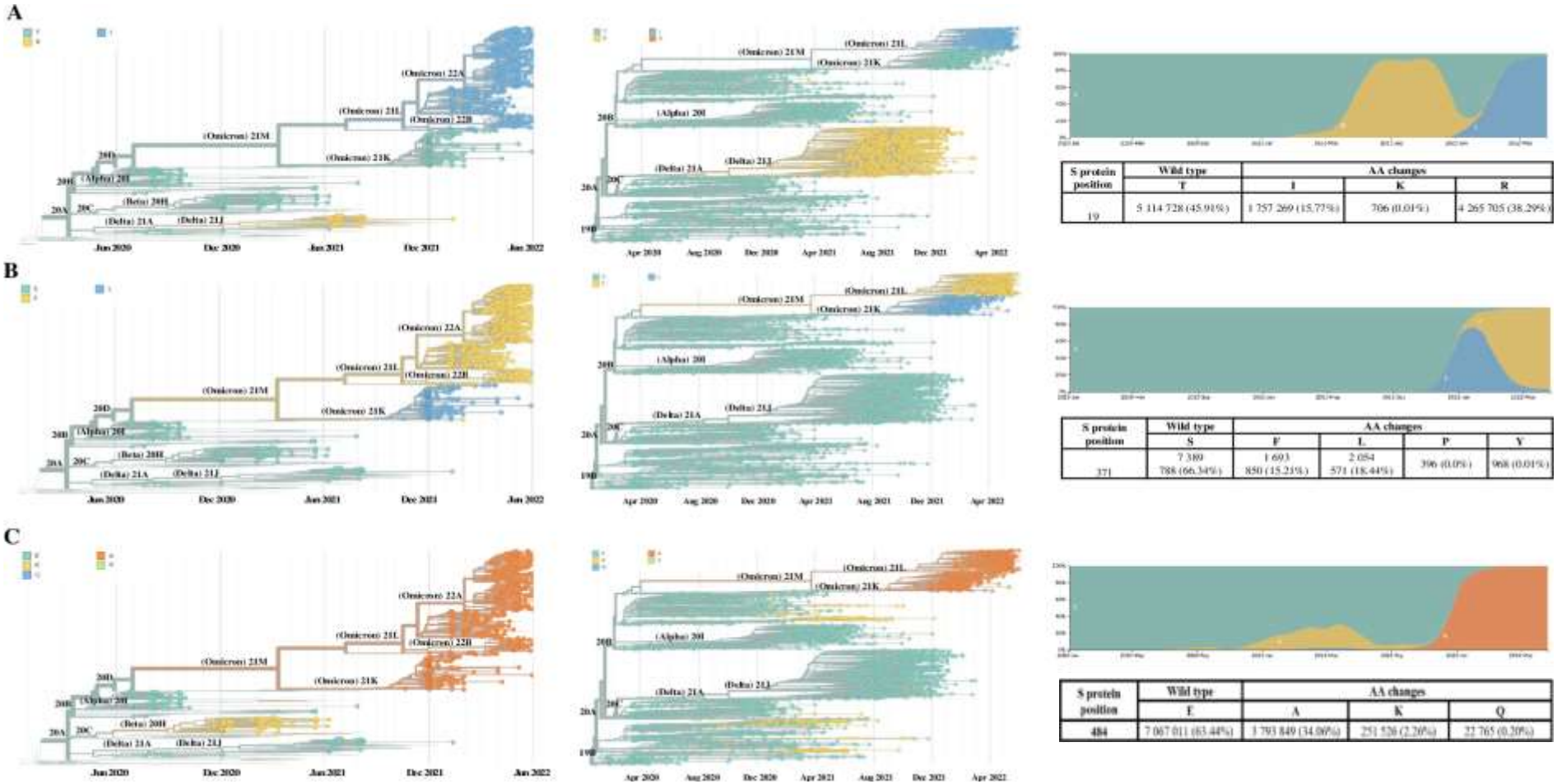

Fig. 4.8: Frequency of mutations occurring at aa positions 19, 371, and 484 in the S protein. Custom NextStrain time tree (left) displays the sequence data from this study from June 2020 to May 2022, with specific aa changes at the heterogeneous positions. The phylogenetic tree (middle) represents the global SARS-CoV-2 sequence data from January 2020 to May 2022, with specific aa changes at the heterogeneous positions (https://nextstrain.org/ncov/gisaid/global/all-time, accessed 22 June 2022). Clades displayed in phylogenetic trees represent the following SARS-CoV-2 pangolin lineages or VOCs (https://covariants.org/variants, accessed 22 June 2022): 20A and 20C include B.1 lineages, 20B includes B.1.1 and its sub-lineages; 20D includes B.1.1.1 lineage, Delta 21A, and Delta 21J includes B.1.617.2 and AY lineages; Alpha 20I includes B.1.1.7; Omicron 21M includes B.1.1.529 lineage; Omicron 21K includes BA.1 lineage; Omicron 21L includes BA.2 lineages and branches out to Omicron 22A (BA.4) and Omicron 22B (BA.5). Graphs on the right show the frequency of the aa changes over time (https://nextstrain.org/ncov/gisaid/global, accessed 22 June 2022). Tables show the frequency of wild-type aa and mutations present globally from GISAID (https://www.gisaid.org/epiflu-applications/covsurver-mutations-app/, accessed 22 June 2022).

Table 4.5: Allele frequencies and prevalence of homogeneous mutations identified among all individuals in the study population

Homogeneous mutations	AF range	n/N (%)	VOC
T19I	0.8-0.9	486/511 (95.1%)	Omicron BA.2/BA.4
T19R	0.8-1.0	610/657 (92.9%)	Delta
L24S	0.1-0.2	2/2 (100%)	Omicron BA.2/BA.4
A67V	0.9-1.0	187/228 (82.0%)	Omicron BA.1
del69/70	0.1	1/1 (100%)	Omicron BA.1/BA.4
D80A	0.8-0.9	112/141 (88.4%)	Beta
T95I	0.8-0.9	206/231 (89.2%)	Omicron BA.1
G142D	0.6-0.9	553/580 (95.3%)	Delta, Omicron BA.1/BA.2/BA.4
del143-145	1.0	1/1 (100%)	Omicron BA.1
del157-158	1.0	2/2 (100%)	Delta
DelL212	1.0	3/4 (75.0%)	Omicron BA.1
V213G	0.7-0.9	505/522 (96.7%)	Omicron BA.2/BA.4
D215G	0.8-0.9	89/134 (72.3%)	Beta
G339D	0.7-1.0	668/673 (99.3%)	Omicron BA.1/BA.2/BA.4
S371F	0.6-0.9	354/389 (91.0%)	Omicron BA.2/BA.4
S371L	0.8	14/22 (63.6%)	Omicron BA.1
S373P	0.7-1.0	637/646 (63.6%)	Omicron BA.1/BA.2/BA.4
S375F	0.7-1.0	627/640 (98.0%)	Omicron BA.1/BA.2/BA.4
T376A	0.8-1.0	385/405 (95.1%)	Omicron BA.2/BA.4
D405N	0.8-0.9	321/380 (84.5%)	Omicron BA.2/BA.4
R408S	0.6-0.9	277/306 (90.5%)	Omicron BA.2/BA.4
K417N	0.8-1.0	547/604 (90.6%)	Beta, Omicron BA.1/BA.2/BA.4
N440K	0.7-1.0	443/467 (94.9%)	Omicron BA.1/BA.2/BA.4
G446S	0.9	157/182 (86.3%)	Omicron BA.1
L452R	0.7-1.0	936/975 (96.0%)	Delta, Omicron BA.4
S477N	0.8-0.9	521/573 (90.9%)	Omicron BA.1/BA.2/BA.4
T478K	0.7-1.0	1303/1347 (96.7%)	Delta, Omicron BA.1/BA.2/BA.4
E484A	0.8-0.9	492/540 (91.1%)	Omicron BA.1/BA.2/BA.4
E484K	0.8-0.9	129/207 (75.8%)	Beta
F486V	0.8-0.9	256/283 (90.5%)	Omicron BA.4
Q493R	0.8-0.9	237/277 (85.6%)	Omicron BA.1/BA.2
G496S	0.8-0.9	181/206 (87.9%)	Omicron BA.1
Q498R	0.8-0.9	507/550 (92.2%)	Omicron BA.1/BA.2/BA.4
N501Y	0.8-1.0	683/738 (92.5%)	Beta, Omicron BA.1/BA.2/BA.4
Y505H	0.8-0.9	503/540 (93.1%)	Omicron BA.1/BA.2/BA.4
T547K	0.7-0.9	189/203 (93.1%)	Omicron BA.1
D614G	0.6-1.0	1692/1697 (99.7%)	Beta, Delta, Omicron BA.1/BA.2/BA.4
H655Y	0.8-1.0	751/780 (96.3%)	Omicron BA.1/BA.2/BA.4
N658S	0.9	67/90 (74.4%)	Omicron BA.4
N679K	0.8-1.0	761/793 (96.0%)	Omicron BA.1/BA.2/BA.4
P681H	0.8-1.0	753/761 (98.9%)	Omicron BA.1/BA.2/BA.4
P681R	0.8-1.0	645/675 (95.6%)	Delta
A701V	0.7-0.8	97/123 (88.0%)	Beta
N764K	0.7-1.0	646/656 (98.5%)	Omicron BA.1/BA.2/BA.4
D796Y	0.8-0.9	605/653 (92.6%)	Omicron BA.1/BA.2/BA.4
N856K	0.8-0.9	181/202 (89.6%)	Omicron BA.1
D950N	0.5-0.9	625/682 (91.6%)	Delta
Q954H	0.7-1.0	650/660 (98.5%)	Omicron BA.1/BA.2/BA.4
N969K	0.7-1.0	687/696 (98.7%)	Omicron BA.1/BA.2/BA.4
L981F	0.8-0.9	188/214 (87.9%)	Omicron BA.1

4.4. Discussion

Intra-host SARS-CoV-2 heterogeneity results in greater viral diversity and viral evolutionary advantage to enhance virus-host adaptation, immune evasion, replication, transmission, and pathogenicity (111,114,115). Diversity studies identified high variability of iSNVs in several regions of the genome such as the ORF and N genes of SARS-CoV-2, with a high frequency of allelic variations, especially within ORF8 and negative selective pressure towards the S gene. The mutation profiles of intra-host quasispecies were also shown to change from the onset of symptoms which could contribute to viral evolution occurring during the course of infection (111,114,115). Our study assessed the intra-host diversity based on the heterogeneity of SNPs across the S gene, not previously described in South Africa, to postulate the impact on host immune responses.

In contrast to earlier studies, our study period from June 2020 to May 2022 encompassed five SARS-COV-2/COVID-19 infection waves in South Africa. The 4th wave, driven by the Omicron BA.1 and BA.2 lineages, had the highest detection rate peaking at 58%, while the lowest peak detection rate (24%) was observed during the Omicron BA.4/BA.5 driven wave of infections (19,20). Although the majority of samples included in this study were from the Gauteng province, the overall distribution of SARS-CoV-2 lineages and VOCs was similar in other Provinces of South Africa (19,21). Individuals that seek CST are usually asymptomatic or present with mild symptoms because of a causative agent other than SARS-CoV-2, therefore a drop in the detection rate was observed. It has been reported that individuals infected with Omicron have reduced severity of infection compared to Delta as well as immunity resulting from natural infection and vaccination within the population (22,23). Of the cases that tested positive for SARS-CoV-2 in this study, both males and females were predominantly from the 25-44 year age group and CST, with the majority belonging to females (59%).

SARS-CoV-2 S protein SNP analysis identified mixed alleles indicative of heterogeneous populations or quasispecies among 5.3% of cases. Heterogeneous genotypes were detected mainly at one or more of the S protein aa positions: 144, 484, 501 and 681, where P681H was the most frequently observed. Overall, sequencing-based variant calling analysis identified 8.8% heterogeneous infections among our study population, which included 3 samples analysed by SNP PCR. These results contrast with a study conducted in 2020 in Australia which reported quasispecies diversity in S protein at a frequency of 12% (13/107), although

their sample size was much smaller (117). We were unable to ascertain if study participants were infected with mixed virus populations, had multiple exposures/infections closely timed, or if the mixed populations observed were due to *de novo* evolution in the infected host during illness.

The lower frequency of heterogeneity reported here for SNP-based approaches could be ascribed to the lower sensitivity of SNP assays. In some cases, more than 50% of samples could not be analysed for SNPs despite acceptable cycle threshold values obtained with the diagnostic tests. Of the heterogeneous SNPs identified, only aa position 144 was truly heterogeneous from sequencing. P681H was the dominant heterogeneous SNP observed during the tail-end of the 3rd COVID-19 wave, but sequencing confirmed that this position had the wild type or mutant at position 681, but not both. Multiple mutations that occur at the same position in the SARS-CoV-2 genome, including P681H and P681R, make genotyping analysis challenging since the probes in the assay for one mutation failed to bind to viral sequence with the adjacent mutations (103). Mutations at this position appear as mixed (wild type/mutant) and moved away from wild type/wild type or mutant/mutant assignment as a result of weak amplification due to non-specific probe activity was evident. In 2021, a study in Birmingham tested K417N, L452R, E484K, N501Y, and P681R; and only identified heterogeneity for E484K (1/42; 0.6%) and P681R (8/42; 19.0%) from the SNP assays, not observed in our findings (103). Another study that tested 13 S protein TaqMan SNP assays, excluding P681H but including additional T20N, del242, and T478K; showed a 100% concordance when comparing the SNP assay to sequencing data (118). However, only 9/150 samples were confirmed using sequencing, and comparisons were based on lineage/VOC calling rather than the frequency of wild type and/or mutation.

Of the samples sequenced, heterogeneity was primarily identified at aa positions 19 (T19IR and T19KR), 371 (S371FP), and 484 (E484AK, E484AQ, E484KQ), similar to the SNP PCR which also identified heterogeneity of the S protein at these aa positions, before the 3rd wave through to the 5th wave of infections in South Africa. We identified T19IR among individuals that tested during CST and had Omicron BA.2.15 infection, while the lineage information for those with T19KR mutations was unknown due to missing sequence data. Amino acid position 19 is present in the N-terminal domain (NTD) of the S gene, and mutations at this position previously identified in Delta and Omicron BA.2 (119,120). The site I ("supersite") on the

NTD is usually the target for NTD-specific neutralising antibodies. However, the T19R mutation disrupts S protein conformation, which reduces virus neutralisation by antibodies (121). When comparing the mutations observed in our study to the global context, high frequencies of T19I were observed among Omicron BA.2 lineages and T19R among Delta. T19K was observed at 0.01% in Alpha VOC that circulated at low frequencies in South Africa (108).

Our data showed that S371FP was identified among Omicron BA.1, BA.2.15, and BA.4 lineages. S371F was observed at much higher frequencies compared to S371P, both of which were identified in the Omicron BA.1 and BA.2. S371L was not associated with heterogeneous infections in our study but was identified as the dominant mutation globally among Omicron BA.1 lineage strains. Mutations at position S371 led to a reduced recognition of the receptor-binding domain (RBD) antibodies (119,120). Similar to our study, S371P, and S371F were reported among Omicron BA.2 lineages in Europe and Italy, and in BA.3 lineage which was sporadic in South Africa (102,120). Despite the fact that the region around S371 to F541 does not have any N-linked glycosylation sites, aa changes at position 371 still decrease neutralising antibody recognition (122). S371P is recognised as a mutation that results in virus escape from neutralisation by antibodies (123), while S371F alters the conformation of the RBD which results in neutralising antibody escape (32). The most frequently observed heterogeneity was among individuals in age groups 5-14 (22.0%), 15-24 (23.3%), and 25-44 (32.9%). Although limited information on HIV-1 status among study participants including out-patients that visited ARV clinics and had heterogeneous mutations, recent data showed that the HIV-1 prevalence among South Africans aged 15 to 49 years remains high at 19.5% (125). A shedding study done in South Africa showed that persons living with HIV/AIDS with low CD4 counts (<200cells/µl) had significantly longer periods of SARS-CoV-2 shedding, ranging between 7 and 43 days, during the 2020 pandemic period in South Africa (126). A longer period of virus shedding provides an opportunity for the virus to evolve in its efforts to evade the SARS-CoV-2 specific immune responses (127–129). Omicron BA.1 was initially detected in HIV-infected patients in Botswana suggesting that individuals with underlying medical conditions such as HIV infection are at a higher risk of SARS-CoV-2 infection and severe disease as well as a possible predisposition for evolving SARS-CoV-2 mutants (130).

We identified E484AK, E484AQ, and E484KQ heterogeneity at very low frequencies among Delta and Beta variants from CST persons and an ARV clinic patient. Unlike our study, E484 mutations were not previously detected in Delta but we could speculate that the virus underwent adaptive changes since E484AK/AQ/KQ were identified among cases just after the Delta (3rd) wave, leading towards the Omicron wave. Globally, E484A was observed at the highest frequency and was primarily observed in Omicron BA.1 and BA.2 lineages. E484K was less frequent but appeared in all lineages over time, except for Omicron. Previously studies reported E484A in Omicron BA.1, BA.2, BA.3, BA.4, and BA.5 lineages; E484K was identified in C.1.2, Beta, Alpha, Gamma, Eta, Lota, and Mu; and E484Q in Kappa and B.1.595 lineages (102,131,132). The E484A, E484K, and E484Q mutations reduce sensitivity to vaccine-induced antibodies through immune escape from class 2 neutralising antibodies, with increased transmissibility due to higher affinity for the host receptor (98,131,133,134). Of note is that E484 was the only site for which vaccine data was available compared to T19 and S371. The E484Q mutation is associated with reduced resistance to neutralisation by post-vaccination sera, and two doses of mRNA vaccines provide protection against E484Q variants in adults (131,132). E484K has shown evidence of neutralisation by the monoclonal antibodies bamlanivumab and etesevimab as well as convalescent sera, but reduced neutralising antibody activity was observed among Pfizer-BioNTech and Moderna vaccine recipients (132).

Other studies reported significant non-synonymous SNVs in the S gene reported early in the pandemic during 2020, which include mutations at nucleotide position G22899T, which translates to amino acid mutation G446V in the RBD, and G24557T which translates to a G999C mutation in the heptad repeat region associated with membrane fusion during virus entry (117). Similarly, another study identified nucleotide substitutions at positions 22899 (translates to G446V), 21575 (L5F associated with increased infectivity in-vitro), and 24198 (A879V with reported reduced sensitivity to neutralisation by convalescent sera in-vitro) which are present in the RBD of the S protein (118). However, these mutations were not observed in our study. Several mutations in the spike protein are responsible for neutralising antibody escape. Omicron lineages are associated with escaping neutralisation from various antibody groups (99). Group D antibodies were affected by mutations at positions 440, 444, 446, and 448 found in the Omicron lineages (99). While group E neutralising antibodies are sensitive to mutations at positions 339, 345 and 346, the frequency of R346K mutations rapidly increased with emerging Omicron lineages. The primary mutations known to evade neutralisation by

antibodies in Beta and Delta include the mutations at positions 484 and 478 (135,136). Of the described mutations, we identified changes at aa homogeneous positions 339, 478, and 484 where 484 was also heterogeneous.

Heterogeneity among our study samples was predominantly observed from June 2021 through to May 2022 covering the Delta to Omicron dominant waves. It was also evident among individuals that sought CST and were expected to be asymptomatic or present with mild influenza-like illness symptoms. However, in this study the majority of cases were from CST for which clinical data were not collected; therefore, we could not conclude that heterogeneous SARS-CoV-2 infections were related to the severity of the disease. Another contributing factor to the observed heterogeneity may be immune selective pressures and adaptation when virus transmission occurs between persons with different HLA backgrounds (117).

The first limitation of the study was the performance of the T20N, P681H, E484K, and E484Q genotyping assays. The cycle-threshold (Ct) values, of samples from diagnostic SARS-CoV-2 testing, were compared to the genotyping assay results to determine the success of each assay. We genotyped most samples that tested positive for SARS-CoV-2 with Ct-values ranging from 13 to 37 from diagnostic PCR assays. However, it is important to note that only two of these assays relate to heterogeneous SNP positions identified by variant calling analysis. We, therefore, concluded that the reduced sensitivity of some Thermo Fisher SNP assays to distinguish wild-type from mutant might be due to the negative impact of other mutations surrounding the SNP of interest as described in a previous study (103). Secondly, incomplete or low-quality S gene sequence coverage was generated from whole genome based NGS due to primer mismatches as a result of novel variants emerging and becoming dominant in each wave of infection. We acknowledge that this problem could have been overcome by updating amplification primers or spiking the primers, but we did not re-sequence samples due to cost. Thirdly, our study is not a longitudinal study but rather describes the heterogeneity across the entire population, looking at the diversity and dominance of heterogeneous mutations within lineages and VOCs over time. Fourth, samples included in this study were primarily from the Gauteng Province, although the lineages and VOCs circulated similarly to that observed in the other Provinces of South Africa, we could not assume that similar heterogeneity would occur throughout the country. A fifth limitation was that we did not have the clinical data at the time of the study for individuals that were tested as part of community testing and from out-patients

and encompassed most of the cohort. Lastly, we did not perform antibody neutralisation assays to determine the significance of immune adaptations during the different waves of infection among individuals infected with the heterogeneous mutations detected.

4.5. Conclusion

Unlike other intra-host diversity studies that focused on mutations of the SARS-CoV-2 genome, our study analysed the S protein aa changes as it is the primary target for neutralising antibodies and this focus for viral immune evasion. Substitutions at aa positions 19, 371, and 484 have previously been shown to reduce virus neutralisation by antibodies. However, we are uncertain if the impact is less or greater among persons infected with SARS-CoV-2 that display heterogeneity in the S protein. Low-frequency variants at heterogeneous aa positions across the S protein provide opportunities for the emergence of SARS-CoV-2 variants that can overcome or outcompete when conditions favour survival and transmission. Our findings show that heterogeneous mutations may be a result of intra-host adaptation and may drive the evolution of the S protein which could contribute to immune evasion. This may contribute to the ongoing emergence of new variants associated with continued outbreaks globally.

We therefore aimed to investigate the persistence and frequency of non-VOC SARS-CoV-2 lineages across the COVID-19 waves in South Africa.

CHAPTER 5

Molecular Epidemiology of SARS-CoV-2 during Five COVID-19 Waves and the Significance of Low-Frequency Lineages

Reference (Appendix C): Subramoney, K., Mtileni, N., Giandhari, J., Naidoo, Y., Ramphal, Y., Pillay, S., Ramphal, U., Maharaj, A., Tshiabuila, D., Tegally, H., Wilkinson, E., de Oliveira, T., Fielding, B. C., & Treurnicht, F. K. (2023). Molecular Epidemiology of SARS-CoV-2 during Five COVID-19 Waves and the Significance of Low-Frequency Lineages. *Viruses*, *15*(5), 1194.
https://doi.org/10.3390/v15051194

5.1. Introduction

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is continually evolving, resulting in the emergence of several lineages and variants of concern (VOC). The first VOC detected was Alpha (B.1.1.7) which originated in the UK (40), followed by Beta (B.1.351), which originated in South Africa (SA) in August 2020 (43,86). Delta (B.1.617.2) was initially detected in India in April 2021 but spread rapidly to other countries, including SA, and led to the emergence of several highly transmissible sub-lineages (137). The most recent VOC identified was Omicron (B.1.1.529), which is divided into BA.1, BA.2, BA.4, and BA.5 lineages and several sub-lineages (138). Omicron BA.3 was not detected in South Africa during our study period. Omicron BA.1 initially circulated in November 2021 and was replaced by BA.2 in early 2022, from which BA.4 and BA.5 emerged by April 2022. Currently, Omicron BA.5 and its sub-lineages are driving the SARS-CoV-2 infection globally (40)). In SA, the National Department of Health implemented stringent preventative measures, including wearing masks, social distancing, and different levels of national lock-down with each COVID-19 wave (139). Longer-term prevention strategies included the rollout of vaccines: five vaccines have been approved for use in SA, including Covishield (Oxford/AstraZeneca), Comirnaty (Pfizer/BioNTech), Janssen (Johnson & Johnsons), COVID-19 vaccine MC Pharma (MC Pharma), and Coronavac (Sinovac) (140).

Since May 2023, Africa has confirmed 9 525 097 SARS-CoV-2 cases and 161 630 genomes sequenced, while South Africa had ~4 million confirmed cases, of which 52 720 genomes were sequenced (141,142). Prior to the circulation of Beta in SA, from March to August 2020, 16 different SARS-CoV-2 lineages were detected in SA, of which B.1.1.54, B.1.1.56, and C.1 drove the first wave of infection (47). The first significant mutation D614G, in the spike (S) protein, first emerged in the original B.1 lineage and subsequently became prevalent in other lineages and VOCs of SARS-CoV-2. B.1.1.54 and B.1.1.56 are sub-lineages of the B.1 lineage that have accumulated additional mutations, resulting in approximately 14 mutations compared to the original Wuhan-Hu1 strain. The C.1 lineage is a separate variant of the virus that was identified with 16 mutations compared to the Wuhan-Hu-1 strain. By mid-September 2020, SA saw approximately 42 different SARS-CoV-2 lineages, with at least 24 lineages circulating within an epiweek during peak infections (47). C.1.1 became the dominant lineage with S protein mutations S477N, A688S, and M1237I and with additional mutations similar to B.1.525 (Q52R and A67V), a globally circulating lineage (86). Convergence was observed between the C.1.2 lineage and other VOCs, but this lineage did not dominate in SA or globally due to its sensitivity to neutralising convalescent plasma antibodies from Beta- and Delta-infected person (143).

The rapid emergence and circulation of novel lineages were not unexpected, as previous reports described the rate of mutations being $0.8–2.4 \times 10^{-3}$ nucleotide substitutions/site/year (144). Such rapid evolution allows the virus to adapt to environmental pressures (i.e., a new host) and to evade vaccine-induced or infection-induced immunity. As VOCs evolved from the Wuhan-Hu-1 strain, SARS-CoV-2 experienced multiple convergent episodes, which enhanced infectivity and disease transmission and improved resistance mechanisms for immune escape (40). Of note, the novel lineages are more likely to evolve from previous VOCs, as observed with Delta and Omicron (40). Limited reports of the circulation, or re-emergence, of the earlier and less frequent lineages during the subsequent COVID-19 waves are available. To better understand the molecular epidemiology of SARS-CoV-2 lineages in SA, we described the circulation of SARS-CoV-2 lineages and VOCs detected between 2020 to 2022 in Gauteng Province, the potential significance of low-frequency lineages and their contribution to the emergence of new lineages.

5.2. Materials and Methods

5.2.1. Study population and samples

SARS-CoV-2 respiratory specimens from individuals of all ages from 30 March 2020 to 13 October 2022 were included in this study. The samples were tested for SARS-CoV-2 at the NHLS at CMJAH, Gauteng, South Africa. The samples received were from the Gauteng province, as well as referrals from the Eastern Cape, KwaZulu-Natal, Limpopo and Western Cape provinces of SA. This study contributed to the national surveillance for SARS-CoV-2 in SA, for which formal patient consent was not required. The study was approved by the University of the Witwatersrand Human Research Ethics Committee (M210119), and all participants' data were anonymised and presented in aggregated form.

5.2.2. Next-generation sequencing of SARS-CoV-2 strains

Amplicon-based sequencing was performed as described in 2.4.4.

5.2.3. Data analysis

Descriptive statistical analysis was conducted across different age groups (<5, 5–14, 15–24, 25–44, 45–60, and >60 years), gender, provinces and patient status. Consensus sequences were analysed using Nextstrain (https://clades.nextstrain.org, accessed 20 February 2023) and pangolin (https://pangolin.cog-uk.io/, accessed 20 February 2023) online tools for variant assignment. Nextstrain custom build time-trees were run using our custom sequence data and metadata. S protein mutations were analysed using the Stanford University Coronavirus Antiviral & Resistance Database tool (https://covdb.stanford.edu/sierra/sars2/by-sequences/, accessed 23 February 2023).

5.3. Results

5.3.1. Study population demographics

A total of 2547/3797 (67.1%) SARS-CoV-2 whole genomes were successfully sequenced, with 10.9% (278/2547) from 30 March to 31 December 2020, 54.9% (1399/2547) from 1 January to 31 December 2021, and 34.2% (870/2547) from 1 January to 13 October 2022 (Table 5.1). From 2020 to 2022, SARS-CoV-2-positive individuals were detected across all age groups. The majority of SARS-CoV-2 infected individuals were among adults aged 25–44 years of age (47.3%; 1162/2547), with similar detection rates observed in 2020 (50.7%; 141/1278), 2021

(43.7%; 611/1399) and 2022 (47.1%; 410/870). The SARS-CoV-2 detection rate observed for females was 60.1% (1531/2547), while for males, it was 37.5% (956/2547) across the study period. Among our study, individuals residing in Gauteng Province represented 97.0% (2472/2547) of samples infrom the City of Johannesburg Metro district. Three percent (68/2547) included samples referred from the Eastern Cape, KwaZulu-Natal, Limpopo and Western Cape for testing at the NHLS Virology laboratory in Gauteng Province. SARS-CoV-2 positive individuals were predominant among those that sought community screening and test services (71.8%; 1828/2547), followed by out-patients (59.3%; 292/492) which were comprised of patients seen at casualty (19.3%; 492/2547), healthcare workers (16.3%; 80/492), and antiretroviral (ARV) clinic attendees (22.0%; 108/492).

Table 5.1: Demographic characteristics of successfully sequenced SARS-CoV-2 samples from 30 March 2020 to 13 October 2022 (N=2547)

Demographics	Years			Total (*n* = 2547)
	2020 (*n* = 278)	2021 (*n* = 1399)	2022 (*n* = 870)	
Age Groups (*n* (%))				
<5	4 (1.4%)	12 (0.9%)	15 (1.7%)	31 (1.2%)
5–14	6 (2.2%)	129 (9.2%)	54 (6.2%)	189 (7.7.%)
15–24	26 (9.4%)	234 (16.7%)	119 (13.7%)	379 (15.4%)
25–44	141 (50.7%)	611 (43.7%)	410 (47.1%)	1162 (47.3%)
45–60	59 (21.2%)	296 (21.2%)	162 (18.6%)	517 (21.0%)
>60	33 (3.2%)	97 (6.9%)	68 (7.8%)	198 (8.1%)
Unknown *	9 (3.2%)	20 (1.4%)	42 (4.8%)	71 (2.9%)
Sex (*n* (%))				
Female	169 (60.8%)	813 (58.1%)	548 (63.0%)	1531 (60.1%)
Male	106 (38.1%)	550 (39.3%)	299 (34.4%)	956 (37.5%)
Unknown *	3 (1.1%)	36 (2.6%)	23 (2.6%)	62 (2.4%)
Province (*n* (%))				
Eastern Cape	5 (1.8%)	54 (3.7%)	0 (0.0%)	59 (2.3.%)
Gauteng	264 (95.0%)	1337 (95.6%)	870 (100.0%)	2471 (97.0%)
KwaZulu-Natal	7 (2.5%)	1 (0.1%)	0 (0.0%)	8 (0.3%)
Limpopo	0 (0.0%)	1 (0.1%)	0 (0.0%)	1 (0.1%)
Western Cape	0 (0.0%)	6 (0.4%)	0 (0.0%)	6 (0.4%)
Unknown *	2 (0.7%)	0 (0.0%)	0 (0.0%)	2 (0.1%)
Patient status (*n* (%))				
Community screening and test	142 (51.5%)	1150 (82.2%)	536 (61.6%)	1828 (71.8%)
Deceased	0 (0.0%)	4 (0.3%)	11 (1.3%)	15 (0.6%)
In-patient	40 (14.4.%)	77 (5.5%)	92 (10.6%)	209 (8.2%)
Out-patient	93 (33.5%)	168 (12.0%)	231 (26.6%)	492 (19.3%)
Unknown *	3/278 (1.1%)	0 (0.0%)	0 (0.0%)	3/2459 (0.1%)

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

5.3.2. Epidemiology of SARS-CoV-2 lineages and VOCs, 2020 to 2022

During the first COVID-19 wave, 24 SARS-CoV-2 lineages were identified among the 278 genomes sequenced in 2020 from epidemiological weeks (epiweeks) 14 to 53. The B.1.1 (16%; 45/278), B.1.1.52 (5%; 13/278) and C.1 (13%; 37/278) lineages were the most prevalent in 2020. Lineages present at frequencies below 5% in the same period included B.l (3%; 8/278), B.1.1.348 (3%; 8/278), B.1.1.1 (2%; 5/278) and C.2 (2%; 6/278) (Figure 5.1A).

In 2020, B.1.1.52 was most prevalent from weeks 15 to 24, with the detection rate decreasing from 100% (1/1) to 33.3% (2/6) (Figure 5.1A). B.1.1 was observed from epiweeks 19 to 40, with detection rates fluctuating between 33.3% (2/6) to 53.8% (7/13) from weeks 27 to 40 (Figure 5.1A). C.1 was dominant from week 16 to 22 and 27 to 41 in 2020, with detection rates ranging from 16.7% to 100%, respectively, while C.2 remained lower at week 31 (11.1%; 1/9) to 32 (5.9%; 1/17) and 35 (50.0%; 1/2) to 39 (11.1%; 1/9) (Figure 5.1A). For lineages detected at lower frequencies, B.1 was scattered from week 14 to 17 (33.3% (1/3) to 50.0% (3/6)), week 30 (20.0%; 1/5), week 31 (11.1%; 1/9) and 36 (12.5%; 1/8); B.1.1.348 was only observed in 2020, from weeks 14 to 22 with detection rate of 66.7% (2/3) down to 20.0% (1/10) (Figure 5.1A). Beta (B.1.351) outcompeted all lineages and became dominant from epiweek 43 in 2020, driving the second wave of infection up to week 15 of 2021 (Figure 5.1A).

In 2021, Beta retained dominance up to week 15 at 84.2% (16/19) to 100% (4/4) and decreased to 2.9% (1/34) by week 28 (Figure 5.1B). Alpha (B.1.17) was also observed during the second wave, with peak detection rates at weeks 20 (50.0%; 2/4) and 22 (46.7%; 7/15) coinciding with Beta and Delta AY.46 (Figure 5.1B). Delta (B.1.617.2) and its sub-lineages outcompeted Beta and Alpha to dominate the third wave. Delta had an overall prevalence of 52.9% (737/1399) in 2021 (Figure 5.1B). Delta AY.45 sub-lineage (36.1%; 505/1399) was predominant from week 26 (58.0%; 51/88) to week 42 (73.3%; 11/15) and decreased to 2.7% (11/74) by week 47. Subsequently, Omicron BA.1 emerged and outcompeted Delta, dominating the beginning of the fourth wave (week 46 to 53) in 2021, with an overall prevalence of 30.2% (422/1399). From weeks 46 to 53, Omicron BA.1 peaked at 59.8% (55/92), followed by BA.1.18, BA.1.19 and BA.1.21 sub-lineages (Figure 5.1B). Omicron BA.2 started to emerge at week 49 in 2021 and continued to co-circulate with BA.1 until the end of 2021. Low-frequency lineages observed in 2021 included re-emerging B.1 and B.1.1. B.1, observed from weeks 1 to 7, with detection rates down to 10.5% (2/19) and decreasing to 0.4% from weeks 26 to 32 (Figure

5.1B). B.1.1 was observed at week 34 (50.0%; 1/2). C.1.2 co-circulated with Beta and Delta (Figure 5.1B), with an overall prevalence of 6.1% (53/870), but was observed from epiweeks 23 to 51 at very low numbers with detection rates ranging from 2.2% (1/46) to 28.6% (2/7).

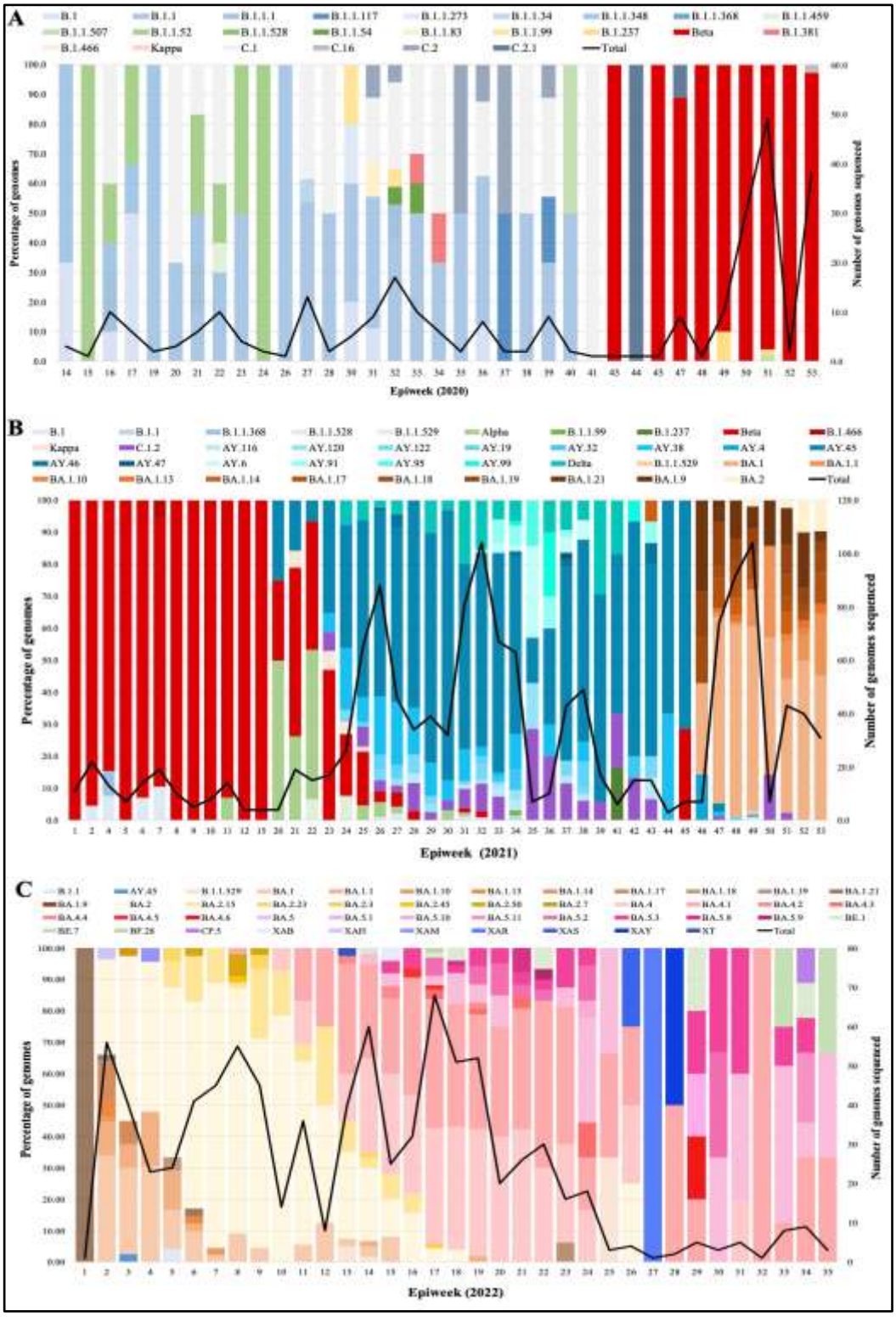

Fig. 5.1: Prevalence of SARS-CoV-2 lineages over time from 2020 to 2022 by epiweek. The bar graph represents the SAR-CoV-2 lineages and VOCs identified in our study cohort. The black line graph represents the

total number of samples that were sequenced during each epiweek. (**A**) Distribution of SARS-CoV-2 lineages in 2020, (**B**) Distribution of SARS-CoV-2 lineages in 2021, and (**C**) Distribution of SARS-CoV-2 lineages in 2022.

In 2022, Omicron BA.2 outcompeted BA.1 from week 3, continuing through the fourth wave (Figure 5.1C). BA.2 retained a prevalence >50% from weeks 3 to 11, peaking in week 7 (84.4%; 38/45). Omicron BA.4 and BA.5 dominated the fifth wave of infections from weeks 13 to 35 in 2022. Omicron BA.4 and BA.4.1 were dominant at similar frequencies from weeks 13 to 26 (15.0% to 53.3%). BA.5 slowly emerged from week 23 but only dominated from week 30 (33.3%; 1/3) to week 35 (33.3%; 1/3), when positivity rates were declining. Subsequently, Omicron sub-lineages continued to circulate, and recombinant strains were identified at an overall prevalence of 2.4% (21/870). BE.1 (0.9%; 8/870) and BE.7 (0.3%; 3/870) recombinants were the most prevalent at 1.5% to 25% but scantily distributed from weeks 14 to 42. Recombinants BF.28, CP.5, XAB, XAM, XAR, XAS, XAY and XT were only identified once from weeks 4 to 35 (Figure 5.1C). In addition, B.1.1 was observed again at week 5 in 2022 in a single case.

5.3.3. Phylogenetic description of VOCs overtime

Compared to global data, our sequences clustered with several VOCs, of which Beta, Delta and Omicron separated into the dominant waves of infection in South Africa as expected. All VOCs evolved from the clades 20C, 20A and 20B, respectively, after the first wave of infections (Figure 5.2A). Among our study samples, Alpha and Kappa were present, with 20B and 20A being the closest ancestors, respectively. An increase in the number of mutations gained across the SARS-CoV-2 whole genome was observed as new variants and their sub-lineages emerged from 2020 to 2022 (Figure 5.2A and 5.2B). From April to October 2020, 20A, 20B, and 20C SARS-CoV-2 clades were dominant and characterised by ~8–30 mutations across their genomes. The number of mutations present in Beta increased from ~20 to 37 from November 2020 to June 2021, respectively. Several Delta sub-lineages were observed from June to November 2021, with Delta 21A having the lowest number of mutations (21–29) and Delta 21J with up to 51 mutations. Omicron B.1.1.529 had ~52–62 mutations; its sub–lineages BA.1 and BA.2 had up to 72 mutations, while BA.4 and BA.5 detected from March to October 2022 had 72 to >80 mutations.

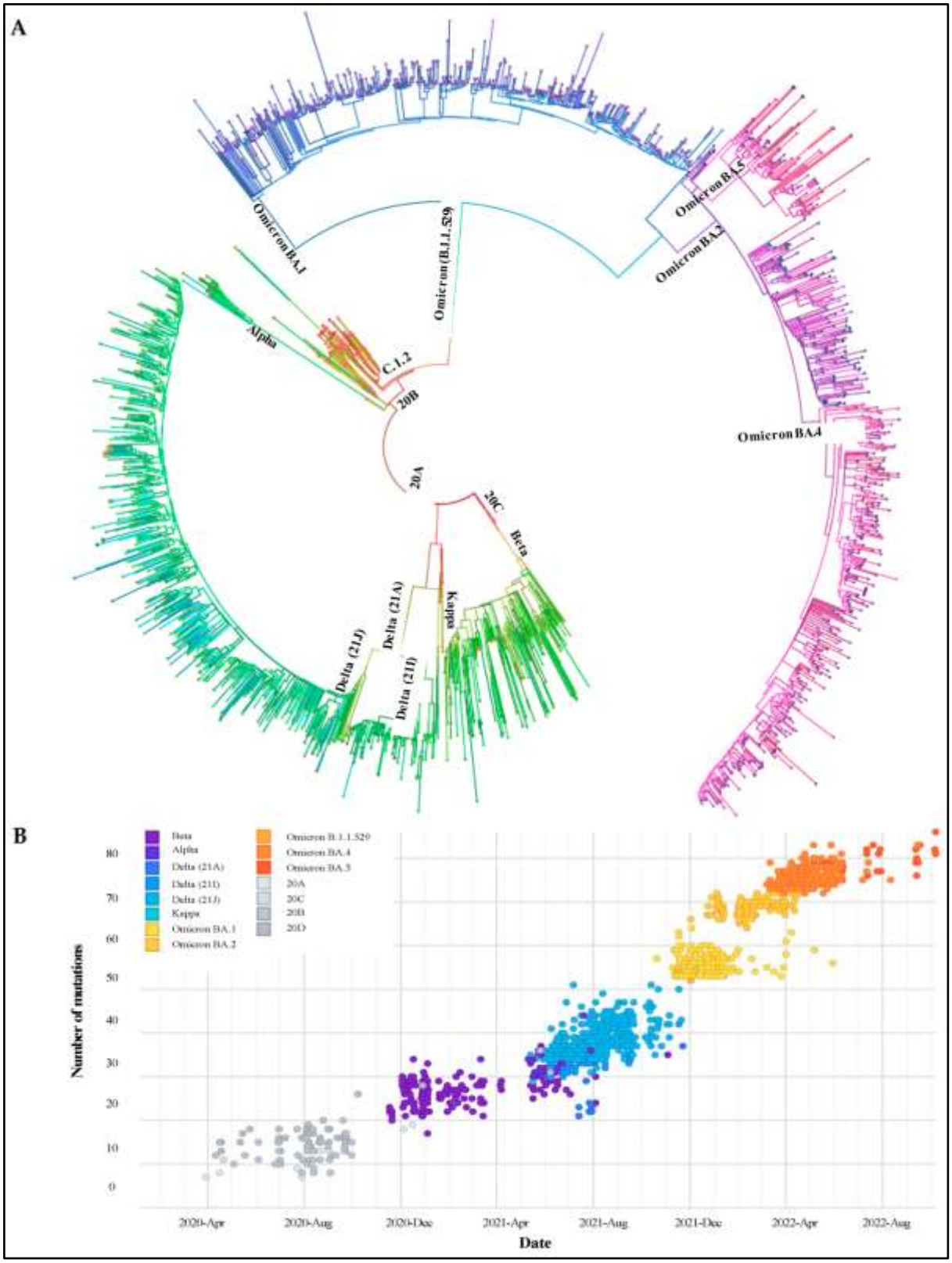

Fig. 5.2: Phylogenetic analysis of SARS-CoV-2 lineages and VOCs detected in South Africa from 2020 to 2022. (**A**) Radial phylogenetic tree displays the lineages identified in South Africa. Tree rooted with the Wuhan-Hu1 strain. Our dataset was compared to the global database embedded in NextStrain. (**B**) Nextclade custom time-tree displaying the increase in the number of mutations observed for each SARS-CoV-2 clade as time progressed.

5.3.4. Genome wide diversity of low-frequency lineages

Seven percent (182/2547) of samples comprised low-frequency lineages, including B.1 (0.7%; 17/2547), B.1.1 (1.8%; 47/2547), B.1.1.348 (0.3%; 8/2547), B.1.1.52 (0.5%; 13/2547), C.1 (1.5%; 38/2547), C.1.2 (2.1%; 53/2547), and C.2 (0.2%; 6/2547). B.1 clustered with the 20A and 20C clades, B.1.1/ B.1.152/ B.1.1.348 clustered with the 20B clade, and C.1/C.1.2/C.2 clustered with the 20D clades (Figure 5.3). As new lineages emerged with time, an increase in the number of mutations was gained across the SARS-CoV-2 whole genome (Figure 5.2A). The B.1 lineage was observed from 2020 through 2021. It was initially identified in March 2020 with 12 mutations which gradually increased to 21 mutations towards the end of April 2020.

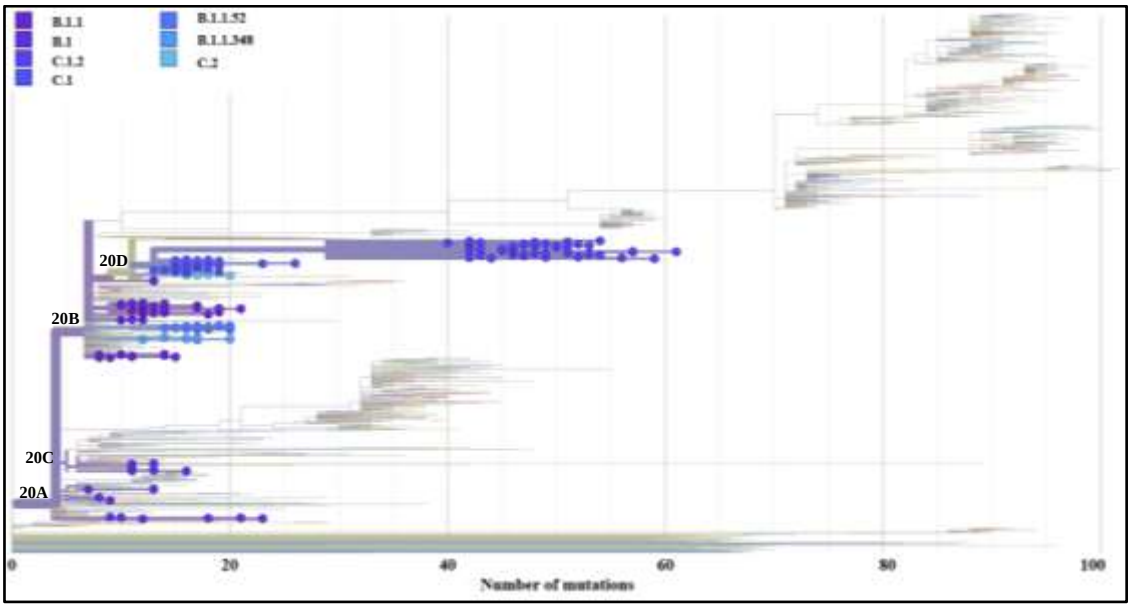

Fig. 5.3: SARS-CoV-2 low-frequency lineages identified in this study. The blue-shaded circles represent the samples designated as low-frequency lineages from our study cohort. The horizontal clusters in the background represent the additional lineages and VOCs identified globally.

B.1 was only observed again in July 2020, with 20 mutations, and decreased to 16 mutations by September 2020 (Figure 5.4b, Table 5.2). In January and February 2021, B.1 only had 9–15 mutations and evolved further from June (21 mutations) to August (23 mutations). From 2020 to 2021, during the first three waves of infection, B.1 retained five mutations, including P323L and G671I in the RNA Dependent RNA Polymerases (RdRp) protein; and C136F, delY144, and D614G in the S protein (Table 5.2). The B.1 lineage was predominantly observed among adults 45–60 years of age diagnosed through community screening and test services

(3/7; 57.1%) and outpatients (2/7; 28.6%) (Table 5.4). B.1.1 was observed consistently from May to September 2020, with mutations fluctuating daily, with at least 7 mutations present and a maximum of 20 mutations (Figure 5.4B, Table 5.3). Six mutations carried over from 2020 to 2021 (during the 1st, 3rd, and 4th waves), including R203K and G204R in the N protein; H208Y in the nsp2 protein; Y138H in the nsp8 protein; R126S in the ORF3a protein; P323L in the RdRp protein; and D614G in the S protein (Table 5.3). This lineage was predominant among younger adults aged 25–44 years from community screening (14/28; 50.05%) and out-patients in casualty (8/28;28.6%) (Table 5.4).

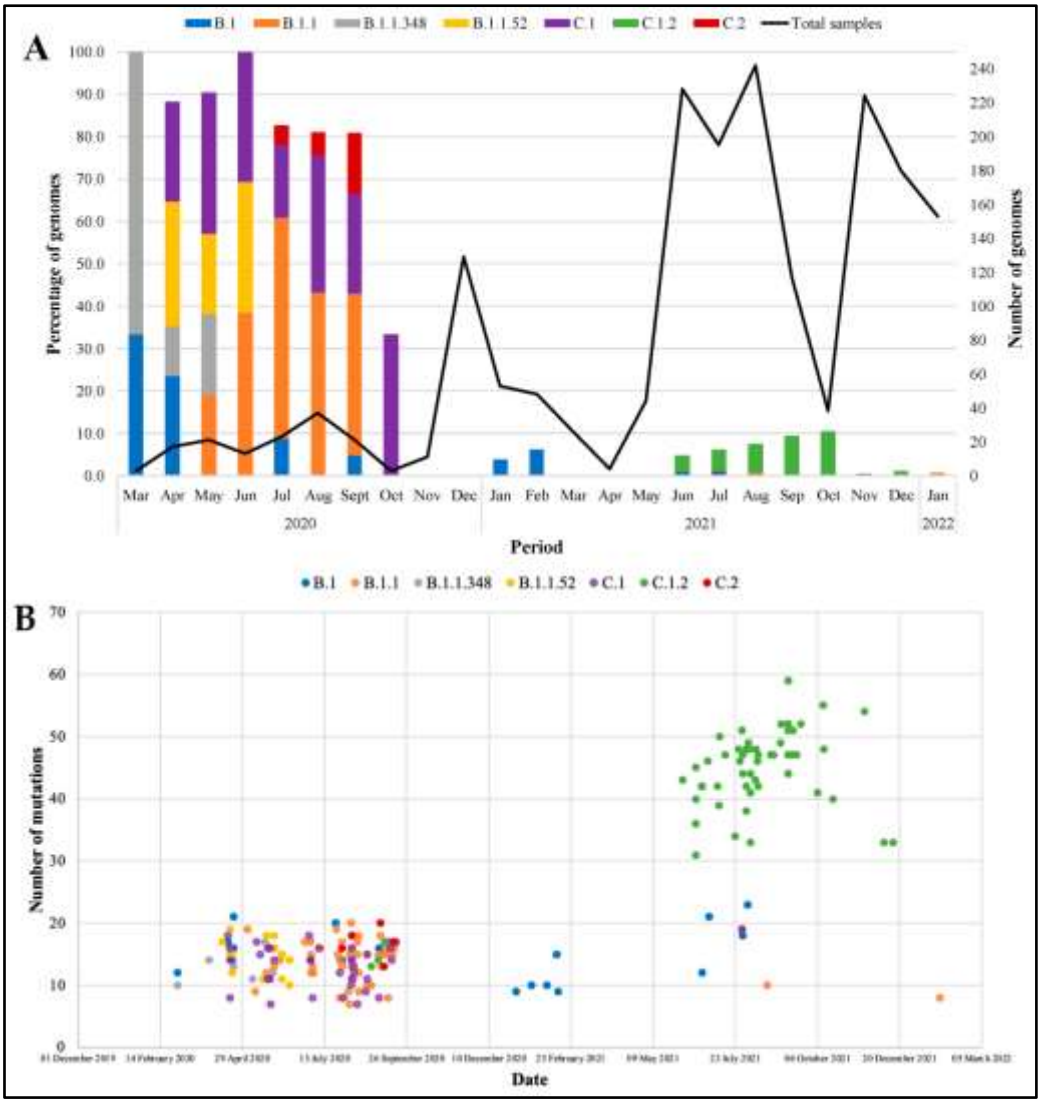

Fig. 5.4: SARS-CoV-2 low-frequency lineages identified over time. (**A**) The bars represent the SAR-CoV-2 lineages, and the line graph represents the total number of samples that were successfully sequenced during each month from 2020 to 2022. (**B**) The shaded dots represent the samples designated as low-frequency lineages. The y-axis represents the total number of mutations identified in each lineage across the whole genome. The x-axis represents the date during which the lineages were observed.

Table 5.2: Mutations detected across the whole genome of B.1 lineage from 2020 to 2021

Genes	3CLpro	E	M					N					nsp2			nsp4				nsp10	nsp13
Positions	**90**	**71**	**3**	**23**	**82**	**210**	**212**	**63**	**194**	**203**	**205**	**377**	**267**	**339**	**596**	**37**	**167**	**399**	**492**	**48**	**77**
Wuhan Hu-1	**K**	**P**	**D**	**V**	**I**	**H**	**S**	**D**	**S**	**R**	**T**	**D**	**N**	**G**	**A**	**T**	**V**	**K**	**T**	**H**	**P**
March 2020 (N=11)																					
April 2020 (N=4)						R	R									S				Y	
July 2020 (N=2)			G						L				del								
September 2020 (N=1)				L					L												
January 2021 (N=2)	R	L																			
February 2021 (N=2)	R			L							I			S							
June 2021 (N=2)					T							Y					L				L
July 2021 (N=1)					T			G		M		Y									L
August 2021 (N=1)					T					M		Y			T				I		L

Genes	nsp14	ORF3a		ORF7a		ORF7b	ORF8		PLpro							RdRp					
Positions	**316**	**26**	**57**	**82**	**120**	**40**	**119**	**120**	**250**	**428**	**488**	**1381**	**1382**	**1639**	**1887**	**323**	**554**	**555**	**556**	**557**	**671**
Wuhan Hu-1	**V**	**S**	**Q**	**V**	**T**	**T**	**D**	**F**	**G**	**T**	**A**	**A**	**K**	**T**	**K**	**P**	**A**	**R**	**T**	**V**	**G**
March 2020 (N=11)														A	N	L					
April 2020 (N=4)			H													L					
July 2020 (N=2)																L					
September 2020 (N=1)	F															L					S
January 2021 (N=2)																L	V				
February 2021 (N=2)			H											A		L					
June 2021 (N=2)					I		del	del		I						L					
July 2021 (N=1)				A		I	del	del	R							L					
August 2021 (N=1)		L					del	del			S	D	S			L					S

Genes	S																						
Positions	**80**	**98**	**136**	**144**	**156**	**157**	**158**	**190**	**215**	**243**	**244**	**417**	**449**	**452**	**478**	**484**	**501**	**614**	**679**	**701**	**716**	**859**	**950**
Wuhan Hu-1	**D**	**S**	**C**	**Y**	**E**	**F**	**R**	**R**	**D**	**A**	**L**	**K**	**Y**	**L**	**T**	**E**	**N**	**D**	**N**	**A**	**T**	**T**	**D**
March 2020 (N=11)																		G					
April 2020 (N=4)		F																G					
July 2020 (N=2)		F																G					
September 2020 (N=1)			F	del				S	G									G					
January 2021 (N=2)	A																	G		V			
February 2021 (N=2)	A																	G		V			
June 2021 (N=2)	A		F	del				S	G	del	del		H			K	Y	G	K		I	N	
July 2021 (N=1)			F	del											K			G					
August 2021 (N=1)	A				del	del	G							R				G					N

*3C-like protease (3CLpro), envelope (E), nucleocapsid (N), non-structural protein (nsp), open-reading frame (ORF), papain-like protease (PLpro), RNA-dependent RNA-polymerase RdRp, spike (S)

Table 5.3: Mutations detected across the whole genome of B.1.1 lineage from 2020 to 2022

Genes	3CLpro			E			N							nsp2						nsp4				nsp6				nsp8	nsp13	nsp14				nsp15		nsp16
Positions	**15**	**89**	**142**	**58**	**19**	**212**	**141**	**145**	**179**	**199**	**203**	**204**	**371**	**57**	**208**	**385**	**537**	**550**	**624**	**83**	**218**	**446**	**457**	**37**	**58**	**168**	**188**	**138**	**85**	**138**	**177**	**324**	**349**	**54**	**257**	**160**
Wuhan Hu-1	**G**	**L**	**N**	**V**	**Q**	**S**	**T**	**H**	**G**	**P**	**R**	**G**	**D**	**E**	**H**	**A**	**Y**	**L**	**P**	**T**	**S**	**A**	**A**	**L**	**M**	**I**	**G**	**Y**	**A**	**A**	**L**	**D**	**K**	**A**	**R**	**K**
May 2020 (N=1)	S		S			R					K	R					D				F							H								
June 2020 (N=1)		F	S								K	R								I								H					N	V		
July 2020 (N=1)									V		K	R	Y		Y				L					F				H		V		Y				
August 2020 (N=1)				I	E						K	R		G		V									I			H	V				N		C	R
September 2020 (N=1)							I	N			K	R			Y			F				S	V			T		H	V		F					
August 2021 (N=1)											K	R			Y												S	H								
January 2022 (N=1)										S	K	R																H								

Genes	ORF3a							ORF7b											ORF8		Plpro								RdRp		
Positions	**15**	**23**	**54**	**121**	**126**	**172**	**191**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**11**	**28**	**126**	**236**	**341**	**493**	**503**	**1119**	**1285**	**1639**	**173**	**257**	**323**
Wuhan Hu-1	**L**	**A**	**A**	**V**	**R**	**G**	**E**	**L**	**S**	**L**	**I**	**D**	**F**	**Y**	**L**	**C**	**F**	**L**	**T**	**A**	**S**	**P**	**I**	**E**	**P**	**F**	**S**	**T**	**R**	**V**	**P**
May 2020 (N=1)							A																			I		A		F	L
June 2020 (N=1)					S																						Y			F	L
July 2020 (N=1)					S		A																	A							L
August 2020 (N=1)		S			S		A												K			S	V		S						L
September 2020 (N=1)	F		S	L	S	C		del	del	del	del	del	del	del	del	del	del	del	K		L			A					C		L
August 2021 (N=1)																															L
January 2022 (N=1)																				V											L

Genes	S																				
Positions	**12**	**69**	**70**	**78**	**98**	**142**	**156**	**157**	**158**	**339**	**417**	**440**	**477**	**478**	**484**	**549**	**554**	**575**	**614**	**655**	**698**
Wuhan Hu-1	**S**	**H**	**V**	**R**	**S**	**G**	**E**	**F**	**R**	**G**	**K**	**N**	**S**	**T**	**E**	**T**	**E**	**A**	**D**	**H**	**S**
May 2020 (N=1)																			G		
June 2020 (N=1)																			G		
July 2020 (N=1)				M	F												D		G		
August 2020 (N=1)		del	del			D	del	del	G		N	K	N	J	A	I			G	Y	L
September 2020 (N=1)	F											K						S	G		
August 2021 (N=1)																			G		
January 2022 (N=1)										D									G		

*3C-like protease (3CLpro), envelope (E), nucleocapsid (N), non-structural protein (nsp), open-reading frame (ORF), papain-like protease (PLpro), RNA-dependent RNA-polymerase RdRp, spike (S)

Table 5.4: Demographic data for participants from whom B.1 and B.1.1 lineages were detected in 2020 to 2022

Patient status	B.1 (N=17)							B.1.1 (N=47)						
	<5 (N=0)	5-14 (N=1)	15-24 (N=0)	25-44 (N=5)	45-60 (N=7)	>60 (N=3)	Unknown (N=1)	<5 (N=0)	5-14 (N=0)	15-24 (N=6)	25-44 (N=28)	45-60 (N=7)	>60 (N=3)	Unknown (N=1)
Community screen and test		1/1 (100%)		5/5 (100%)	3/7 (57.1%)	2/3 (66.7%)				2/6 (33.3%)	14/28 (50.0%)	5/9 (55.6%)	2/3 (66.7%)	
In-patient					2/7 (28.6%)					2/6 (33.3%)	2/28 (7.1%)	1/9 (11.1%)		
Cardiac unit										1/6 (16.7%)	1/28 (3.6%)			
Neurology					1/7 (14.3%)									
Other					1/7 (14.3%)					1/6 (16.7%)	1/28 (3.6%)	1/9 (11.1%)		
Out-patient					1/7 (14.3%)					1/6 (16.7%)	12/28 (42.9%)	3/9 (33.3%)	1/3 (33.3%)	1/1 (100%)
ARV clinic											1/28 (3.6%)	1/9 (11.1%)		
Healthcare worker										1/6 (16.7%)	2/28 (7.1%)		1/3 (33.3%)	
Paediatric											1/28 (3.6%)			
Casualty					1/7 (14.3%)						8/28 (28.6%)	2/9 (22.2%)		
Unknown						1/3 (33.3%)	1/1 (100%)							
Total	0/17 (0.0%)	1/17 (5.9%)	0/17 (0.0%)	5/17 (29.4%)	7/17 (41.2%)	3/17 (6.4%)	0/17 (0.0%)	0/47 (0.0%)	0/47 (0.0%	6/47 (12.8%)	28/47 (59.6%)	9/47 (19.2%)	3/47 (6.4%)	1/47 (2.1%)

B.1.1.348 was observed from March to May 2020, with <20 mutations seen throughout this period, with 10 mutations gained in March, 16 in April and 17 in May (Figure 5.4B, Table 5.5). Six mutations were observed during the first wave, including S212R in the M protein, R203K and G204K in the N protein, del141-143 in the nsp1 protein, P323L in the RdRp protein, and D614G in the S protein. B.1.1.348 was predominantly detected in 25–44-year-old adults from community screening (Table 5.7). B.1.1.52 was present from April to June 2020, with a maximum of 19 mutations present in April 2020, which was stable until May 2020 (16–18 mutations) and declined by June 2020 (10–15 mutations) (Figure 5.4B, Table 5.6). Eight important mutations were identified in B.1.1.52 during the first wave as shown in Table 5.6. Mutations included H210R and S212R in the M protein, R203K and G204R in the N protein, Y138H in the nsp8 protein, T1639A in the PLpro protein, P323L in the RdRp protein, and D614G in the S protein. More than 50% of individuals were infected with B.1.1.52 (7/13; 53.8%) among community screening and out-patients from ARV clinics and casualty (Table 5.7). D614G was the only S protein mutation that was present with more than a single occurrence among the B.1.1.348 (87.5%; 7/8) and B.1.1.52 (92.9%; 13/14) lineages.

Table 5.5: Mutations detected across the whole genome of B.1.1.348 lineage in 2020

Genes	E	M			N		nsp1			nsp2	nsp4	PLpro		RdRp	S	
Positions	**19**	**125**	**210**	**212**	**203**	**204**	**141**	**142**	**143**	**424**	**492**	**428**	**1639**	**323**	**95**	**614**
Wuhan Hu-1	**L**	**H**	**H**	**S**	**R**	**G**	**K**	**S**	**F**	**G**	**T**	**T**	**T**	**P**	**T**	**D**
March 2020 (N=2)				R	K	R	del	del	del	C				L		G
April 2020 (N=2)		P	R	R	K	R						I	A	L		G
May 2020 (N=4)	P			R	K	R	del	del	del	C	A		A	L	I	G

*envelope (E), membrane (M), nucleocapsid (N), non-structural protein (nsp), papain-like protease (PLpro), RNA-dependent RNA-polymerase RdRp, spike (S)

Table 5.6: Mutations detected across the whole genome of B.1.1.52 lineage from 2020

Genes	3CLpro		M		N		nsp1	nsp2	nsp5	nsp6	nsp8	nsp14	orf3a		Plpro		RdRp		S
Positions	**142**	**196**	**210**	**212**	**203**	**204**	**84**	**424**	**218**	**220**	**138**	**349**	**23**	**255**	**666**	**1639**	**257**	**323**	**614**
Wuhan Hu-1	**N**	**T**	**H**	**S**	**R**	**G**	**V**	**G**	**S**	**F**	**Y**	**K**	**A**	**V**	**T**	**T**	**V**	**P**	**D**
April 2020 (N=5)		M	R	R	K	R		C	F		H		S	del	S	A		L	G
May 2020 (N=4)	S		R	R	K	R		C		V	H					A	F	L	G
June 2020 (N=4)			R	R	K	R	I				H	N		del		A		L	G

*3C-like protease (3CLpro), membrane (M), nucleocapsid (N), non-structural protein (nsp), open-reading frame (ORF), papain-like protease (PLpro), RNA-dependent RNA-polymerase RdRp, spike (S)

Table 5.7: Demographic data for participants from whom B.1.1.438 and B.1.1.52 lineages were identified in 2020 to 2022

Patient status	B.1.1.348 (N=8)							B.1.1.52 (N=13)						
	<5 (N=0)	5-14 (N=0)	15-24 (N=1)	25-44 (N=3)	45-60 (N=2)	>60 (N=1)	Unknown (N=1)	<5 (N=1)	5-14 (N=1)	15-24 (N=0)	25-44 (N=7)	45-60 (N=3)	>60 (N=0)	Unknown (N=1)
Community screen and test			1/1 (100%)	3/3 (100%)	1/2 (50.0%)		1/1 (100%)	1/1 (100%)	1/1 (100%)		2/7 (28.6%)	2/3 (66.7%)		
In-patient											1/7 (14.3%)			
ICU											1/7 (14.3%)			
Out-patient					1/2 (50.0%)						4/7 (57.1%)	1/3 (33.3%)		1/1 (100%)
ARV clinic						1/1 (100%)					2/7 (28.6%)			
Healthcare worker					1/2 (50.0%)	1/1 (100%)						1/3 (33.3%)		1/1 (100%)
Casualty											2/7 (28.6%)			
Total	0/8 (0.0%)	0/8 (0.0%)	1/8 (12.5%)	3/8 (37.5%)	2/8 (25.0%)	1/8 (12.5%)	1/8 (12.5%)	1/13 (7.7%)	1/13 (7.7%)	0/13 (0.0%)	7/13 (53.8%)	3/13 (23.1%)	0/13 (0.0%)	1/13 (7.7%)

C.1 emerged in April 2020 and circulated until September 2020. The lineage had <20 mutations overall, but >15 mutations were observed throughout the year and were present again in July 2021 with 19 mutations (Figure 5.4B, Table 5.10). During the 1st, 2nd, and 3rd waves from 2020 to 2021, 7 mutations were retained in the C.1 lineage: G15S in the 3CLpro; R203K, G204R, and Q384H in the N protein; del255 in the ORF3a, T428I in the PLpro, and P323L in the S protein. C.1 was predominantly identified among adults 25–44 years of age (22/38; 57.9%), primarily from community screening and out-patients from ARV clinics, healthcare workers, and casualty (Table 5.8). C.1.2 was observed from June to December 2021 with 31 to 59 mutations, increasing rapidly from August to September 2021 (Figure 5.4B, Table 5.11). The C.1.2 samples had ~23 mutations in the S protein, of which >80% of sequences had similar mutations. In 2021, during the 3rd to 4th waves, several mutations carried over including the

following: G15S and T24I in the 3CLpro, L21I and S68F/P in the E protein; L29F and I82T in the M protein; P13L, R203K, G204R and Q384H in the N protein; E102K in the nsp1 protein; L438P in the nsp4 protein; del106–108 in the nsp6 protein; S216L, T229N, del255 and L275F in the ORF3a; T428I and T819I in the PLpro protein, and P323L in the RdRp protein; P9L, P25L, C136F, delY144, R190S, D215G, del243-244, Y449H, E484K, N501Y, L585F, D614G, H655Y, N679K, T716I, and T859N in the S protein (Figure 5.4B, Table 5.11). In 2020, C.2 had 16 mutations present in July, which increased to 18 in August and 20 at the beginning of September and lost mutations in the middle of September (13–17 mutations) (Figure 5.4B). C.2 only had two primary mutations in the S protein (D614G and N679K), in 2020 that, were also detected in other lineages (Table 5.12). In 2020, during the first wave, eight mutations were detected in C.2. Mutations included G15S in the 3CLpro protein, R203K and G204R in the N protein, A85S in the nsp14 protein, T428I in the PLpro protein, L302S and P323L in the RdRp, and D614G in the S protein. Similar to other low-frequency lineages, C.1.2 and C.2 were predominant among adults aged 25–44 years from community surveillance (24/29; 82.8% and 1/3; 33.3%, respectively) and out-patients (4/29; 13.8% and 1/3; 33.3%, respectively) (Table 4.8 and 4.9).

Table 5.8: Demographic data for participants for whom C.1 and C.1.2 lineages were detected in2020 to 2022

Patient status	C.1 (N=38)							C.1.2 (N=53)						
	<5 (N=0)	5-14 (N=1)	15-24 (N=3)	25-44 (N=22)	45-60 (N=7)	>60 (N=3)	Unknown (N=1)	<5 (N=0)	5-14 (N=1)	15-24 (N=8)	25-44 (N=29)	45-60 (N=10)	>60 (N=5)	Unknown (N=0)
Community screen and test		1/1 (100%)	3/3 (100%)	11/22 (50.0%)	1/7 (14.2%)					8/8 (100%)	24/29 (82.8%)	10/10 (100%)	4/5 (80.0%)	
In-patient				2/22 (9.1%)	2/7 (28.6%)	1/3 (33.3%)			1/1 (100%)		1/29 (3.5%)			
Chronic dialysis					1/7 (14.3%)									
Other				2/22 (9.1%)	1/7 (14.3%)	1/3 (33.3%)			1/1 (100%)		1/29 (3.5%)			
Out-patient				9/22 (40.9%)	4/7 (57.1%)	2/3 (66.7%)					4/29 (13.8%)		1/5 (20.0%)	
ARV clinic				2/22 (9.1%)										
Healthcare worker				2/22 (9.1%)							1/29 (3.5%)			
Casualty				5/22 (22.7%)	4/7 (57.1%)	2/3 (66.7%)					3/29 (10.3%)		1/5 (20.0%)	
Unknown							1/1 (100%)							
Total	0/38 (0.0%	1/38 (2.6%)	3/38 (7.9%)	22/38 (57.9%)	7/38 (18.4%)	3/38 (7.9%)	1/38 (2.6%)	0/6 (0.0%)	1/6 (16.7%)	8/53 (15.1%)	29/53 (54.7%)	10/53 (18.9%)	5/53 (9.4%)	0/53 (0.0%)

Table 5.9: Demographic data for participants for whom C.2 lineages were detected in 2020 to 2022

Patient status	C.2 (N=6)						
	<5 (N=0)	15-24 (N=1)	15-24 (N=0)	25-44 (N=3)	45-60 (N=1)	>60 (N=1)	Unknown (N=0)
Community screen and test		1/1 (100%)		1/3 (33.3%)			
In-patient				1/3 (33.3%)	1/1 (100%)		
Oncology					1/1 (100%)		
Rheumatology/nephrology				1/3 (33.3%)			
Out-patient				1/3 (33.3%)		1/1 (100%)	
Casualty				1/3 (33.3%)		1/1 (100%)	
Total	0/6 (0.0%)	1/6 (16.7%)	0/6 (0.0%)	3/6 (50.0%)	1/6 (16.7%)	1/6 (16.7%)	0/6 (0.0%)

Table 5.10: Mutations detected across the whole genome of C.1 lineage from 2020 to 2021

Genes	3CLpro		M		N								nsp1		nsp2	nsp4			nsp6	nsp8			nsp13			nsp14		ORF3a				ORF8		PLpro			RdRp	S
Positions	15	38	8	68	34	70	141	203	204	272	382	384	49	156	200	218	438	457	37	29	138	178	18	392	403	305	516	202	240	254	255	24	62	428	945	1639	323	98
Wuhan Hu1	G	C	I	A	G	Q	T	R	G	Q	L	Q	G	D	P	S	L	A	L	G	Y	P	A	R	A	I	T	V	P	G	V	S	V	T	K	T	P	S
April 2020 (N=4)	S			del				K	R		S	H		N		F					H										del			I		A	L	
May 2020 (N=7)	S	G					I	K	R			H	C	N	S																del			I		A	L	
June 2020 (N=4)	S		V					K	R	R		H												C							del			I			L	
July 2020 (N=4)	S				W			K	R			H							F			S			V	V					del		M	I			L	
August 2020 (N=12)	S					R		K	R			H					F	T	F	S			T				I			R	del			I	R		L	
September 2020 (N=5)	S					R		K	R			H																del	S		del			I	R		L	F
October 2020 (N=1)								K	R																						del			I		A	L	
July 2021 (N=1)	S							K	R			H																			del			I			L	

*3C-like protease (3CLpro), membrane (M), nucleocapsid (N), non-structural protein (nsp), open-reading frame (ORF), papain-like protease (PLpro), RNA-dependent RNA-polymerase RdRp, spike (S)

Table 5.11: Mutations detected across the whole genome of C.1.2 lineage from 2020 to 2021

Genes	3CLpro				E		M		N							nsp1							
Positions	15	24	90	108	21	68	29	82	13	80	186	203	204	383	384	85	87	102	110	141	142	143	166
Wuhan Hu-1	G	T	K	P	L	S	L	I	P	P	S	R	G	P	Q	M	E	E	H	K	S	F	S
June 2021 (N=7)	S	I			I	F	F	T	L		P	K	R		H			K					
July 2021 (N=10)	S	I			I	P	F	T	L	R		K	R	L	H	del		K		del	del	del	
August 2021 (N=16)	S	I		S		P	F	T	L	R		K	R	L	H			K					
September 2021 (N=11)	S	I			I	P	F	T	L	R	P	K	R	L	H			K	L				G
October 2021 (N=4)	S		R		I	F	F	T	L	R		K	R	L	H			K					
November 2021 (N=1)	S					F		T	L			K	R		H			K					
December 2021 (N=2)	S	I			I		F	T	L			K	R		H			K					

Genes	nsp2			nsp4				nsp6				nsp10	nsp14	nsp15	nsp16	ORF3a												
Positions	157	181	429	33	233	289	438	37	106	107	108	111	68	33	160	18	33	34	67	95	155	181	216	223	229	238	255	275
Wuhan Hu-1	V	P	T	M	V	I	L	L	S	G	F	T	G	T	K	G	A	T	K	L	D	E	S	T	T	D	V	L
June 2021 (N=7)					I		P		del	del	del				R						Y		L		N		del	F
July 2021 (N=10)	I		I				P		del	del	del					V								I	N	Y	del	
August 2021 (N=16)		T	I	C			P		del	del	del								T			A		I	N	Y	del	
September 2021 (N=11)			I				P	F	del	del	del	I					S	M		F				I	N	Y	del	
October 2021 (N=4)			I				P		del	del	del			I									L		N		del	F
November 2021 (N=1)						V	P		del	del	del												L				del	F
December 2021 (N=2)									del	del	del		A										L		N		del	F

Genes	ORF7a	ORF7b	ORF8				Plpro											Rdrp				
Positions	43	40	7	16	120	121	93	126	153	237	385	428	819	822	864	1274	1693	140	184	323	337	744
Wuhan Hu-1	N	T	L	F	F	I	S	S	P	T	I	T	T	P	T	H	K	D	Q	P	G	E
June 2021 (N=7)										I		I	I			Y		A		L		D
July 2021 (N=10)		I							S			I	I	S						L		
August 2021 (N=16)	K		S		V	L	F	P			V	I	I	S	A		N			L	S	
September 2021 (N=11)			S	L								I	I	S	A					L		
October 2021 (N=4)			S									I	I							L		
November 2021 (N=1)												I	I						R	L		
December 2021 (N=2)												I	I							L		

Genes	S																														
Positions	9	25	26	67	75	105	136	143	144	152	190	215	241	242	243	244	440	449	477	478	484	501	585	614	655	679	681	716	859	879	1101
Wuhan Hu-1	P	P	P	A	G	I	C	V	Y	W	R	D	L	L	A	L	N	Y	S	T	E	N	L	D	H	N	P	T	T	A	H
June 2021 (N=7)	L	L					F		del		S	G			del	del		H			K	Y	F	G	Y	K		I	N		
July 2021 (N=10)	L	L					F	A	del	R	S	G			del	del		H		K	K	Y		G	Y	K	H	I	N	T	Q
August 2021 (N=16)	L	L		V			F	A	del	R	S	G			del	del		H		K	K	Y	F	G	Y	K	H	I	N	T	Q
September 2021 (N=11)	L	L				V	F		del	R	S	G			del	del		H		K	K	Y		G	Y	K		I	N	T	Q
October 2021 (N=4)	L	L	S				F		del		S	G			del	del		H	N	K	K	Y	F	G	Y	K		I	N	T	Q
November 2021 (N=1)	L	L					F		del		S	G	del	del	del			H			K	Y	F	G	Y	K		I	N		
December 2021 (N=2)	L	L					F		del		S	G	del	del	del	del		H			K	Y		G	Y	K		I	N		

*3C-like protease (3CLpro), envelope (E), membrane (M), nucleocapsid (N), non-structural protein (nsp), open-reading frame (ORF), papain-like protease (PLpro), RNA-dependent RNA-polymerase RdRp, spike (S)

Table 5.12: Mutations detected across the whole genome of C.2 lineage from 2020

Genes	3CLpro	N		nsp2	nsp4	nsp6	nsp14	nsp15	nsp16	ORF3a	ORF7a	PLpro	Rdrp		S			
Positions	**15**	**203**	**204**	**274**	**264**	**37**	**85**	**44**	**182**	**23**	**63**	**428**	**302**	**323**	**614**	**623**	**679**	**1219**
Wuhan Hu-1	**G**	**R**	**G**	**L**	**L**	**L**	**A**	**E**	**K**	**A**	**F**	**T**	**L**	**P**	**D**	**A**	**N**	**G**
July 2020 (N=1)	S	K	R		I		S					I	S	L	G			V
August 2020 (N=2)	S	K	R	F			S	V			del	I	S	L	G	V	K	
September 2020 (N=3)	S	K	R			F	S		N	S		I	S	L	G			

*3C-like protease (3CLpro), nucleocapsid (N), non-structural protein (nsp), open-reading frame (ORF), papain-like protease (PLpro), RNA-dependent RNA-polymerase RdRp, spike (S)

5.4. Discussion

The SARS-CoV-2 global pandemic is fueled by several variants of concern or interest that displayed more efficient transmission or immune evasion properties. As VOCs evolved from the Wuhan-Hu1, SARS-CoV-2 has undergone convergence resulting in novel lineages that could have evolved from previous lineages or VOCs (40). Our study describes the diversity of SARS-CoV-2 lineages and VOCs from the first to fifth COVID-19 waves in SA and the significance of lineages observed at low frequencies.

SARS-CoV-2 infected individuals were predominantly observed among adults aged 25–60 years and females, with no noticeable difference in the prevalence from 2020, 2021 or 2022. Globally, similar observations were made, with higher rates of infection observed among younger adults >25 years of age, with the majority of resultant deaths observed among older adults >60 years (13,145–148). The detection rate of SARS-CoV-2 almost doubled for females (60.1%) when compared to males (37.5%). Of note was that the majority of cases were from the City of Johannesburg Metro district in Gauteng, which is expected since it is the hub of major public hospitals in Johannesburg, clinics and a variety of SARS-CoV-2 community screening sites. Approximately 72.0% of samples sequenced belonged to individuals that sought community screen and test services, followed by out-patients, and therefore history and symptom profiles of these individuals were not recorded as was for in-patients. Of note was that community screening was 30.7% and 20.6% greater in 2022 compared to 2020 and 2021, respectively.

We identified 24 lineages, apart from the dominating VOCs in SA, from 2020 to 2022, which descended from the Wuhan-Hu1 strain, including B.l (3%; 8/278), B.1.1 (16%; 45/278), B.1.1.1 (2%; 5/278), B.1.1.348 (3%; 8/278), B.1.1.52 (5%; 13/278), C.1 (13%; 37/278), and C.2 (2%; 6/278), among which the D614G mutation was highly prevalent. Overall for SA, B1.1, B.1.1.52 and C.1 were detected at frequencies of 11%, 14% and 23%, respectively, before the emergence of Beta in 2020 (149,150), which is higher than observed for Gauteng in this

study except for B.1.1. These three lineages (B.1, B.1.1 and C.1) were the only low-frequency lineages that were not completely displaced and circulated during more than one wave. The B.1 lineage was predominant across Africa at the beginning of the pandemic (86). B.1 gained up to 21 mutations across the genome over time from 2020 to 2021. B.1 was detected at a high prevalence of 46% in the United States of America (USA) and 11% in Turkey, whereas for Africa, it was 26.7% compared to 2% overall for SA and 3% in this study (55,72). This lineage was responsible for the 2020 SARS-CoV-2 outbreak in Northern Italy (151). Similar to our study, B.1 was detected in Spain through waves 2 to 4 (152). B.1 was detected in less than 1% of SARS-CoV-2 genomes from Peru in 2021, whereas it was detected at a high frequency of 46.3% in Colombia (153,154).

The number of mutations in B.1.1 decreased from 20 to 8 by 2022. This lineage dominated in the UK at 25% and at 17% in the USA (55,72). B.1.1 was also reported as one of the dominant (32.9%) SARS-CoV-2 lineages in Spain during the first wave in 2020, whereas B.1 was present at 2.7% during the same time (152). In Peru, B.1.1 was detected at a frequency of 1%, while in Chile, it was reported as one of five lineages that dominated early in the pandemic (154,155). B.1.1.348 was detected in Argentina during the first wave of SARS-CoV-2 infections at a low frequency of 3.7% and was present in Peru at 3% and at a higher frequency of 9.7% in Colombia (154,156,157). Chile also reported B.1.1.348 as one of five dominant lineages early in the pandemic (155). B.1.1.348 was also detected in Spain in a few cases (152).

Globally C.1, C.1.2 and C.2 had a cumulative prevalence of <0.5%, with SA accounting for 94.6% (437/462), 89.5% (307/343), and 45.8% (33/72) of the data, respectively (55,72). C.1 had ~16 mutations in 2020 which remained somewhat consistent with up to 19 mutations by 2021, which was similar to findings in previous studies from SA with the majority of sequences geographically representative of KwaZulu-Natal and the Free State Provinces (47,158). However, the latter study detected B.1.1.54 (28.8%; 320/1 111) and B.1.1.56 (9.4%; 104/1 111) lineages (~14 mutations) from the parent lineage B.1.1 at high prevalence during the first wave across five SA provinces, but the majority of sequences were from KwaZulu-Natal (47). Whereas the B.1.1.53, B.1.1.54, B.1.1.117, B.1.237 and B.1.381, from the parent lineage B.1 and B.1.1, were the most prevalent in the Free State but were not detected in our study (158). This may suggest that B.1.1.348, B.1.1.52 and C.2 were more specific to Gauteng, especially within the City of Johannesburg Metro. C.1.2 (20D clade) evolved from the C.1 lineage (87)

and circulated at low frequencies in 2021 during the Delta wave. C.1.2 was a variant of interest for SA during 2021. It was only observed from June to December 2021 at a frequency of <10% during each month but never outcompeted other VOCs. C.1.2 was observed across all nine provinces of SA with <4% monthly prevalence from November 2021. While our study showed a slightly higher prevalence of 6.1% (53/870) and represented ~32% of all the South African C.1.2 strains described (87), C.1.2 had 23 highly prevalent mutations that occurred concurrently (co-evolved), with the exception of L585F (17%) and D936H (5%) observed in a previous study (143).

Significant mutations in the S protein were detected in lower frequency lineages and also characteristic in VOCs that were dominant in SA. Mutations including D614G, D215G, N440K, S477N, T478K, E484K and H655Y are responsible for enhancing ACE2 binding affinity, transmissibility or immune escape (50,75). The low-frequency lineages were primarily observed among adults aged 25–44 years from community screening and out-patients. There were several important mutations worth noting that were detected during more than one SARS-CoV-2 wave among the low-frequency lineages in this study. The S68F substitution in the E protein can also assist with stabilising the protein structure (159,160). A significant M protein mutation was the I82T, which is involved in increasing structural stability and was most prevalent in the B.1.575 lineages in the USA in 2020 (161,162). The N protein mutations, including P13L, R203K and G204R/K, may assist with increasing the transmission of the SARS-CoV-2 but also be associated with reduced severity of disease and, therefore, lower mortality rates when compared to individuals with the wild-type N protein (163–165).

In the nsp1 protein, the E102K mutation encourages attachment to viral RNA, increasing replication and thereby increasing infectivity (166). While del141-143 in nsp1, also reported in Omicron BA.4 lineages during the fifth wave in SA and described in Delta AY.63 in Norway, extends the duration of infection (138,167). R126S in ORF3a compromises the viral structure by reducing its stability (168). P323L was observed across all low-frequency lineages and was previously described as responsible for maintaining the RdRp protein structure and may enhance mutation rates (169,170). Among the S protein mutation in the low-frequency lineages, the most significant were the following: N501Y which encourages affinity to human ACE-2 (49); del69-70 and delY144 may influence the antigenic properties of the S protein (NTD); furthermore, del69-70 escapes neutralising antibodies, increases viral replication and

transmissibility (49,171–173); E484K assists the virus by escaping host immunity and increased binding affinity to the hosts ACE-2 receptor (173); D614G was identified in all VOCs and found to improve ACE-2 receptor binding thereby increasing viral transmissibility (174–176); H655Y was also detected in Gamma and Omicron, and increases virulence by encouraging furin cleavage, evading immunity and increasing transmission; N679K increases glycosylation at the S1/S2 furin cleavage site which could prevent syncytia formation (163,177).

The circulation of SARS-CoV-2 VOC were similar to that reported previously in SA during the first five COVID-19 waves (89,92). Wuhan-Hu1 ancestral strains driving the first wave (2020: epiweeks 14–42), Beta (2020: epiweeks 43 to 2021: epiweek 25), Delta (2021: epiweeks 20 to 45), Omicron BA.1/BA.2 (2021: epiweek 46 to 2022: epiweek 12) and Omicron BA.4/BA.5 (2022: epiweeks 13 to 42) driving the 2nd, 3rd, 4th and 5th waves, respectively. Several studies have previously reported on the circulation of VOCs in SA, which, except for Alpha, drove the major waves of infection from 2020 to 2022 (43,75,85,86,108). We identified Alpha in 2021 between the Beta and Delta waves; however, it was never a dominant lineage and only circulated scantly across 9 weeks. Beta evolved from the B.1.1 lineage (20C clade), with up to 37 mutations gained across the genome by the end of the second wave. After its initial detection in SA, Beta spread across 20 countries in Africa (86). This VOC had eight mutations in the S protein, with the most significant being K417N, E484K and N501Y within the RBD (43,55,56). Beta was not completely displaced after the second wave since we detected it at low frequencies later in 2021 towards the tail-end of the Delta wave, as well as in several other provinces in SA and <0.1% globally (89,178).

We identified 14 different Delta sub-lineages, with AY.45 dominating from epiweeks 20 to 45 in 2021. Similarly, Delta AY.45 (21J clade) had an overall prevalence of 80% (522/650) by the end of October 2021 (epiweek 45), with Western Cape being the only province with a similar prevalence of B.1.617.2 (21A clade) and AY.45 (89). Phylogenetic analysis showed that Delta descended from the B.1 lineage (20A clade) and gained up to 51 mutations across the genome by the end of the third wave, with 9–10 mutations in the S protein (56). L452R and P681R were the common mutations identified in Delta sub-lineages, while T478K was the primary mutation in the S protein among the Delta lineages, which is responsible for interacting with the human ACE-2 receptor, increasing its infectivity (54–56). From our study, Beta and Delta

circulated for up to 30 consecutive weeks; however, the detection rate of Delta was much greater than Beta, suggesting Delta has higher rates of viral replication and transmission (54).

Omicron outcompeted Delta unexpectedly and at rapid rates in 2021, with a detection rate of 30.2% (422/1399) in 2021 and 100% (870/870) in 2022. Similar events were observed globally (55). Phylogenetic analysis showed that Omicron lineages and sub-lineages disseminated from the 20B clade. We identified 37 Omicron sub-lineages from epiweek 43 in 2021 to epiweek 42 in 2022, including several recombinants (BF.28, CP.5, XAB, XAM, XAR, XAS, XAY and XT) sporadically detected with a prevalence of 2.4% (21/870). BQ.1, XD, XE, XF and XBB recombinants also identified in our study were reported to display a significant increase in infectivity and immune evasion (179,180). As Omicron evolved from BA.1 to BA.5, the virus gained >80 mutations across its genome, up to 34 mutations in the S protein. Evidence of convergence was present in Omicron, which carried over a few important mutations from Beta (K417N, N501Y), Alpha (P681R and del69/70) and Delta (T478K and P681H) (56,88,181). These important mutations were known to increase the spike trimer stability and were present in the absence of host immune pressure (181). BA.1 was outcompeted by BA.2, and the del69/70 fell away, which was the distinguishing feature between BA.1 and BA.2 (100,138). Omicron BA.2 was then outcompeted by BA.4 and BA.5, which gained the del69/70, L452R, F486V and Q493 wild type (138).

The main limitation identified in this study was the inconsistency of sequence data in 2020. The number of genomes sequenced in 2020 was limited since routine surveillance commenced late in 2020, and retrospective sequencing had to be performed. However, our dataset proves to be a good representation due to random selection, as the major VOC events observed in SA are reported accurately when compared to the national surveillance data. In addition, the emergence of novel lineages and VOC primer mismatches resulted in challenges in obtaining high-quality sequence data with adequate coverage across the genome. To overcome this, primer optimisation was required to prevent incorrect lineage/VOC assignment. Artic primer version used was changed from V3 to V4.1 for samples sequenced from 2021 to 2022 and not retrospectively for samples from 2020 due to the high costs of sequencing. Therefore, in this study, only genome sequences with 65%-99% coverage across the genome and with >80% coverage for the S proteins were included.

5.5. Conclusions

Our study showed that with the emergence of novel lineages and VOCs, the number of mutations increased simultaneously, reducing protection against vaccine-induced or natural immune response and other environmental selective pressures. In addition, we identified that B.1, B.1.1, and C.1 lineages were able to retain mutations over time and still maintained their standing during the five COVID-19 waves in SA, although at much lower frequencies. These lineages presented with mutations that were carried over with the VOCs. It is possible that low-frequency lineages, together with VOCs circulating, can lead to possible convergence and recombination, which may result in the next novel lineage or variant that may further increase transmissibility, infectivity and escape vaccine-induced or natural host immunity. Accounting for diversity in vaccine design and universal vaccine design strategies may be necessary to ensure effective vaccines to protect against all variants of SARS-CoV-2.

We therefore aimed to generate an appropriate vaccine construct derived from SARS-CoV-2 S protein gene sequences from our South African population, which accounts for increased diversity by maximizing inclusion of targeted CTL epitopes present in the dataset. This strategy may generate a potential S immunogen capable of inducing broad protective immune responses against previous and emerging variants or lineages of SARS-CoV-2.

CHAPTER 6

Utilising SARS-CoV-2 sequence diversity to design a potential spike (S) immunogen

6.1. Introduction

The S protein is one of four immunogenic proteins in SARS-CoV-2 (182,183). It comprises the S1 domain which houses the receptor binding domain (RBD) and the N-terminal domain (NTD) while the S2 domain contains the fusion peptide (182). The S protein, specifically the RBD and NTD regions, is the primary target for neutralising antibodies and facilitates viral entry via the angiotensin-converting enzyme 2 (ACE2) receptor on the host cells (183–185). Once the innate immune response is activated the progression of adaptive response is evident. Cytotoxic T cells including CD8+ T cells and CD4+ T cells play an important role in adaptive immune response to SARS-CoV-2 by inhibiting the spread of the virus (183). The CD4+ T cells are activated in response to initial infection at high levels, while CD8+ T cells are detected after infection (185). The combination of CD4+ and CD8+ T cell responses are usually observed during mild disease which is evidence of a protective immune response against COVID-19 (185). Increased levels of CD8+ T cells have been detected in individuals with moderate disease severity which may be an indication of correlation between adaptive cellular immune response and well developed disease management systems (185). In contrast, CD8+ and CD4 +T cells are significantly reduced in individuals with severe SARS-CoV-2 infection (185). However, neutralising antibody levels is greater among individuals infected with SARS-CoV-2 that become critical (185). SARS-CoV-2 T cells display cytotoxic properties during the acute phase of infection which is linked to clinical markers of disease severity (185).

Mutations in the RBD and NTD of the S protein promote viral fitness and improve transmissibility of the virus between hosts as described with the N501Y in the Alpha, Beta and Gamma variants (186,187). Some substitutions in Beta (K417N, E484K) and Gamma (K417T, E484K) are responsible for escaping natural or vaccine mediated immune responses. In South Africa, Beta was shown to escape class 1 and 2 antibodies (188) and sera from patients infected with Beta showed cross-neutralising antibody activity against wild type and Gamma (189). In another South African study, CD4+T cells that targeting the wild type did not respond to

epitopes from the Beta variant in the same manner, which was likely a result of the increased number of mutations (190). VOCs can evade the hosts immunity through which it escapes the CD8+T cell response (191). Upon vaccination against SARS-CoV-2 the T cell epitopes induce humoral and cellular responses that prevents the viraling infectivity. A previous study showed that Beta and Alpha could not evade T cell mediated response after the vaccination with the Pfizer BNT162b2, unlike some of the Omicrons variants that still evade the immune response after vaccination (191,192). A previous study showed that individuals recovering from SARS-CoV-2 wild type developed antibodies that targeted several epitopes of the S protein in emerging VOC (193). The latter study reported 12 monoclonal antibodies (mAbs) that had cross-neutralising activity against Alpha, Gamma and Delta by primarily targeting epitopes in RBD and NTD. Class 2 mAbs (RBD) specifically neutralise Alpha and Delta variants, while class 3 mAbs neutralise Alpha and Gamma (193). Unlike other VOCs, more than 15 mutations detected in the RBD of Omicron and its sub-lineages, enhances transmission and infection rates, target neutralising antibodies and vaccine derived epitopes, making it necessary to improve vaccine design (194). Interactions with Omicron RBD and the ACE2 are enhanced through mutations including Q403RK, K417N, N440K, G446S, S447N, T478K, E484A, Q498R, and N501Y, and encourage immune evasion and resistance to neutralising activity from convalescent sera, vaccine-induced antibodies and mAbs (191,194,195). In South Africa, individuals vaccinated with Pfizer BNT162b2 prior to the emergence of Omicron, gained some protection against Omicron as residual neutralisation was evident (196).

Global data has shown that SARS-CoV-2 to some extent has reduced cellular immune evasion with time, suggesting that T cell responses induced through natural infection or vaccination assist with controlling emerging VOCs. Furthermore, natural infection and mRNA vaccines elicit CD4+ T cell response that can protect individuals against severe disease. With the rise in Omicron cases and the continual emergence of novel sub-lineages, the CD8+ T cell response is reduced substantially compared to CD4+ T cells which is evidence of immune evasion. Data is currently only available for the immune responses targeting Beta and very limited data is available for Omicron in South Africa. This information is necessary to derive more significant and targeted S protein epitopes for vaccine design improvement for South Africa. We therefore aimed to use predicted population level CTL epitopes to maximise the inclusion of variant epitopes using a mosaic vaccine design tool (81,197).

6.2. Methods

6.2.1. Sequence data

South African SARS-CoV-2 samples sequenced for this study, from March 2020 to April 2022 (N=2371), were downloaded from GISAID. Sequence data can be accessed from GISAID using EPI_SET ID: 231005gn (doi:10.55876/gis8.231005gn). GISAID data access agreement policy (https://www.gisaid.org/about-us/acknowledgements/sharing-policy/) was adhered to. Sequences were aligned using AliView software and translated into amino acids and the S protein region was trimmed.

b) Methods used as described in 2.4.6 to 2.4.11 and further summarised in Fig 6.1.

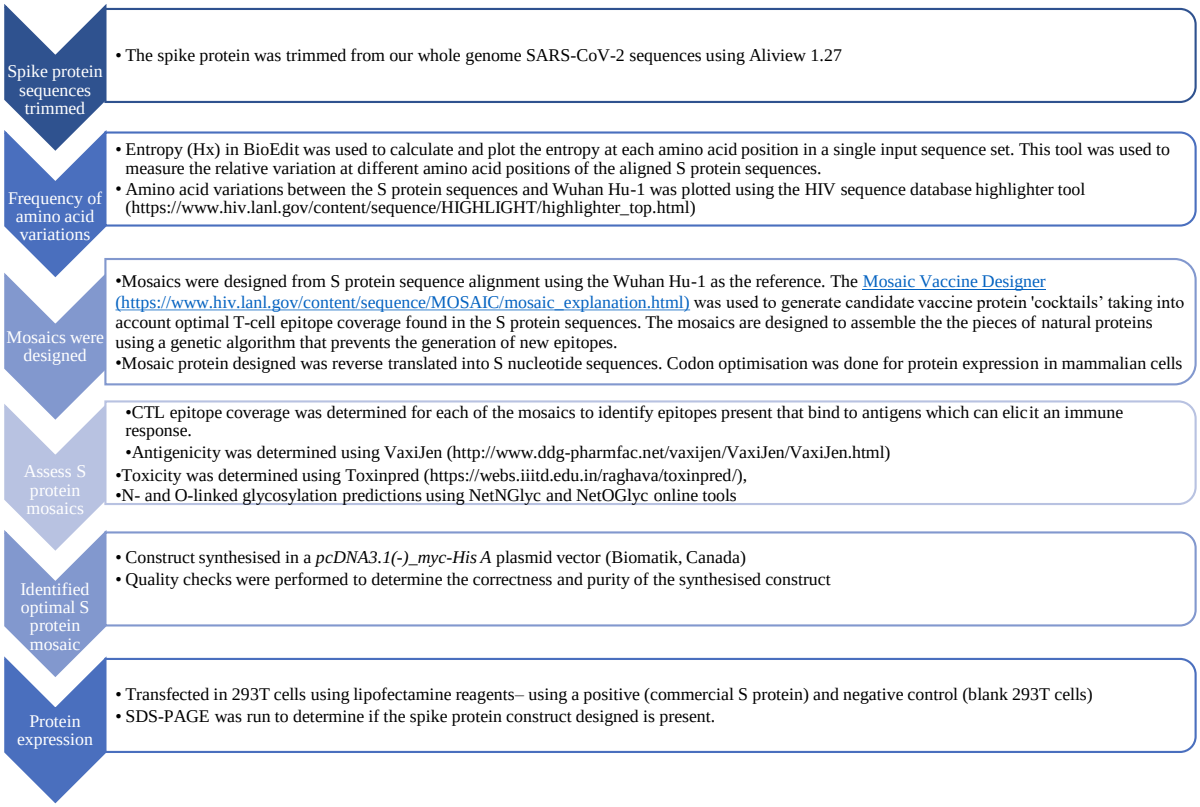

Fig 6.1: Workflow described to design and test the spike protein construct

6.3. Results

6.3.1. Diversity of the S protein sequences

Of the 2371 sequences available from this study, only unique sequences (N=1732) were included in this study. The sequence data included a diverse array of lineages and variants: wild type lineages (B and C lineages), Beta, Delta, Omicron BA.1, Omicron BA.2, and recombinants (Fig 6.2). The diversity of the sequences was analysed based on the amino acid

variations across the aligned S protein in comparison to the Wuhan Hu-1 reference sequence (Fig 6.4). Amino acid positions from 1 to 570 and from 650 to 1 070 had an increase in entropy indicating higher diversity in these regions (Fig.6.3). The region from 1 to 530 includes the NTD and RBD of the S1 subunit. Overall, the dataset provided good sequence coverage across the S protein region compared to the reference sequence.

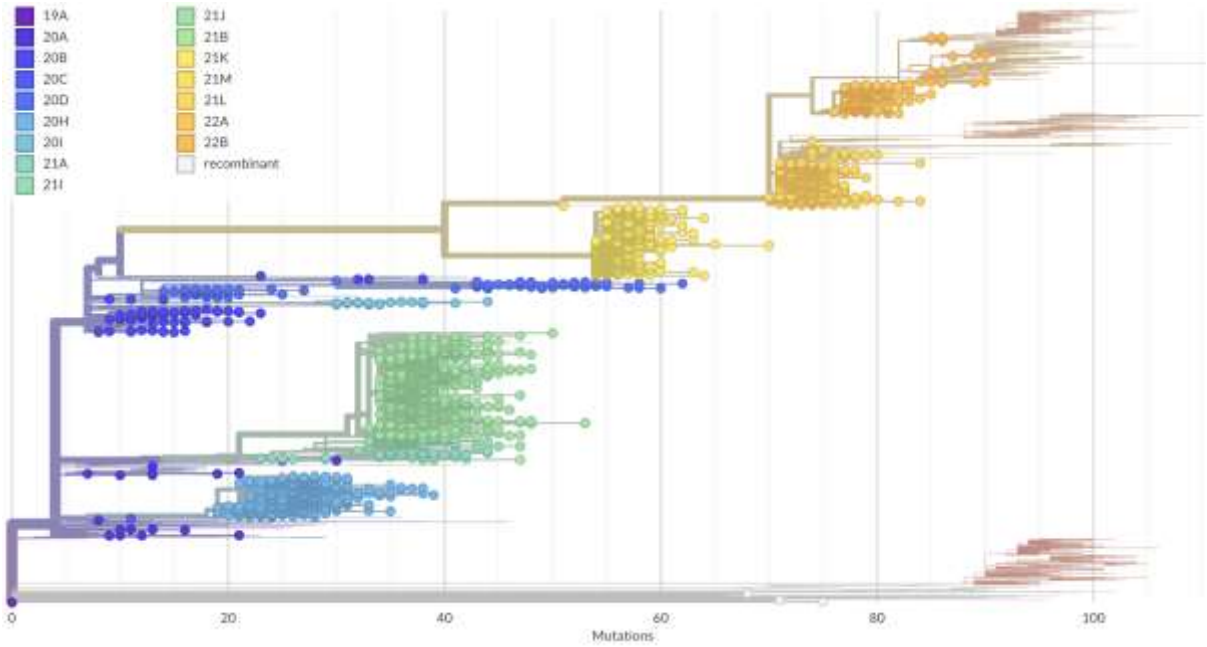

Fig. 6.2: Phylogenetic tree displaying our sequence data from the wild type to the Omicron waves, March 2020 to April 2022. The data from this study is represented as coloured circles under its respective clusters compared to global data. The additional lines without circles represent the global sequence data.

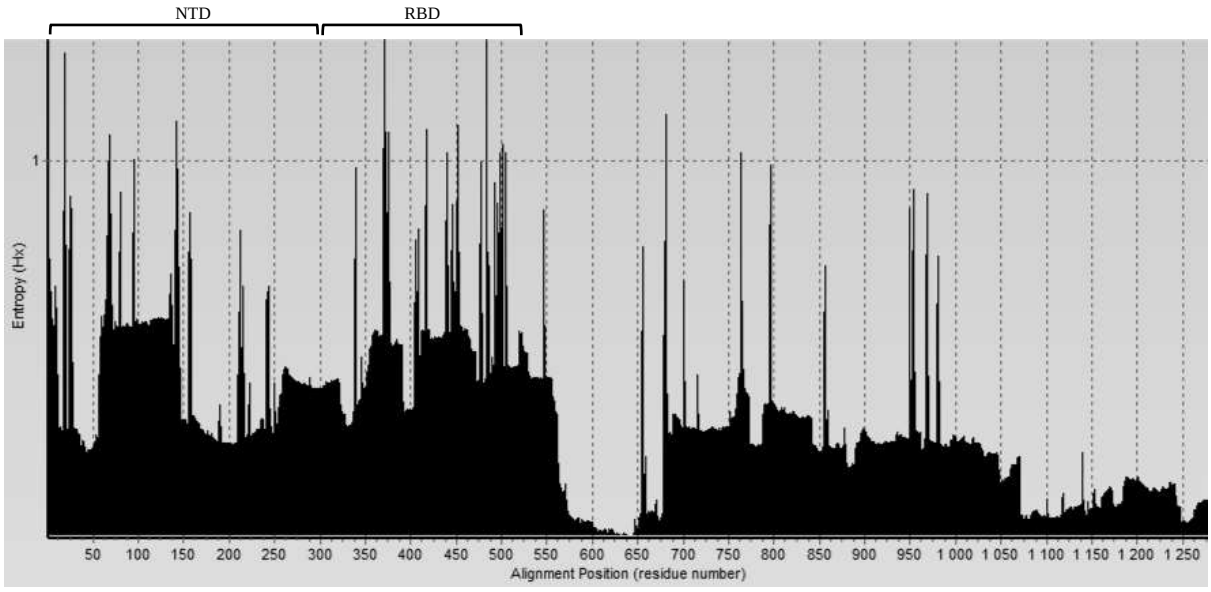

Fig. 6.3: Entropy plot of amino acid sequence coverage compared to the Wuhan-Hu1 reference strain. Data includes the spike protein of all South African sequences from 2020 to 2022, from our study. The black shaded areas represent the amino acid similarities at an amino acid position. The grey area represents amino acid variations compared to the Wuhan-Hu1 strain.

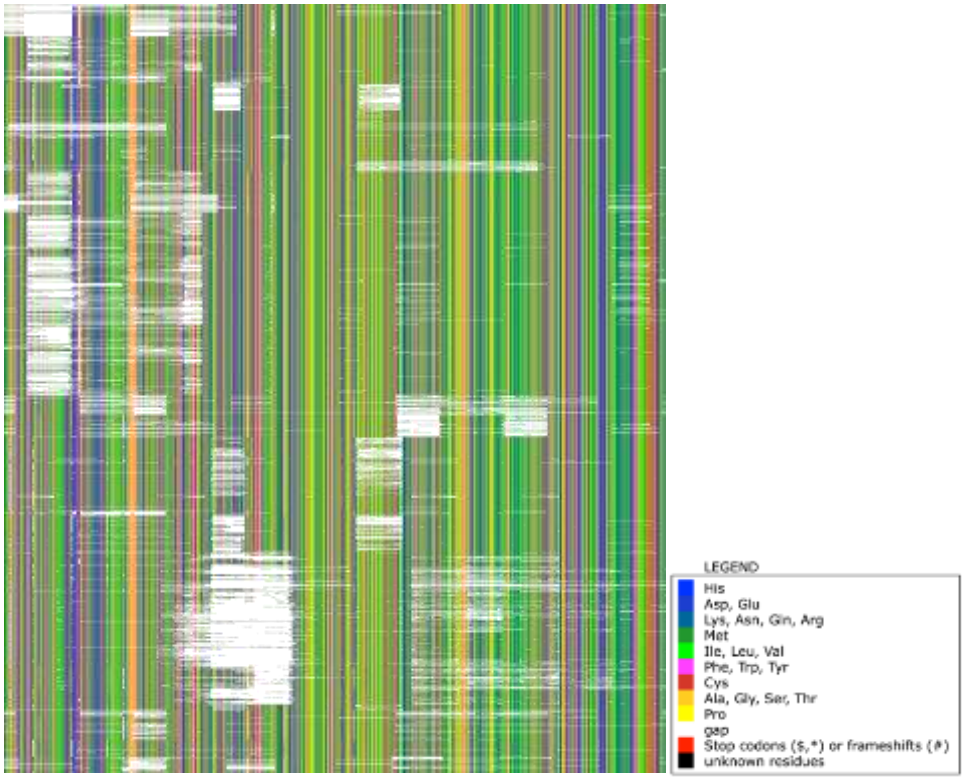

Fig 6.4: Spike protein sequence data aligned against the Wuhan Hu-1 reference (first row) sequence. The amino acid positions are represented on the horizontal axis and the vertical axis represents each spike protein sequence. Each colour indicates a different amino acid. White represents gaps present in the sequence data.

6.3.2. Generation of a mosaic S peptide cocktail and identifying the representative three best natural S sequence matches

Figure 6.5 displays the CTL epitope coverage across each S protein sequence in the dataset for peptide cocktails derived from the Wuhan sequence as reference and each of the three best natural sequences individually as well as combined. The epitope coverage tool calculates the fraction of all the epitope-length peptides in the S protein sequences that are covered or matched by the peptides in the protein cocktail. Best 1 had greater coverage and greater unique 9-mer epitope coverage compared to Wuhan Hu-1 (Fig 6.5 and 6.6b). A slight decrease in unique 9-mer epitopes were observed in Best 2 (Fig 6.5 amnd 6.6c). However, Best 2 shows slightly improved coverage for some regions not well covered by Best 1 or Wuhan (Fig 6.5). The black regions (Fig 6.5) marked poor CTL epitope recognition in Best 3 represented by missing information in the S protein sequence, also indicated in Fig 6.5. It is clear that the peptide cocktail generated from the combination of all 3 best natural sequences (indicated by Best 1,2,3) provided optimal CTL epitope coverage against all sequences in the dataset (Fig 6.5 and 6.6f). The mosaic peptide cocktail generated from 3 best natural sequences increased the overall predicted exact match CTL epitope coverage from ~95% to 98% (Fig 6.5).

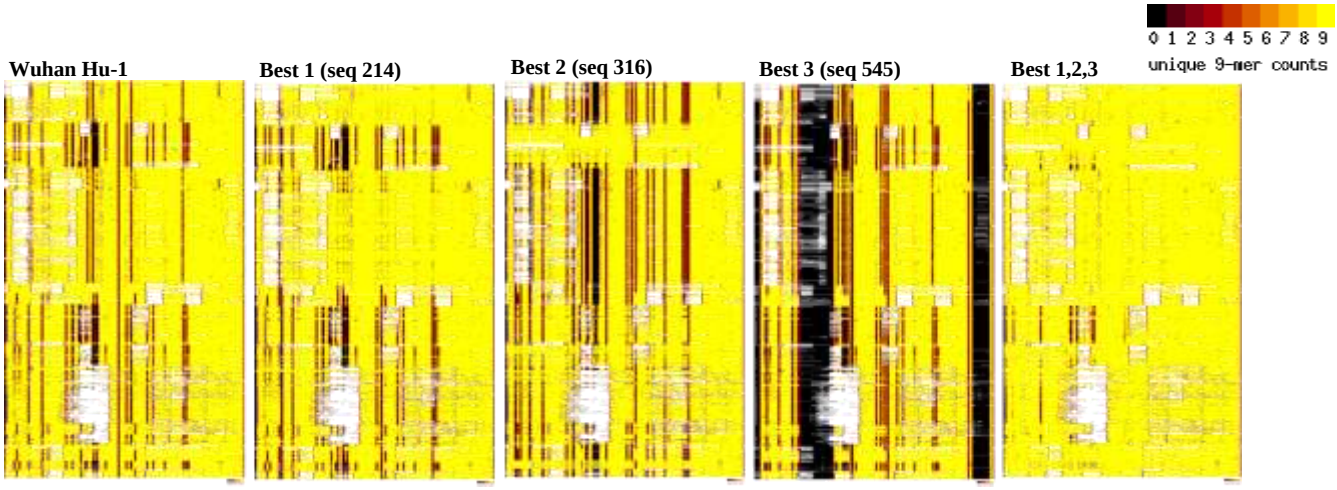

Fig. 6.5: The unique 9-mer coverage of the spike peptides with mosaic alignments representing 4 possible scenarios for the 3 best natural sequences (214, 316, 545) identified as representative of the mosaic peptide cocktail. Best 1 represents seq 214 or hCoV-19/SouthAfrica/CERI-KRISP-K027930/2021, Best 2 represents seq 316 or hCoV-19/SouthAfrica/CERI-KRISP-K033221/2021, Best 3 represents seq 545 or hCoV-19/SouthAfrica/KRISP-K018784/2021**).** The dark areas indicate poor CTL epitope coverage associated with the amino acid mutations in the spike protein of SARS-CoV-2. Identical sequences were removed, errors, gaps and non-specific nucleotides were removed.

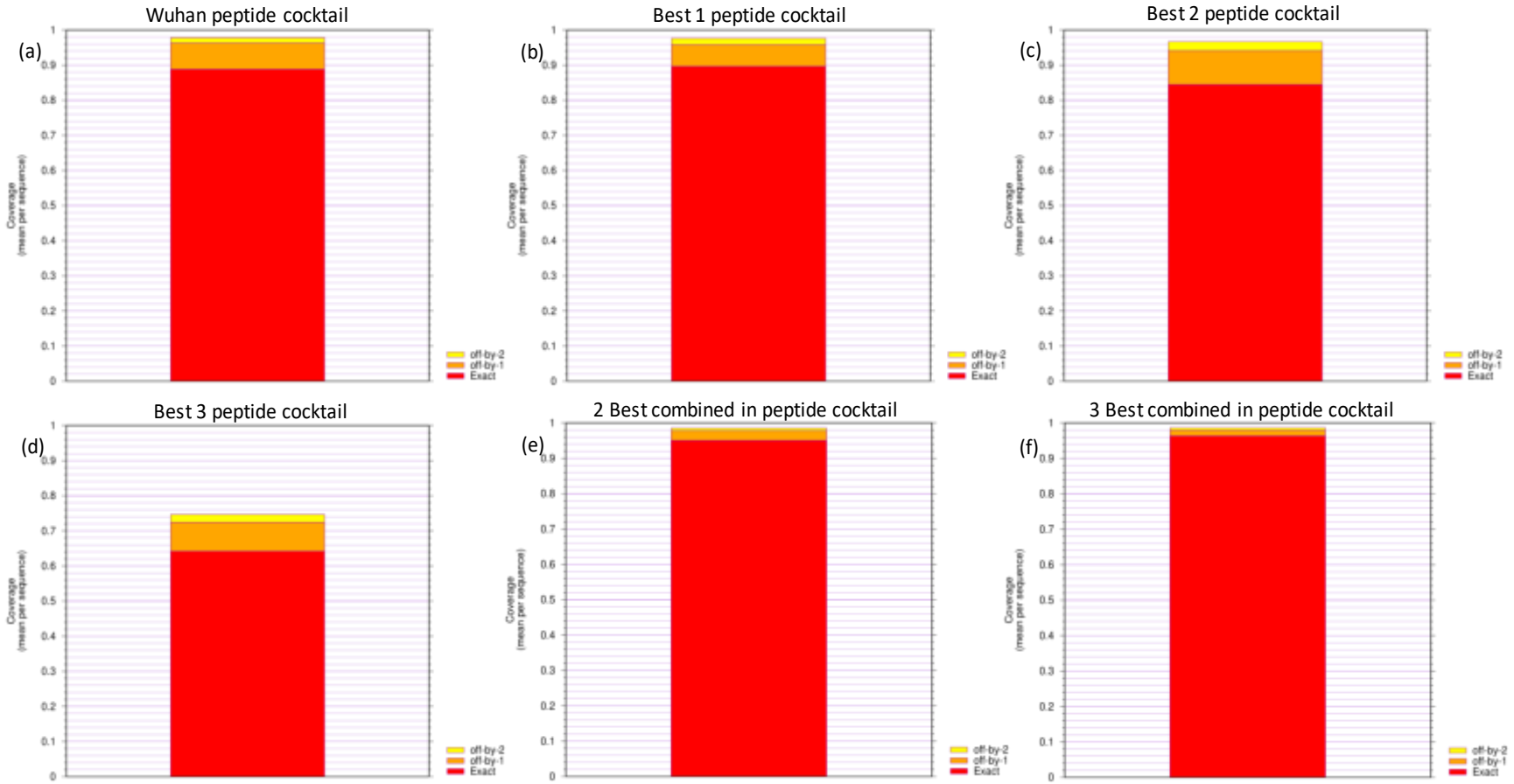

Fig. 6.6: Epitope coverage of spike peptide mosaic. **Graphs a to f display the mean coverage of potential 9-mer T cell epitopes over the S protein sequence covered by CTL epitope in a mosaic cocktail. The coverage is described as the average number of potential epitopes per natural strain that are matched per natural strain.** The results indicate the coverage (mean per-sequence) of S protein. The red area (exact) represents 100% match i.e., 9 amino acids out of 9 amino acids; the orange area (off-by-1) represents a match at 8 amino acids out of 9 amino acids; the yellow area (off-by-2) represents a match at 7 amino acids out of 9 amino acids.

To further assess the optimal S immunogen, positional 9-mer epitope coverage was analysed. It is important to identify an immunogen with high epitope coverage to determine if the selected immunogen elicits a host immune response. Mosaic peptides generated by combination of the 3 best natural matched sequences improved predicted positional epitope coverage (Fig 6.7). The positional epitope coverage for the selected S immunogen for 9-mer peptide epitopes had an exact match with coverage of 75 to 100% across most of the amino acid positions. Positions 200, 380, and 400 had 8/9 match with a coverage of 62%, 58%, and 60%, respectively (Fig 6.7). The optimal S immunogen was named SC2M2 and was subjected to further analysis.

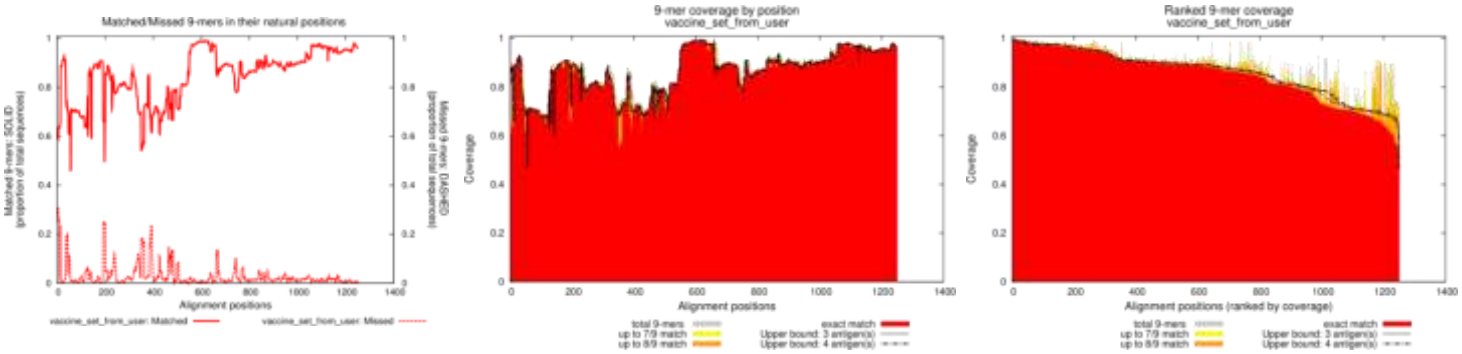

Fig. 6.7: Graphical display of predicted positional 9-mer epitope coverage for the three best natural sequences that will cover CTL epitopes present in the derived mosaic protein. The x-axis displays alignment positions where each value represents a column in the sequence alignment, ranked from left to right by exact-match proportion (descending). The y-axis represents the proportion of background sequences at a position that has 9-mers matching at different level in the sequences of the antigen/peptide set. Graph on the right represent the matched 9-mers compared to upper bounds (ranked): proportion of 9-mers in the test set (S protein of SARS-CoV-2 vaccine strains) that exactly matches the antigen/peptide set (S protein mosaic), across the alignment. The red area (exact) represents 100% match i.e., 9 amino acids out of 9 amino acids; the orange area (off-by-1) represents a match at 8 amino acids out of 9 amino acids; the yellow area (off-by-2) represents a match at 7 amino acids out of 9 amino acids.

6.3.3. Codon optimised S immunogen

The optimal S protein mosaic was derived from a combination of the 3 best natural S protein sequences (Fig 6.8) and reverse translated to nucleotides. Codon optimisation was performed on the nucleotide sequence to promote the expression of S immunogen (SC2M2) in mammalian cells. Prior to optimisation, the SC2M2 sequence had a GC content of 48.59 and a codon adaptive index (CAI) value of 0.27 with a very unstable relativeness adaptiveness ranging from 0.01 to 0.99 (Fig 6.9). The CAI is the measure of relative adaptiveness of the codon usage of a gene compared to the codon usage of an expressed gene. The codon optimisation was performed using the JCAT codon optimiser tool (http://www.jcat.de/) for which we obtained a

GC content of 63.6 and a codon adaptive index (CAI)-value of 0.95 for the improved sequence, with >90% coverage (Fig 6.9). The final codon optimised SC2M2 nucleotide sequence is shown in Fig 6.10.

```
ASMFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGT
KRFDNPVLPFNDGVYFASIEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVIKVCEFQFCNDPFLDVYYHKNNKSWME
SGFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRDLPQGFSALEPLVDLPIG
INITRFQTLLALHRSYLTPGDSSSGWTAGAAAYYVGYLQPRTFLLKYNENGTITDAVDCALDPLSETKCTLKSFTVEK
GIYQTSNFRVQPTESIVRFPNITNLCPFDEVFNATRFASVYAWNRKRISNCVADYSVLYNLAPFFTFKCYGVSPTKLN
DLCFTNVYADSFVIRGDEVRQIAPGQTGNIADYNYKLPDDFTGCVIAWNSNKLDSKVSGNYNYLYRLFRKSNLKPFER
DISTEIYQAGNKPCNGVAGFNCYFPLRSYSFRPTYGVGHQPYRVVVLSFELLHAPATVCGPKKSTNLVKNKCVNFNFN
GLKGTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQGVNCTEVPV
AIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEYVNNSYECDIPIGAGICASYQTQTKSHRRARSVASQSIIAYTMSLG
VENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLKRALTGIAVEQDKNT
QEVFAQVKQIYKTPPIKYFGGFNFSQILPDPSKPSKRSFIEDLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFK
GLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKI
QDSLSSTASALGKLQDVVNHNAQALNTLVKQLSSKFGAISSVLNDIFSRLDKVEAEVQIDRLITGRLQSLQTYVTQQL
IRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPAICHDGKAHFP
REGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGD
ISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYIWLGFIAGLIAIVMVTIMLCCMTSCCSCLKG
CCSCGSCCKFDEDDSEPVLKGVKLHYTDI
```

Fig. 6.8: Protein sequence of the optimised S protein mosaic (SC2M2) obtained from the 3 best natural protein sequences

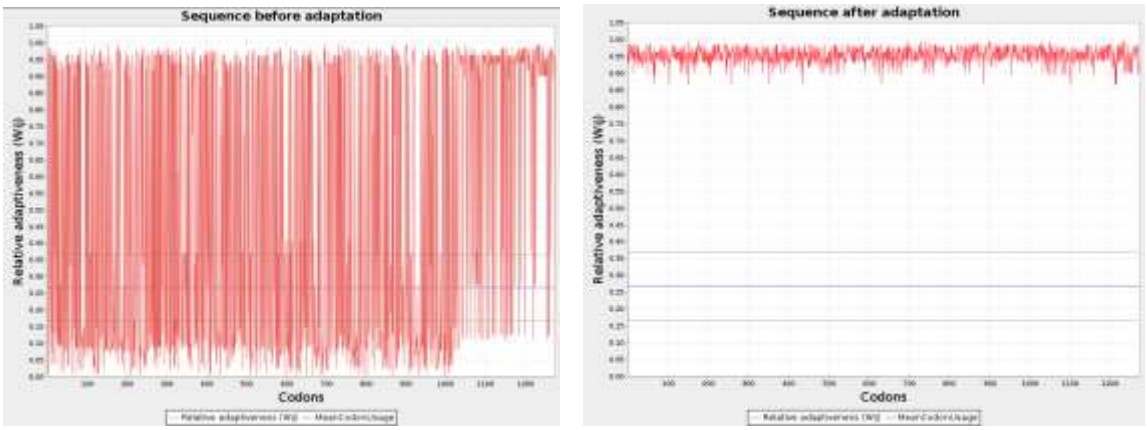

Fig. 6.9: Codon usage and coverage of the SC2M2 before and after codon optimisation. Left displays unadapted codon usage. Right represents the SC2M2 adapted codon including the elimination of restriction enzyme cleavage sites. The blue line represents the mean codon usage of SAR-CoV-2 (0.27). The gray lines above and below the mean codon usage indicate the standard deviation of the codon usage (0.17-0.37).

GCTAGCATGTTCGTATTTCTCGTGCTGCTCCCATTGGTGAGTTCACAGTGTGTGAACTTGACCACGCGCACACAGTTGCCCCCGGCATACACAAATAG
CTTTACTAGGGGTGTATACTATCCCGATAAAGTTTTCAGGTCTAGTGTGCTCCATTCCACCCAAGACTTGTTCCTCCCCTTCTTTTCTAATGTAACCT
GGTTTCACGCTATCCACGTAAGCGGAACTAACGGTACCAAGAGATTCGACAATCCCGTTCTGCCATTTAATGATGGTGTTTATTTTGCTTCTATTGAG
AAAAGCAATATCATTCGCGGTTGGATCTTTGGTACGACTCTCGACAGTAAAACCCAATCTCTGCTTATTGTAAATAACGCTACGAATGTGGTAATCAA
AGTATGTGAGTTCCAGTTTTGCAACGATCCCTTTCTTGATGTGTACTACCATAAAAACAATAAAAGTTGGATGGAAAGCGGCTTTCGTGTCTACAGTA
GTGCAAATAACTGTACGTTTGAATATGTCTCACAGCCCTTTTTGATGGACCTGGAAGGGAAGCAGGGTAATTTTAAAAATTTGAGGGAATTTGTTTTC
AAAAACATCGACGGATACTTTAAGATCTACTCAAAACACACGCCCATAAATCTTGTGAGGGATCTTCCACAAGGTTTCAGTGCGCTCGAGCCGCTTGT
GGACCTCCCTATTGGTATAAATATCACCCGATTCCAGACGCTCCTGGCGCTTCATAGATCTTACTTGACCCCGGGAGACAGCTCTTCAGGCTGGACAG
CGGGTGCCGCAGCGTACTATGTGGGATACTTGCAGCCACGAACCTTTCTTCTGAAGTATAATGAGAATGGTACCATTACCGACGCGGTGGATTGTGCG
TTGGACCCGCTCTCTGAGACAAAATGTACCCTGAAGTCATTCACCGTTGAAAAGGGAATCTATCAGACCTCTAACTTCAGAGTGCAACCAACAGAGTC
AATCGTTCGGTTTCCAAACATCACCAATCTGTGCCCCTTTGATGAAGTCTTTAATGCTACGAGGTTCGCGTCTGTCTATGCCTGGAATCGAAAGAGAA
TTTCTAATTGCGTTGCAGACTACTCCGTGCTGTACAATCTCGCACCTTTCTTTACGTTTAAATGCTACGGTGTAAGCCCAACCAAGTTGAACGATTTG
TGCTTTACCAACGTTTATGCCGACTCCTTTGTGATCAGGGGGGACGAAGTAAGGCAGATAGCACCCGGCCAGACGGGAAACATCGCCGATTATAATTA
CAAATTGCCAGACGATTTTACTGGTTGTGTTATAGCTTGGAATTCCAATAAGCTCGACTCCAAAGTGAGTGGAAACTATAACTACCTTTACCGGCTTT
TCCGCAAAAGCAACTTGAAACCGTTCGAACGAGATATTTCAACCGAAATCTATCAGGCAGGGAACAAGCCATGTAACGGAGTCGCTGGCTTTAACTGC
TATTTTCCGCTTAGGTCCTATTCTTTCCGACCGACCTACGGCGTCGGACACCAGCCTTACCGAGTCGTAGTCCTTTCCTTCGAGTTGCTCCACGCCCC
AGCTACTGTATGCGGTCCGAAGAAAAGCACCAACCTCGTTAAGAACAAATGTGTGAATTTTAACTTCAATGGCTTGAAGGGCACCGGCGTACTGACGG
AATCTAACAAGAAATTCCTCCCATTTCAACAATTCGGGCGAGACATCGCGGACACTACAGACGCTGTTCGAGACCCACAAACGCTTGAGATTCTGGAC
ATTACACCCTGCTCATTCGGGGGGGTCTCAGTCATAACACCAGGGACTAATACATCAAACCAAGTCGCAGTACTGTACCAAGGGGTTAATTGTACAGA
GGTGCCGGTCGCTATACACGCCGATCAACTTACTCCCACCTGGAGAGTATACAGTACGGGGAGTAACGTGTTTCAGACGAGGGCTGGCTGTTTGATAG
GCGCTGAGTATGTGAACAATAGCTATGAGTGCGACATACCAATCGGGGCAGGGATTTGTGCGTCATATCAAACTCAAACGAAGAGTCATCGACGCGCT
CGCAGCGTAGCTTCTCAATCAATAATAGCCTACACCATGAGTCTTGGTGTTGAAAACTCCGTTGCCTACAGTAACAACTCTATTGCGATCCCAACTAA
CTTCACCATTTCCGTGACCACCGAAATCCTTCCAGTGTCAATGACTAAGACAAGCGTGGATTGCACTATGTATATTTGTGGGGACTCCACTGAGTGTA
GTAATCTCTTGCTGCAGTATGGGTCATTTTGTACGCAGCTTAAGAGGGCATTGACTGGAATTGCCGTTGAACAAGATAAAAATACTCAAGAGGTATTT
GCACAAGTGAAGCAAATCTATAAGACGCCGCCGATAAAATATTTTGGGGGCTTTAATTTTAGTCAGATACTGCCTGATCCATCCAAGCCCAGCAAGCG
ATCCTTTATCGAGGACCTTCTCTTCAACAAAGTGACGCTGGCTGACGCCGGTTTCATCAAACAGTATGGAGATTGTCTTGGTGATATCGCGGCGCGAG
ACCTTATTTGCGCTCAAAAGTTTAAGGGGCTTACTGTTCTCCCTCCTCTCCTGACAGATGAGATGATAGCACAGTACACCTCCGCCTTGCTTGCCGGT
ACAATTACCAGTGGCTGGACATTCGGGGCTGGAGCGGCACTCCAAATTCCTTTCGCCATGCAGATGGCGTATCGCTTTAACGGGATAGGAGTTACGCA
AAATGTCCTTTACGAGAACCAGAAGCTGATTGCTAATCAATTTAATTCTGCGATTGGAAAAATACAAGATTCACTCAGCTCCACTGCATCTGCCTTGG
GAAAGCTTCAAGATGTGGTAAATCATAACGCCCAGGCTCTGAATACCCTCGTGAAACAACTCTCCAGTAAGTTTGGCGCCATCAGCTCAGTTCTTAAC
GATATCTTTTCACGACTTGATAAAGTCGAAGCTGAAGTCCAGATCGACAGACTGATAACAGGAAGACTGCAGAGTCTGCAAACCTATGTAACCCAACA
GTTGATTCGAGCTGCTGAGATCAGAGCCAGTGCAAACCTCGCTGCGACGAAGATGTCTGAATGCGTGCTGGGCCAGAGCAAGCGCGTGGATTTTTGCG
GCAAGGGCTACCACCTGATGAGCTTTCCGCAGAGCGCGCCGCACGGCGTGGTGTTTCTGCACGTGACCTACGTGCCGGCGCAGGAGAAGAACTTTACC
ACCGCGCCGGCGATCTGCCACGATGGCAAGGCGCACTTTCCGCGCGAGGGCGTGTTTGTGAGCAACGGCACCCACTGGTTTGTGACCCAGCGCAACTT
TTACGAGCCGCAGATCATCACCACCGATAACACCTTTGTGAGCGGCAACTGCGATGTGGTGATCGGCATCGTGAACAACACCGTGTACGATCCGCTGC
AGCCGGAGCTGGATAGCTTTAAGGAGGAGCTGGATAAGTACTTTAAGAACCACACCAGCCCGGATGTGGATCTGGGCGATATCAGCGGCATCAACGCG
AGCGTGGTGAACATCCAGAAGGAGATCGATCGCCTGAACGAGGTGGCGAAGAACCTGAACGAGAGCCTGATCGATCTGCAGGAGCTGGGCAAGTACGA
GCAGTACATCAAGTGGCCGTGGTACATCTGGCTGGGCTTTATCGCGGGCCTGATCGCGATCGTGATGGTGACCATCATGCTGTGCTGCATGACCAGCT
GCTGCAGCTGCCTGAAGGGCTGCTGCAGCTGCGGCAGCTGCTGCAAGTTTGATGAGGATGATAGCGAGCCGGTGCTGAAGGGCGTGAAGCTGCACTAC
ACCGATATC

Fig. 6.10: Codon optimised SC2M2 nucleotide sequence

6.3.4. Predicted antigenicity and toxicity profile of the SC2M2 immunogen

VaxiJen (http://www.ddg-pharmfac.net/vaxijen/VaxiJen/VaxiJen.html) was used to predict the presence of protective antigens based on the protein sequence. The SC2M2 immunogen was found to have an overall protective antigen prediction of 0.47 and was classified as a probably antigen. Similar to the Wuhan-Hu1 reference sequence which had an antigen prediction of 0.46 and was classified as a probable antigen.

The toxicity was predicted using the Toxinpred (https://webs.iiitd.edu.in/raghava/toxinpred/, (198)). The toxicity profile of the SC2M2 was determined by generating 50 overlapping peptide sequences at a time and predicting if the peptide is toxic or not. Toxin prediction was based on the support vector machine (SVM) threshold value i.e., if the SVM value is negative then the peptide is considered non-toxic and if the SVM value is positive then peptide is

considered toxic. Thirty-five of 1650 combinations of overlapping peptide sequences were identified as toxic. The predicted toxic peptides were detected from amino acid position 1188 to the end of the sequence (position 1273).

Table 6.1: Toxic peptides detected from the optimised S protein immunogen

Peptide Sequence	SVM score	Prediction	Hydrophobicity
EVAKNLNESLIDLQELGKYEQYIKWPWYIWLGFIAGLIAIVMVTIMLCCM	0.05	Toxin	0.09
AKNLNESLIDLQELGKYEQYIKWPWYIWLGFIAGLIAIVMVTIMLCCMTS	0.01	Toxin	0.09
NLNESLIDLQELGKYEQYIKWPWYIWLGFIAGLIAIVMVTIMLCCMTSCC	0.56	Toxin	0.10
LNESLIDLQELGKYEQYIKWPWYIWLGFIAGLIAIVMVTIMLCCMTSCCS	0.62	Toxin	0.11
NESLIDLQELGKYEQYIKWPWYIWLGFIAGLIAIVMVTIMLCCMTSCCSC	0.66	Toxin	0.10
ESLIDLQELGKYEQYIKWPWYIWLGFIAGLIAIVMVTIMLCCMTSCCSCL	0.79	Toxin	0.13
SLIDLQELGKYEQYIKWPWYIWLGFIAGLIAIVMVTIMLCCMTSCCSCLK	0.78	Toxin	0.12
LIDLQELGKYEQYIKWPWYIWLGFIAGLIAIVMVTIMLCCMTSCCSCLKG	0.81	Toxin	0.12
IDLQELGKYEQYIKWPWYIWLGFIAGLIAIVMVTIMLCCMTSCCSCLKGC	0.97	Toxin	0.11
DLQELGKYEQYIKWPWYIWLGFIAGLIAIVMVTIMLCCMTSCCSCLKGCC	1.56	Toxin	0.10
LQELGKYEQYIKWPWYIWLGFIAGLIAIVMVTIMLCCMTSCCSCLKGCCS	1.61	Toxin	0.11
QELGKYEQYIKWPWYIWLGFIAGLIAIVMVTIMLCCMTSCCSCLKGCCSC	1.51	Toxin	0.10
ELGKYEQYIKWPWYIWLGFIAGLIAIVMVTIMLCCMTSCCSCLKGCCSCG	1.51	Toxin	0.12
LGKYEQYIKWPWYIWLGFIAGLIAIVMVTIMLCCMTSCCSCLKGCCSCGS	1.56	Toxin	0.12
GKYEQYIKWPWYIWLGFIAGLIAIVMVTIMLCCMTSCCSCLKGCCSCGSC	1.51	Toxin	0.11
KYEQYIKWPWYIWLGFIAGLIAIVMVTIMLCCMTSCCSCLKGCCSCGSCC	2.04	Toxin	0.11
YEQYIKWPWYIWLGFIAGLIAIVMVTIMLCCMTSCCSCLKGCCSCGSCCK	2.11	Toxin	0.11
EQYIKWPWYIWLGFIAGLIAIVMVTIMLCCMTSCCSCLKGCCSCGSCCKF	2.20	Toxin	0.12
QYIKWPWYIWLGFIAGLIAIVMVTIMLCCMTSCCSCLKGCCSCGSCCKFD	2.22	Toxin	0.12
YIKWPWYIWLGFIAGLIAIVMVTIMLCCMTSCCSCLKGCCSCGSCCKFDE	2.29	Toxin	0.12
IKWPWYIWLGFIAGLIAIVMVTIMLCCMTSCCSCLKGCCSCGSCCKFDED	2.33	Toxin	0.11
KWPWYIWLGFIAGLIAIVMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDD	2.36	Toxin	0.08
WPWYIWLGFIAGLIAIVMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDS	2.25	Toxin	0.10
PWYIWLGFIAGLIAIVMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSE	2.23	Toxin	0.08
WYIWLGFIAGLIAIVMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEP	2.17	Toxin	0.08
YIWLGFIAGLIAIVMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPV	2.08	Toxin	0.08
IWLGFIAGLIAIVMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVL	2.01	Toxin	0.09
WLGFIAGLIAIVMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLK	1.87	Toxin	0.05
LGFIAGLIAIVMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKG	1.85	Toxin	0.05
GFIAGLIAIVMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGV	1.81	Toxin	0.05
FIAGLIAIVMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVK	1.82	Toxin	0.02
IAGLIAIVMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKL	1.80	Toxin	0.02
AGLIAIVMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLH	1.79	Toxin	0.00
GLIAIVMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHY	1.79	Toxin	-0.00
LIAIVMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT	1.72	Toxin	-0.01

N-glycosylation sites were predicted based on the threshold value > 0.5 (Fig. 6.11). The SC2M2 construct has eighteen N-glycosylation sites while the Wuhan Hu-1 S protein had

eleven N-glycosylation sites based on a threshold above 0.5 (Fig.6.11C/D). Compared to the Wuhan Hu-1 S protein, the SC2M2 had six additional glycosylation sites from positions 0 to 400. From positions 400 to 600, the Wuhan Hu-1 S protein had four glycosylation sites not present in the SC2M2. While the SC2M2 had six additional glycosylation sites between positions 1000 to 1200 which was not present in the Wuhan Hu-1 S protein.

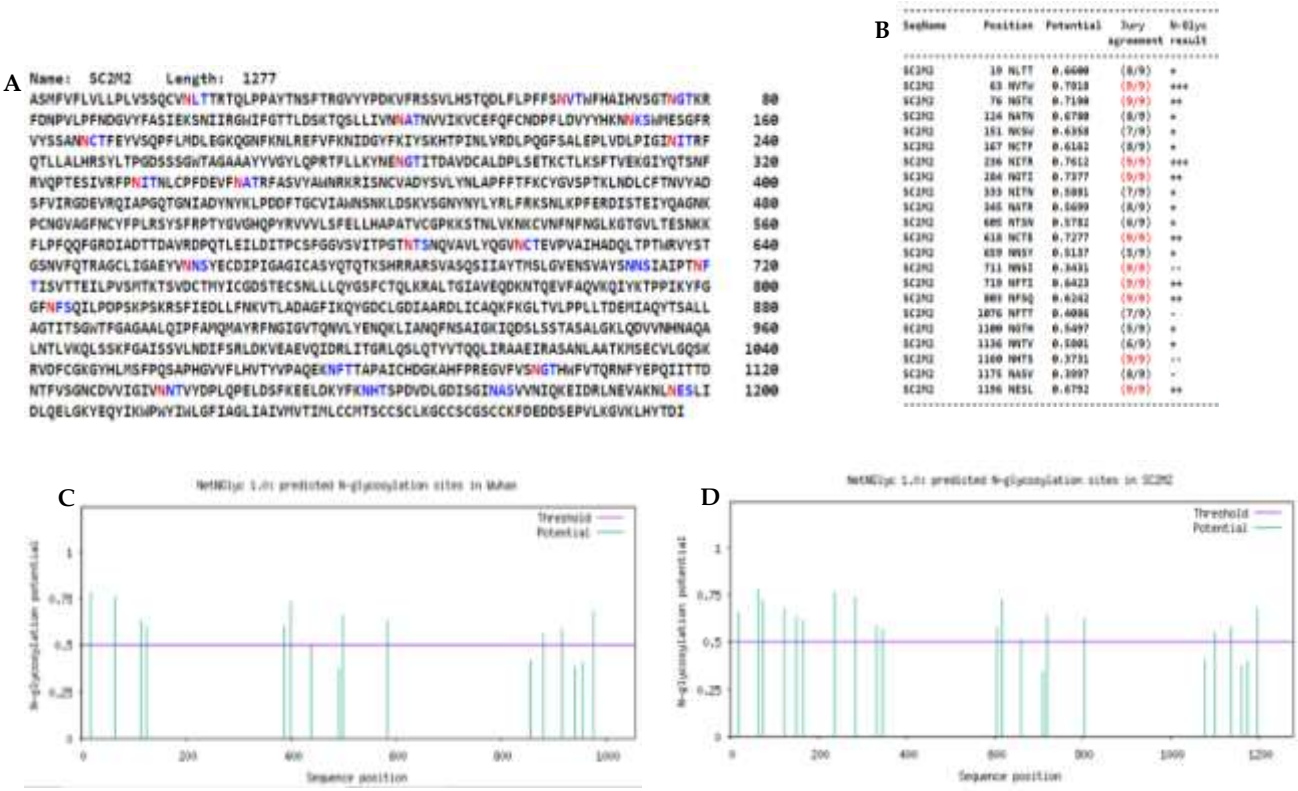

Fig. 6.11: N-linked glycosylation site predications. A) The red text represents the N-glycosylated across the SC2M2 construct. B) Displays the summary of theN-glycosylation predications based on a threshold of >0.5. C) Graphical representation of the N-glycosylation sites present in the Wuhan Hu-1 S protein. D) Graphical representation of the N-glycosylation sites present in the SC2M2 construct.

Sites with a threshold ≥0.5 were predicted as O-glycosylation sites and referred to a positive. SC2M2 has 4 O-glycosylation sites at positions 24, 31, 675 and 688 (Table 6.2). This differed from the Wuhan Hu-1 S protein that has 3 O-glycosylation sites at positions 455 and 460 (Table 6.3).

Table 6.2: O-linked glycosylation predications for SC2M2 S protein construct

#seqname	source	feature	start	end	score	strand	frame	comment		
SC2M2	netOGlyc-4.0.0.13	CARBOHYD	2	2	0.0229985	.	.			
SC2M2	netOGlyc-4.0.0.13	CARBOHYD	14	14	0.214118	.	.			
SC2M2	netOGlyc-4.0.0.13	CARBOHYD	15	15	0.0648594	.	.			
SC2M2	netOGlyc-4.0.0.13	CARBOHYD	21	21	0.445284	.	.			
SC2M2	netOGlyc-4.0.0.13	CARBOHYD	22	22	0.282598	.	.			
SC2M2	netOGlyc-4.0.0.13	CARBOHYD	24	24	0.5	.	.	#POSITIVE		
SC2M2	netOGlyc-4.0.0.13	CARBOHYD	31	31	0.528794	.	.		#POSITIVE	
SC2M2	netOGlyc-4.0.0.13	CARBOHYD	661	661	0.0168086	.	.			
SC2M2	netOGlyc-4.0.0.13	CARBOHYD	675	675	0.607275	.	.		#POSITIVE	
SC2M2	netOGlyc-4.0.0.13	CARBOHYD	678	678	0.184069	.	.			
SC2M2	netOGlyc-4.0.0.13	CARBOHYD	680	680	0.411348	.	.			
SC2M2	netOGlyc-4.0.0.13	CARBOHYD	682	682	0.352187	.	.			
SC2M2	netOGlyc-4.0.0.13	CARBOHYD	688	688	0.679353	.	.		#POSITIVE	
SC2M2	netOGlyc-4.0.0.13	CARBOHYD	691	691	0.456444	.	.			

Table 6.3: O-linked glycosylation predications for Wuhan Hu-1

#seqname	source	feature	start	end	score	strand	frame	comment		
WUHAN	netOGlyc-4.0.0.13	CARBOHYD	455	455	0.520173	.	.		#POSITIVE	
WUHAN	netOGlyc-4.0.0.13	CARBOHYD	458	458	0.130477	.	.			
WUHAN	netOGlyc-4.0.0.13	CARBOHYD	460	460	0.566343	.	.		#POSITIVE	

6.3.5. Synthesis of the mosaic S construct

The SC2M2 codon optimised nucleotide sequence (Fig 6.13) was cloned in a pCDNA3.1(-) myc-HisA plasmid vector by Biomatik (Fig 6.12). Together with the plasmid, the construct was 9298 bp in length with cloning sites NheI (GCTAGC) and EcoRV (GATATC). All quality checks performed on the construct to ensure correctness and purity, passed as shown in Table 6.4. Additional QC was performed by Biomatik to ensure the correct base calling was achieved (Fig 6.13A). High fluorescent intensity was observed for each nucleotide base which was an indication of good quality bases (Fig 6.13B). The construct was reconstituted in 50µl TBE buffer and stored at -80°C until required.

Table 6.4: Quality control checks performed on the SC2M2 construct by Biomatik

Test Items	Specifications	Results
Sequencing Alignment	Sequencing data consistent with target	[X] Pass
Vector Sequence	Flanking sequence of cloning sites are correct	[X] Pass
Restriction Digest	Insert size is correct and no contaminated bands	[X] Pass
ORF Across Junction	Correct and consistent with target	N/A
PCR Amplification	Correct and no contaminated bands	[X] Pass
Endotoxin Level	Verified, <0.1 EU/µg (Endo-Free Preps Only)	N/A
Appearance	Clear, no foreign particles	[X] Pass
DNA Purity	Purity (A 260/A280 = 1.8 - 2.0)	[X] Pass
DNA Quantity	Actual yield (by A 260)	5µg/5µg

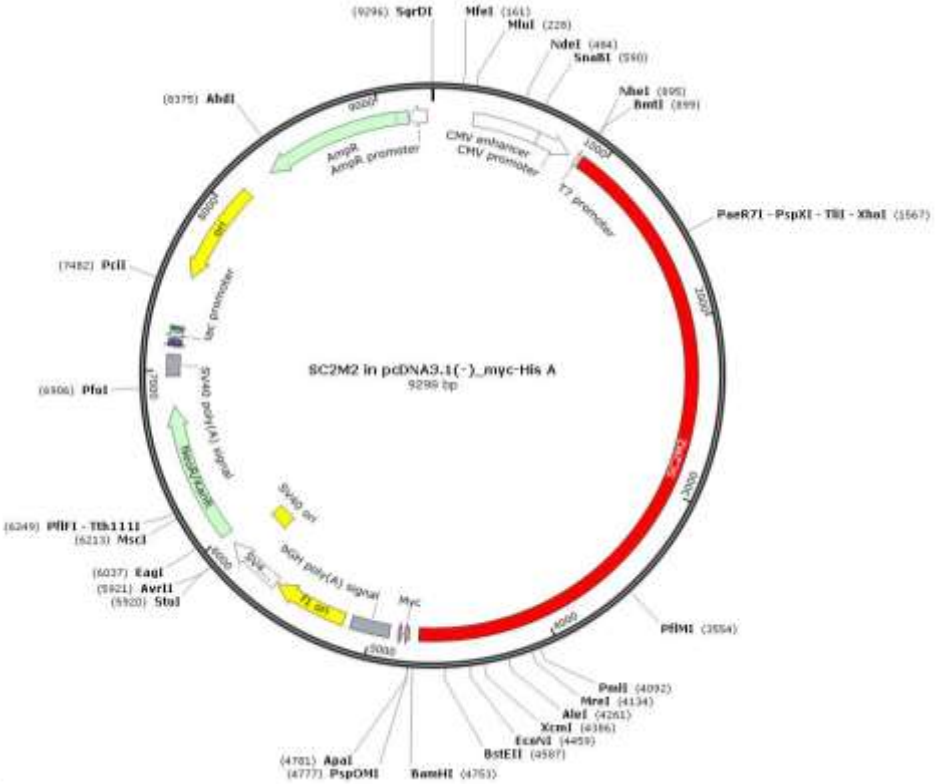

Fig. 6.12: Map of the selected optimal spike protein construct (SC2M2) cloned in pcDNA3.1(-) myc-His A. Diagram created by snapgene.

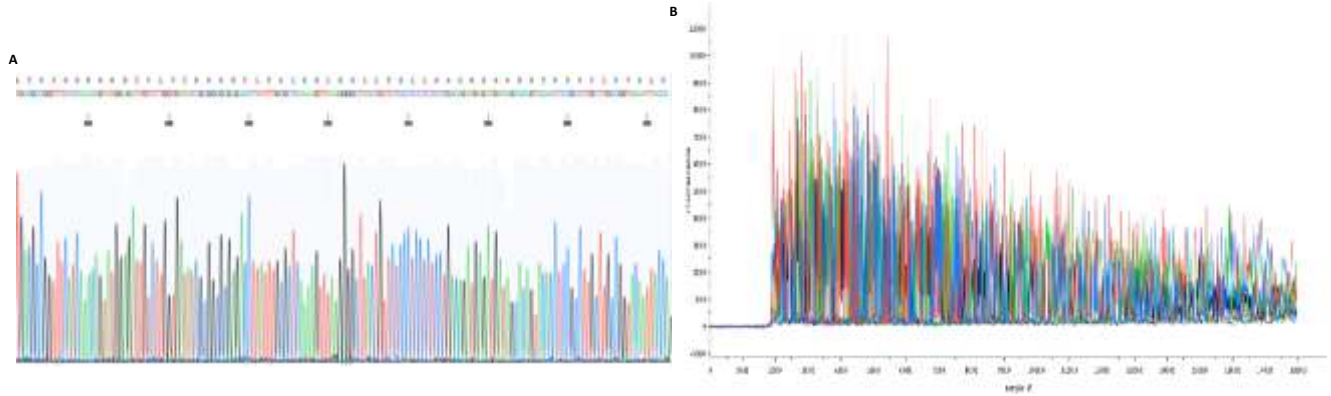

Fig 6.13: SC2M2 base calls (A) and fluorescent intensity (B) of each nucleotide base checked by Biomatik.

6.3.6. Detection of expressed S protein construct

Post transfection, the S protein supernatants were used to detect the presence of the S protein in the SC2M2 construct. The S protein was run through an SDS-PAGE to determine if the SC2M2 protein was successfully separated and detected. The SARS-CoV-2 S protein SC2M2 construct was approximately 110kDa in size which is about the size of the S1 and S2 fusion

proteins (Fig. 6.1). The presence of the complete S polyprotein (~180kDa) was not evident from the gel.

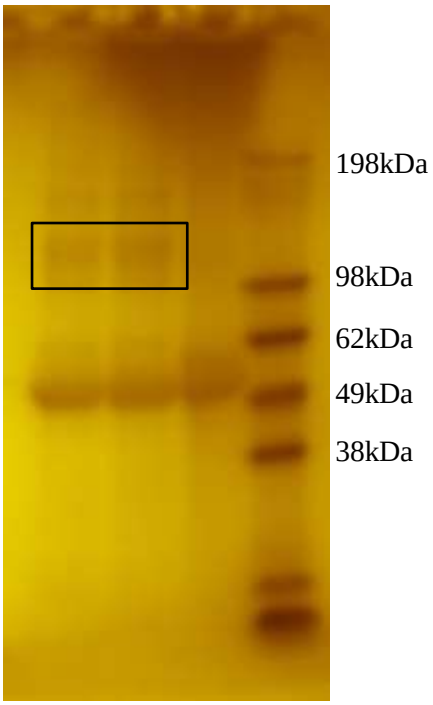

Fig. 6.14: SDS-PAGE showing the protein separation of the SC2M2 construct. The seeblue plus 2 ladder was used for the marker on the right of the gel. The black rectangle indicates the spike protein band present in SC2M2 supernatant (well 1 and 2).

6.4. Discussion

In South Africa, as of October 2023 the virus has led to >4 million cases and 102 595 deaths, which is a cause for concern (https://covid19.who.int/region/afro/country/za, accessed 18 October 2023). The increase in cases and deaths are a result of the emergence of SARS-CoV-2 novel lineages and variants which resulted in the rapid generation of several mutations, especially among the Omicron sub-lineages. The S protein is necessary for SARS-CoV-2 entry into host cells and is therefore the main region targeted by vaccines (199). Antibodies that act against the S protein prevent viral entry which inhibits viral replication (199). The S protein is the perfect target for vaccine design since it is highly antigenic and has the ability to elicit the hosts immune response. Several vaccines that target the S protein are currently being used however with the increasing number of mutations, challenges have been encountered to establish an appropriate vaccine that provides effective protection against most, if not all, lineages and VOCs.

The entry of a mosaic antigen through a viral vector promotes intracellular expression and antigen processing, resulting in more efficient activation of CD8+ T cells (200). In doing so, mosaics have the ability to improve humoral and cellular immune responses against targeted infectious agents, as described in previous studies (183,200,201). A previous study used immunoinformatics tools to design vaccines based on immunodominant SARS-CoV-2 S protein epitopes and primary immune escape mutations, similar to our study (202). The mosaic antigens targets epitopes that confer greater breadth of known viral sequences and therefore induce broad T cell epitope coverage from immunisation with a mosaic (200). Therefore the use of mosaics instead of natural sequences can improve T cell responses (201). We generated an optimal S protein mosaic based on South African sequence data. The construct designed in this study has excellent predicted CTL epitope coverage and comprises protective antigens. Theoretically, the protective antigens are targeted by the hosts immune system to induce a response against SARS-CoV-2. Although there was evidence of toxicity across peptides towards the latter region of the S protein, the use of shorter peptides (<50) could resolve some of the predicted toxic properties. We did however identify non-toxic peptides within the NTD and RBD regions, located in the first 550 amino acid positions. The construct has six predicted N-linked glycosylation sites present within the region of the NTD and RBD. The O-linked glycosylation sites were present within the NTD sequence of the SC2M2 construct based on prediction tools used for analysis. Glycosylation predictions with the NTD and RBD regions could potentially enhance the binding affinity of the virus to the hosts ACE-2 receptor and improves viral infectivity (203,204). The presence of glycosylation sites within the immunogenic epitopes of the S protein can result in concealing the epitopes from the host and in turn escaping the hosts immune responses (204). It is therefore important that vaccine design also targets glycosylation sites within the S protein.

Previous studies have describe the NTD and RBD as being highly diverse and the primary regions that harbour antibody escape mutations (50,199,205). These regions are therefore important targets for vaccine design. Majority of the primary mutations located in the RBD and NTD are associated with antibody neutralising activity. Previous studies described mutations in the NTD in VOC prevent monoclonal antibody attachment (205,206). Such mutations include deletions at position 69/70 and 144 shared in both Omicron and Alpha, and D215G, del242-244 and R246I in Beta. Amino acid residues 159 to 170 and 230 to 233 are described as conserved, indicating that regions may be associated with RBD/NTD interactions (205).

When designing therapeutics, it may be important to consider including neutralising antibodies that interact with NTD residues that are conserved across various VOCs and lineages.

This study had a few limitations to take into consideration for future studies. First the mosaic design did not include variants from Omicron BA.5 onwards. However, our data did include recombinants of Omicron BA.1 and BA.2 sub-lineages which had several mutations associated with immune escape and therefore still serve as an appropriate construct. Second, the complete S polyprotein was not observed but the S1 and S2 regions of the protein which may be due to the short incubation period during transfection. This could be overcome by performing transfection in flasks rather than 6-well plates and incubating for up to 6 days. Third, we did not assess the ability of the construct to elicit an immune response and determine its effect on neutralising antibodies. A previous study reported that antibody escape provides the means to the evolution of novel SARS-CoV-2 variants (205).

6.5. Conclusion

The strategy used in this study to design a novel S protein construct (SC2M2) and the predictions made based on the sequence data from our study from 2020 to 2022 show that our approach was reliable. This encompasses VOCs from Beta through to the Omicron BA.4 waves. Our construct accounted for increased diversity by maximizing inclusion of targeted CTL epitopes present across the dataset. The construct designed has excellent CTL epitope coverage, protective antigens, and non-toxic peptides in the NTD and RBD which are the regions where the most significant mutations are detected. Additional N- and O-linked glycosylation sites were observed in the SC2M2 construct compared to the Wuhan Hu-1 S protein, specifically within the NTD and RBD regions. If these glycosylation sites are targeted by an antibody response elicited by the S immunogen this could potentially improve the efficacy of the construct. Our immunogen may prove to serve as a reliable candidate vaccine, however further studies are required to determine if an immune response is elicited by SC2M2. The latter could be achieved by inoculating mice with varying concentrations of the immunogen and testing the mice sera neutralising activity.

CHAPTER 7
GENERAL DISCUSSION AND CONCLUSIONS

SARS-CoV-2 emerged in 2020 in South Africa and continues to circulate with the emergence of novel SARS-CoV-2 VOCs and lineages. From 2020 to 2022, we described and characterised the viruses circulating in the five COVID-19 waves during which new mutations in the S protein has enhanced its ability to escape neutralising antibodies induced by the dominant vaccine immune epitopes in the S protein and increase transmissibility and infectivity. Although sequencing is the gold standard for detection of VOCs, we showed that SNP genotyping by PCR proved reliable to differentiate between Alpha, Beta, Delta and Omicron BA.1 in low- to middle- income countries were sequencing resources may be scarce or unavailable. However, for the detection of additional Omicron sub-lineages, additional SNP assays will be required that target unique mutations within the S protein of each sub-lineage.

Furthermore, we showed that the number of mutations increased as novel lineages and VOCs emerged, which resulted in reduced protection against vaccine-induced or natural immune response and other environmental selective pressures. We also identified several low frequency lineages of which the B.1, B.1.1, and C.1 lineages retained unique mutations and mutations similar to VOCs during more than one wave in South Africa. This could result in convergence and recombination between lineages and VOCs, which may result in the next novel lineage or variant that may further increase transmissibility, infectivity and escape vaccine-induced or natural host immunity. It is therefore important to analyse low-frequency lineages together with VOCs across the globe to determine their impact on different populations.

In addition, we described the significance of heterogeneity identified within the SARS-CoV-2 S protein at an intra-host level. Our study identified mutations in S protein at positions 19, 371, and 484 which were previously reported to reduce virus neutralisation by antibodies. Our findings showed that heterogeneous mutations may be due to intra-host adaptation and could encourage the evolution of the S protein further contributing to immune evasion. This may contribute to the ongoing emergence of new variants associated with continued outbreaks globally. However, we could not determine if heterogeneity in the S protein has a positive or negative impact on individuals infected with SARS-CoV-2.

Finally, we designed a novel S expression construct (SC2M2) based on SARS-CoV-2 strains characterised in this study from 2020 to 2022. This encompasses VOCs from Beta through to the Omicron BA.4 waves. Our construct accounted for increased diversity by maximising inclusion of targeted CTL epitopes present across the dataset, has excellent CTL epitope coverage, protective antigens, and non-toxic peptides in the S protein. This novel S construct may prove to serve as a reliable candidate for use and improvement SARS-CoV-2 vaccines. However, further studies to determine if immune responses are elicited *in vivo* by novel S proteins expressed by SC2M2 are pending.

To conclude, our study findings have shown that (i) rapid detection of emerging VOCs was possible with the use of SNP genotyping assays, (ii) low frequency lineages that harbor mutations from VOCs circulated over time together with VOCs, (iii) heterogeneity within the S protein identified in our study was previously reported to encourage escape from neutralising antibodies and the evolution of SARS-CoV-2, and lastly (iv) the novel S construct based on initial and more recent circulating VOCs and lineages described in this study could be evaluated further for use in development of improved SARS-CoV-2 vaccines.

REFERENCES

1. Zardecki C, Dutta S, Goodsell DS, Lowe R, Voigt M, Burley SK. PDB-101 : Educational resources supporting molecular explorations through biology and medicine. Protein Sci. 2021;31:129–40.
2. Artika IM, Dewantari AK, Wiyatno A. Molecular biology of coronaviruses: current knowledge. Heliyon [Internet]. 2020;6:1–22. Available from: https://doi.org/10.1016/j.heliyon.2020.e04743
3. Yin Y, Wunderink RG. MERS, SARS and other coronaviruses as causes of pneumonia. Respirology. 2018;23(2):130–7.
4. Fehr AR, Perlman S. Coronaviruses Methods and Protocols. Maier HJ, Bickerton E, Britton P, editors. Springer Science; 2015. 1–282 p.
5. Chen J. Pathogenicity and transmissibility of 2019-nCoV—A quick overview and comparison with other emerging viruses. Microbes Infect [Internet]. 2020;22:69–71. Available from: https://doi.org/10.1016/j.micinf.2020.01.004
6. Yan Y, Shin WI, Pang YX, Meng Y, Lai J, You C, et al. The first 75 days of novel coronavirus (SARS-CoV-2) outbreak: Recent advances, prevention, and treatment. Int J Environ Res Public Health. 2020;17(2323):1–23.
7. Mackay IM, Arden KE. MERS coronavirus: Diagnostics, epidemiology and transmission. Virol J [Internet]. 2015;12(222):1–21. Available from: http://dx.doi.org/10.1186/s12985-015-0439-5
8. Rokni M, Ghasemi V, Tavakoli Z. Immune responses and pathogenesis of SARS-CoV-2 during an outbreak in Iran: Comparison with SARS and MERS. Rev Med Virol. 2020;30:1–6.
9. World Health Organization. Middle East respiratory syndrome coronavirus (MERS-CoV). 2022. Middle East respiratory syndrome coronavirus. Available from: https://www.who.int/health-topics/middle-east-respiratory-syndrome-coronavirus-mers#tab=tab_1 (accessed 18 March 2023)
10. Al-Tawfiq JA, Memish ZA. Lack of seasonal variation of Middle East Respiratory Syndrome Coronavirus (MERS-CoV). Travel Med Infect Dis. 2019;27:125–6.
11. WHO. WHO coronavirus disease (COVID-19) dashboard. Geneva: World Health Organization [Internet]. 2023. Available from: https://covid19.who.int/ (accessed: 20 March 2023)

12. Fadaka AO, Sibuyi NRS, Adewale OB, Bakare OO, Akanbi MO, Klein A, et al. Understanding the epidemiology, pathophysiology, diagnosis and management of SARS-CoV-2. J Int Med Res. 2020;48(8):1–23.
13. O'Driscoll M, Ribeiro Dos Santos G, Wang L, Cummings DAT, Azman AS, Paireau J, et al. Age-specific mortality and immunity patterns of SARS-CoV-2. Nature [Internet]. 2021;590:140–5. Available from: http://dx.doi.org/10.1038/s41586-020-2918-0
14. Peckham H, de Gruijter NM, Raine C, Radziszewska A, Ciurtin C, Wedderburn LR, et al. Male sex identified by global COVID-19 meta-analysis as a risk factor for death and ITU admission. Nat Commun [Internet]. 2020;11:1–10. Available from: http://dx.doi.org/10.1038/s41467-020-19741-6
15. World Health Organization. Confirmed and probable COVID-19 cases and deaths by age [Internet]. 2023. Available from: https://app.powerbi.com/view?r=eyJrIjoiYWRiZWVkNWUtNmM0Ni00MDAwLTljYWMtN2Ew NTM3YjQzYmRmIiwidCI6ImY2MTBjMGI3LWJkMjQtNGIzOS04MTBiLTNkYzI4MGFmYjU5MCIsI mMiOjh9 (accessed 21 March 2023)
16. Chilamakuri R, Agarwal S. COVID-19: Characteristics and Therapeutics. Cells. 2021;10:1–29.
17. Lofti M, Hamblin MR, Rezaei N. COVID-19: Transmission, prevention, and potential therapeutic opportunities. Clin Chim Acta [Internet]. 2020;508:254–66. Available from: www.elsevier.com/locate/cca Review
18. Acter T, Uddin N, Das J, Akhter A. Evolution of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) as coronavirus disease 2019 (COVID-19) pandemic: A global health emergency. Sci Total Environ J. 2020;730:2–15.
19. World Health Organisation [Internet]. 2021. Available from: https://www.who.int/news/item/26-11-2021-classification-of-omicron-(b.1.1.529)-sars-cov-2-variant-of-concern
20. Mehmood I, Ijaz M, Ahmad S, Ahmed T, Bari A, Abro A, et al. SARS-CoV-2: An Update on Genomics, Risk Assessment, Potential Therapeutics and Vaccine Development. Int J Environ Res Public Health. 2021;18:1–23.
21. Ng WH, Tipih T, Makoah NA, Vermeulen JG, Goedhals D, Sempa JB, et al. Comorbidities in SARS-CoV-2 patients: A systematic review and meta-analysis. Am Soc Microbiol. 2021;12(1):1–12.

22. Kanwugu ON, Adadi P. HIV/SARS-CoV-2 coinfection: A global perspective. J Med Virol. 2021;93(2):726–32.
23. Liu C, Zhao Y, Okwan-Duodu D, Basho R, Cui X. COVID-19 in cancer patients: risk, clinical features, and management. Cancer Biol Med. 2020;17(3):519–27.
24. Caillard S, Chavarot N, Bertrand D, Kamar N, Thaunat O, Moal V, et al. Occurrence of severe COVID-19 in vaccinated transplant patients. Kidney Int. 2021;100:477–9.
25. Choi B, Choudhary M, Sparks JR, Padera RF. Persistance and evolution of SARS-Cov-2 in an immunocompromised host. New Engl J Med J Med. 2020;1–4.
26. Camprubí D, Gaya A, Angeles M, Martí-soler H, Soriano A, Mosquera M, et al. Persistent replication of SARS-CoV-2 in a severely immunocompromised patient treated with several courses of remdesivir. Int J Infect Dis. 2021;104:379–81.
27. Kemp SA, Collier DA, Datir RP, Ferreira IATM, Gayed S, Jahun A, et al. SARS-CoV-2 evolution during treatment of chronic infection. Nature. 2021;592:277–82.
28. Zhang D, Lou X, Yan H, Pan J, Mao H, Tang H, et al. Metagenomic analysis of viral nucleic acid extraction methods in respiratory clinical samples. BMC Genomics. 2018;19:773.
29. Prevention C for DC and. Post-COVID Conditions. Centers Dis Control Prev. 2021;1–2.
30. Morris SB, Schwartz NG, Patel P, Abbo L, Beauchamps L, Balan S, et al. Case Series of Multisystem Inflammatory Syndrome in Adults Associated with SARS-CoV-2 Infection — United Kingdom and United States, March–August 2020. MMWR Morb Mortal Wkly Rep. 2020;69:1450–6.
31. Liu Y, Liu J, Xia H, Zhang X, Fontes-Garfias CR. Neutralizing Activity of BNT162b2-Elicited Serum. N Engl J Med. 2021;384:1–3.
32. Sadoff J, Gray G, Vandebosch A, Cárdenas V, Shukarev G, Grinsztejn B, et al. Safety and Efficacy of Single-Dose Ad26.COV2.S Vaccine against Covid-19. N Engl J Med. 2021;384:2187–201.
33. SAMRC. Women Scientists lead the South African leg of the Phase 3 Janssen Efficacy COVID-19 vaccine trial [Internet]. 2020. Available from: www.samrc.ac.za/media-release/women-scientists-lead-south- african-leg-phase-3-janssen-efficacy-covid- 19-vaccine
34. Mengstu S, Berha AB. Safety and Efficacy of COVID-19 Vaccine in Africa : Systematic Review. Infect Drug Resist. 2023;16:3085–100.

35. Mabuka T, Naidoo N, Ncube N, Yiga T, Ross M, Kurehwa K, et al. The Impact of SARS-CoV-2 Lineages (Variants) and COVID-19 Vaccination on the COVID-19 Epidemic in South Africa : Regression Study. J Med Internest Res [Internet]. 2023;4:1–26. Available from: https://med.jmirx.org/2023/1/e34598
36. Kabir MT, Uddin MS, Hossain MF, Abdulhakim JA, Alam MA, Ashraf GM, et al. nCOVID-19 Pandemic: From Molecular Pathogenesis to Potential Investigational Therapeutics. Front Cell Dev Biol. 2020;8:1–23.
37. Ou X, Liu Y, Lei X, Li P, Mi D, Ren L, et al. Characterization of spike glycoprotein of SARS-CoV-2 on virus entry and its immune cross-reactivity with SARS-CoV. Nat Commun [Internet]. 2020;11:1–12. Available from: http://dx.doi.org/10.1038/s41467-020-15562-9
38. Rabaan AA, Al-Ahmed SH, Haque S, Sah R, Tiwari R, Malik YS, et al. SARS-CoV-2, SARS-CoV, and MERS-CoV: A comparative overview. Infez Med. 2020;28:174–84.
39. Chan YA, Zhan SH. The Emergence of the Spike Furin Cleavage Site in SARS-CoV-2. Mol Biol Evol. 2021;39:1–7.
40. Carabelli AM, Peacock TP, Thorne LG, Harvey WT, Hughes J, de Silva TI, et al. SARS-CoV-2 variant biology: immune escape, transmission and fitness. Nat Rev Microbiol [Internet]. 2023; Available from: https://doi.org/10.1038/s41579-022-00841-7
41. Lokman SM, Rasheduzzaman M, Salauddina A, Baruaa R, Tanzinaa AY, Rumia MH, et al. Exploring the genomic and proteomic variations of SARS-CoV-2 spike glycoprotein: A computational biology approach. Infect Genet Evol. 2020;84(January):1–10.
42. Aleem A, Bari A, Samad A, Slenker AK. Emerging Variants of SARS-CoV-2 And Novel Therapeutics Against Coronavirus (COVID-19). In: In: StatPearls [Internet] Treasure Island (FL): StatPearls Publishing. 2021.
43. Tegally H, Wilkinson E, Giovanetti M, Iranzadeh A, Fonseca V, Jennifer G, et al. Detection of a SARS-CoV-2 variant of concern in South Africa. Nature [Internet]. 2021;592:438–43. Available from: http://dx.doi.org/10.1038/s41586-021-03402-9
44. Afrin SZ, Paul SK, Begum JA, Nasreen SA, Ahmed S, Ahmad FU, et al. Extensive genetic diversity with novel mutations in spike glycoprotein of severe acute respiratory syndrome coronavirus 2 , Bangladesh in late 2020. New Microbes New Infect [Internet]. 2021;41:1–5. Available from: https://doi.org/10.1016/j.nmni.2021.100889

45. Banoun H. Evolution of SARS-CoV-2 : Review of Mutations , Role of the Host Immune System. 2021;145:392–403.
46. Volz E, Hill V, Mccrone JT, Thomson EC, Rambaut A, Connor TR. Evaluating the Effects of SARS-CoV-2 Spike Mutation D614G on Transmissibility and Pathogenicity. Cell. 2021;184:64–75.
47. Tegally H, Wilkinson E, Lessells RJ, Giandhari J, Pillay S, Msomi N, et al. Sixteen novel lineages of SARS-CoV-2 in South Africa. Nat Med [Internet]. 2021;27(March). Available from: http://dx.doi.org/10.1038/s41591-021-01255-3
48. CDC. Centres for Disease Control and Prevention [Internet]. 2021. Available from: https://www.cdc.gov/coronavirus/2019-ncov/variants/variant-classifications.html?CDC_AA_refVal=https%3A%2F%2Fwww.cdc.gov%2Fcoronavirus%2F201 9-ncov%2Fvariants%2Fvariant-info.html#anchor_1632154493691
49. Peacock TP, Penrice-Randal R, Hiscox JA, Barclay WS. SARS-CoV- 2 one year on: Evidence for ongoing viral adaptation. J Gen Virol. 2021;102:1–17.
50. Harvey WT, Carabelli AM, Jackson B, Gupta RK, Thomson EC, Harrison EM, et al. SARS-CoV-2 variants, spike mutations and immune escape. Nat Rev Microbiol [Internet]. 2021;19:409–24. Available from: http://dx.doi.org/10.1038/s41579-021-00573-0
51. Tzou PL, Tao K, Kosakovsky Pond SL, Shafer RW. Coronavirus Resistance Database (CoV-RDB): SARS-CoV-2 susceptibility to monoclonal antibodies, convalescent plasma, and plasma from vaccinated persons. PLoS One [Internet]. 2022;17:1–20. Available from: http://dx.doi.org/10.1371/journal.pone.0261045
52. Wang P, Nair MS, Liu L, Iketani S, Luo Y, Guo Y, et al. Antibody resistance of SARS-CoV-2 variants B.1.351 and B.1.1.7. Nature [Internet]. 2021;593:130–5. Available from: http://dx.doi.org/10.1038/s41586-021-03398-2
53. Fontanet A, Autran B, Lina B, Kieny MP, Karim SSA, Sridhar D. SARS-CoV-2 variants and ending the COVID-19 pandemic. Lancet. 2021;397:952–4.
54. Dhawan M, Sharma A, Priyanka, Thakur N, Rajkhowa TK, Choudhary OP. Delta variant (B.1.617.2) of SARS-CoV-2: Mutations, impact, challenges and possible solutions. Hum Vaccines Immunother [Internet]. 2022;18(5):1–12. Available from: https://doi.org/10.1080/21645515.2022.2068883
55. Gangavarapu K, Latif AA, Mullen JL, Alkuzweny M, Hufbauer E, Tsueng G, et al. Outbreak.info genomic reports: scalable and dynamic surveillance of SARS-CoV-2

variants 2 and mutations. medRxiv Prepr. 2022;1–20.

56. Gangavarapu K, Alaa Abdel Latif, Julia Mullen, Alkuzweny M, Hufbauer E, Tsueng G, et al. outbreak.info. 2023 [cited 2023 Feb 13]. Lineage comparison. Available from: https://outbreak.info/compare-lineages?pango

57. Willett BJ, Grove J, Maclean OA, Wilkie C, Lorenzo G De, Furnon W, et al. SARS-CoV-2 Omicron is an immune escape variant with an altered cell entry pathway. Nat Microbiol. 2022;7:1161–79.

58. Chatterjee S, Bhattacharya M, Nag S, Chakraborty C. A Detailed Overview of SARS-CoV-2 Omicron: Its Sub-Variants, Mutations and Pathophysiology, Clinical Characteristics, Immunological Landscape, Immune Escape, and Therapies. Viruses. 2023;15:1–27.

59. Makoni M. South Africa responds to new SARS-CoV-2 variant. Lancet (London, England) [Internet]. 2021;397(10271):267. Available from: http://dx.doi.org/10.1016/S0140-6736(21)00144-6

60. Lopez Bernal J, Andrews N, Gower C, Gallagher E, Simmons R, Thelwall S, et al. Effectiveness of Covid-19 Vaccines against the B.1.617.2 (Delta) Variant. N Engl J Med. 2021;1–10.

61. Callaway E. Delta Coronavirus Variant : Scientists brace for impact. Nature. 2021;595:17–8.

62. Gray G, Collie S, Goga A, Garrett N, Champion J, Seocharan I. Effectiveness of Ad26.COV2.S and BNT162b2 Vaccines against Omicron Variant in South Africa. N Engl J Med. 2022;1–3.

63. NOVAVAX. GlobalNewswire. 2020. p. 2020–2 Novavax Initiates Efficacy Trial of COVID-19 Vaccine in South Africa. Available from: https://ir.novavax.com/news-releases/news-release-details/novavax-initiates-efficacy-trial-covid-19-vaccine-south-africa

64. Shinde V, Bhikha S, Hoosain Z, Archary M, Bhorat Q, Fairlie L, et al. Efficacy of NVX-CoV2373 Covid-19 Vaccine against the B.1.351 Variant. N Engl J Med. 2021;384:1899–909.

65. Folegatti PM, Ewer KJ, Aley PK, Angus B, Becker S, Belij-Rammerstorfer S, et al. Safety and immunogenicity of the ChAdOx1 nCoV-19 vaccine against SARS-CoV-2: a preliminary report of a phase 1/2, single-blind, randomised controlled trial. Lancet. 2020;396:467–78.

66. Madhi SA, Baillie V, Cutland CL, Voysey M, Koen AL, Fairlie L, et al. Efficacy of the ChAdOx1 nCoV-19 Covid-19 Vaccine against the B.1.351 Variant. N Engl J Med. 2021;384:1885–98.
67. Noh JY, Jeong HW, Shin EC. SARS-CoV-2 mutations, vaccines, and immunity: implication of variants of concern. Signal Transduct Target Ther [Internet]. 2021;6(1):3–4. Available from: http://dx.doi.org/10.1038/s41392-021-00623-2
68. Chen J, Gao K, Wang R, Wei GW. Prediction and mitigation of mutation threats to COVID-19 vaccines and antibody therapies. Chem Sci. 2021;12(20):6929–48.
69. Grifoni A, Weiskopf D, Ramirez SI, Mateus J, Dan JM, Moderbacher CR, et al. Targets of T Cell Responses to SARS-CoV-2 Coronavirus in Humans with COVID-19 Disease and Unexposed Individuals. Cell. 2020;181:1489–501.
70. Li J, Wang J, Kang AS, Sacitharan PK. Mapping the T cell response to COVID-19. Nat Signal Transduct Target Ther. 2020;5:111–2.
71. Tarke A, Sidney J, Methot N, Yu ED, Zhang Y, Dan JM, et al. Impact of SARS-CoV-2 variants on the total CD4+ and CD8+ T cell reactivity in infected or vaccinated individuals. Cell Reports Med. 2021;2:1–12.
72. Khare S, Gurry C, Freitas L, Schultz MB, Bach G, Diallo A, et al. GISAID's Role in Pandemic Response. China CDC Wkly. 2021;3:1049–51.
73. Quick J. protocols.io. 2020. nCoV-2019 sequencing protocol v3 (LoCost). protocols.io. Available from: https://www.protocols.io/view/ncov-2019-sequencing-protocol-v3-locost-bh42j8ye Josh
74. Reiling SJ, Loranger K, Roy AM, Chen SH, Ragoussis I. McGill Genome Centre. 2022. nCoV-2019 McGill Artic PCR Protocol, V4.1 at 63C V.2. Available from: https://www.protocols.io/view/ncov-2019-mcgill-artic-pcr-protocol-v4-1-at-63c-ewov18e4ygr2/v2
75. Viana R, Moyo S, Amoako DG, Tegally H, Scheepers C, Althaus CL, et al. Rapid epidemic expansion of the SARS-CoV-2 Omicron variant in southern Africa. Nature [Internet]. 2022;1–29. Available from: http://www.ncbi.nlm.nih.gov/pubmed/35042229
76. Bartolo AJ, Holman O, Wilbraham S, Grosch S, Ottenwälder B, Laubscher D. Oxford Nanopore PCR Tiling of SARS-CoV-2 Virus with Rapid Barcoding and Midnight RT PCR Expansion. Genomics. 2022;1–6.
77. Freed N, Silander O. Vol. 3, protocols.io. 2020. p. 1–14 nCoV-2019 sequencing

protocol (RAPID barcoding, 1200bp amlicon) V.3.

78. Cleemput S, Dumon W, Fonseca V, Abdool Karim W, Giovanetti M, Carlos Alcantara L, et al. Genome Detective Coronavirus Typing Tool for rapid identification and characterization of novel coronavirus genomes. Bioinformatics [Internet]. 2020;11:3552–5. Available from: https://www.genomedetective.com/app/typingtool/virus/
79. Hyrax BioSciences [Internet]. Exatype : automated SARS-CoV-2 variant typing directly from NGS sequence data Exatype. Available from: https://ngs.exatype.com/dashboard?product=SARSCoV2
80. Aksamentov I, Roemer C, Hodcroft EB, Neher RA. Nextclade: clade assignment, mutation calling and quality control for viral genomes. J Open Source Softw [Internet]. 2021;6:3773. Available from: https://doi.org/10.21105/joss.03773
81. Fischer W, Perkins S, Theiler J, Bhattacharya T, Yusim K, Funkhouser R, et al. Polyvalent vaccines for optimal coverage of potential T-cell epitopes in global HIV-1 variants. Nat Med. 2007;13(1):100–6.
82. Gupta R, Brunak S. Prediction of glycosylation across the human proteome and the correlation to protein function. Pacific Symp Biocomput. 2002;7:310–22.
83. Steentoft C, Vakhrushev SY, Joshi HJ, Kong Y, B VCM, Schjoldager KTB, et al. Precision mapping of the human O-GalNAc glycoproteome through SimpleCell technology. EMBO J. 2013;32:1478–88.
84. Daniloski Z, Jordan TX, Ilmain JK, Guo X, Bhabha G, TenOever BR, et al. The spike D614G mutation increases SARS-CoV-2 infection of multiple human cell types. Elife. 2020;10:1–16.
85. Tegally H, Wilkinson E, Althaus CL, Giovanetti M, San JE, Giandhari J, et al. Rapid replacement of the Beta variant by the Delta variant in South Africa. medRxiv [Internet]. 2021;1–26. Available from: https://www.medrxiv.org/content/10.1101/2021.09.23.21264018v1%0Ahttps://www.medrxiv.org/content/10.1101/2021.09.23.21264018v1.abstract%0Ahttps://www.medrxiv.org/content/10.1101/2021.09.23.21264018v1%0Ahttps://www.medrxiv.org/content/10.1101/2021.09.23.212
86. Wilkinson E, Giovanetti M, Tegally H, San J, Lessels R, Cuadros D, et al. A year of genomic surveillance reveals how the SARS-CoV-2 pandemic unfolded in Africa. Science (80-). 2021;431:423–31.

87. Scheepers C, Everatt J, Amoako DG, Tegally H, Wibmer CK, Mnguni A, et al. Emergence and phenotypic characterization of C.1.2, a globally detected lineage that rapidly accumulated mutations of concern. Nat Commun. 2022;13:1–9.
88. Karim SSA, Karim QA. Omicron SARS-CoV-2 variant: a new chapter in the COVID-19 pandemic. Lancet [Internet]. 2021;(21):1–3. Available from: http://www.ncbi.nlm.nih.gov/pubmed/34871545
89. NGS-SA. Network for Genomic Surveillance in South Africa (NGS-SA). SARS-CoV-2 Seq Updat 3 December 2021 [Internet]. 2021;1–29. Available from: https://www.nicd.ac.za/diseases-a-z-index/disease-index-covid-19/sars-cov-2-genomic-surveillance-update/
90. Gao SJ, Guo H, Luo G. Omicron variant (B.1.1.529) of SARS-CoV-2, a global urgent public health alert! J Med Virol [Internet]. 2021;1–2. Available from: http://www.ncbi.nlm.nih.gov/pubmed/34850421
91. Pulliam J, Schalkwyk C van, Govender N, Gottberg A von, Cohen C, Groome MJ, et al. Increased risk of SARS-CoV-2 reinfection associated with emergence of the Omicron variant in South Africa. medRxiv. 2021;1–43.
92. NGS-SA. NGS-SA. Network for Genomic Surveillance in South Africa (NGS-SA). SARS-CoV-2 Seq Update 14 January 2022. Glob Challenges [Internet]. 2022;1–32. Available from: https://www.nicd.ac.za/diseases-a-z-index/disease-index-covid-19/sars-cov-2-genomic-surveillance-update/
93. Brown KA, Gubbay J, Hopkins J, Patel S, Buchan SA, Daneman N, et al. S-Gene Target Failure as a Marker of Variant B.1.1.7 among SARS-CoV-2 Isolates in the Greater Toronto Area, December 2020 to March 2021. ournal Am Med Assoc. 2021;325:2115–6.
94. Quick J. nCoV-2019 sequencing protocol V.3. Spanish J Leg Med [Internet]. 2020;https://ww:1–17. Available from: https://www.protocols.io/view/ncov-2019-sequencing-protocol-v3-locost-bp2l6n26rgqe/v3
95. NICD. Monthly respiratory pathogens surveillance report South Africa week 5 2022. 2022.
96. Cocherie T, Zafilaza K, Leducq V, Marot S, Calvez V, Marcelin AG, et al. Epidemiology and Characteristics of SARS-CoV-2 Variants of Concern: The Impacts of the Spike Mutations. Microorganisms. 2023;11(30).
97. Nörz D, Grunwald M, Tang HT, Olearo F, Günther T, Robitaille A, et al. Rapid

automated screening for sars-cov-2 b.1.617 lineage variants (delta/kappa) through a versatile toolset of qPCR-based snp detection. Diagnostics. 2021;11:1818.

98. Dejnirattisai W, Huo J, Zhou D, Zahradník J, Supasa P, Liu C, et al. SARS-CoV-2 Omicron-B.1.1.529 leads to widespread escape from neutralizing antibody responses. Cell. 2022;185:467–84.
99. Cao Y, Wang J, Jian F, Xiao T, Song W, Yisimayi A, et al. Omicron escapes the majority of existing SARS-CoV-2 neutralizing antibodies. Nature. 2021;1–24.
100. Wolter N, Jassat W, Walaza S, Welch R, Moultrie H, Groome M, et al. Early assessment of the clinical severity of the SARS-CoV-2 omicron variant in South Africa: a data linkage study. Lancet [Internet]. 2022;399:437–46. Available from: http://dx.doi.org/10.1016/S0140-6736(22)00017-4
101. Scott LE, Hsiao NY, Dor G, Hans L, Marokane P, da Silva MP, et al. How South Africa Used National Cycle Threshold (Ct) Values to Continuously Monitor SARS-CoV-2 Laboratory Test Quality. Diagnostics. 2023;13:1–13.
102. NGS-SA. Network for Genomic Surveillance in South Africa (NGS-SA): SARS-CoV-2 Sequencing Update 22 July 2022. 2022.
103. Ashford F, Best A, Dunn S, Ahmed Z, Siddiqui H, Melville J, et al. SARS-CoV-2 testing in the community: Testing positive samples with the TaqMan SARS-CoV-2 2 Mutation Panel to find variants in real-time. medRxiv Prepr [Internet]. 2021;1–13. Available from: doi: https://doi.org/10.1101/2021.11.17.21266297
104. Takemae N, Doan YH, Momose F, Saito T, Kageyama T. Development of new SNP genotyping assays to discriminate the Omicron variant of SARS-CoV-2. Japanese J Infect Dis Dev. 2022;
105. Yamasoba D, Kimura I, Nasser H, Morioka Y, Nao N, Ito J, et al. Virological characteristics of SARS-CoV-2 BA.2 variant. bioRxiv [Internet]. 2022;2022.02.14.480335. Available from: http://biorxiv.org/content/early/2022/02/15/2022.02.14.480335.abstract
106. Fung TS, Liu DX. Human Coronavirus : Host-Pathogen Interaction. In: Annual Review of Microbiology. 2019. p. 529–60.
107. WHO. Tracking SARS-CoV-2 variants [Internet]. 2022. Available from: https://www.who.int/activities/tracking-SARS-CoV-2-variants
108. Subramoney K, Mtileni N, Bharuthram A, Davis A, Kalenga B, Rikhotso M, et al. Identification of SARS-CoV-2 Omicron variant using spike gene target failure and

genotyping assays, Gauteng, South Africa, 2021. J Med Virol. 2022;1–9.

109. Wang Y, Wang D, Zhang L, Sun W, Zhang Z, Chen W, et al. Intra-host variation and evolutionary dynamics of SARS-CoV-2 populations in COVID-19 patients. Genome Med. 2021;13(1):1–13.
110. Voloch CM, Da Silva Francisco R, De Almeida LGP, Brustolini OJ, Cardoso CC, Gerber AL, et al. Intra-host evolution during SARS-CoV-2 prolonged infection. Virus Evol. 2021;7(2):1–11.
111. Lythgoe KA, Hall M, Ferretti L, de Cesare M, MacIntyre-Cockett G, Trebes A, et al. SARS-CoV-2 within-host diversity and transmission. Science (80-). 2021;372:1–10.
112. San JE, Ngcapu S, Kanzi AM, Tegally H, Fonseca V, Giandhari J, et al. Transmission dynamics of SARS-CoV-2 within-host diversity in two major hospital outbreaks in South Africa. Virus Evol. 2021;7:1–16.
113. Enis Afgan, Dannon Baker, Bérénice Batut, Marius van den Beek, Dave Bouvier, Martin Čech, John Chilton, Dave Clements, Nate Coraor, Björn Grüning, Aysam Guerler, Jennifer Hillman-Jackson, Vahid Jalili, Helena Rasche, Nicola Soranzo, Jeremy Goecks, James T and DB. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. Nucleic Acids Res. 2018;46:W537–W544.
114. Gao R, Zu W, Liu Y, Li J, Li Z, Wen Y, et al. Quasispecies of SARS-CoV-2 revealed by single nucleotide polymorphisms (SNPs) analysis. Virulence. 2021;12:1209–26.
115. Yao H, Lu X, Chen Q, Xu K, Chen Y, Cheng M, et al. Patient-derived SARS-CoV-2 mutations impact viral replication dynamics and infectivity in vitro and with clinical implications in vivo. Cell Discov [Internet]. 2020;6:1–16. Available from: http://dx.doi.org/10.1038/s41421-020-00226-1
116. Madhi SA, Kwatra G, Myers JE, Jassat W, Dhar N, Mukendi CK, et al. South African Population Immunity and Severe Covid-19 with Omicron Variant. medRxiv Prepr. 2021;1–13.
117. Armero A, Berthet N, Avarre JC. Intra-host diversity of sars-cov-2 should not be neglected: Case of the state of Victoria, Australia. Viruses. 2021;13:1–15.
118. Neopane P, Nypaver J, Shrestha R, Beqaj SS. SARS-CoV-2 variants detection using TaqMan SARS-CoV-2 mutation panel molecular genotyping assays. Infect Drug Resist. 2021;14:4471–9.
119. Abbas Q, Kusakin A, Sharrouf K, Jyakhwo S, Komissarov AS. Follow-up

investigation and detailed mutational characterization of the SARS-CoV-2 Omicron variant lineages (BA.1, BA.2, BA.3 and BA.1.1). bioRxiv [Internet]. 2022;1–37. Available from: https://www.biorxiv.org/content/10.1101/2022.02.25.481941v1%0Ahttps://www.biorxiv.org/content/10.1101/2022.02.25.481941v1.abstract

120. Desingu PA, Nagarajan K, Dhama K. Emergence of Omicron third lineage BA.3 and its importance. J Med Virol. 2022;94:1808–10.
121. Eslami S, Glassy MC, Ghafouri-fard S. A comprehensive overview of identified mutations in SARS CoV-2 spike glycoprotein among Iranian patients. Gene. 2022;813:1–11.
122. Núñez-Muñoz L, Marcelino-Pérez G, Calderón-Pérez B, Pérez-Saldívar M, Acosta-Virgen K, González-Conchillos H, et al. Recombinant antigens based on non-glycosylated regions from rbd sars-cov-2 as potential vaccine candidates against covid-19. Vaccines. 2021;9:3–15.
123. Huo J, Mikolajek H, Le Bas A, Clark JJ, Sharma P, Kipar A, et al. A potent SARS-CoV-2 neutralising nanobody shows therapeutic efficacy in the Syrian golden hamster model of COVID-19. Nat Commun. 2021;12:1–18.
124. Miller NL, Clark T, Raman R, Sasisekharan R. A structural dynamic explanation for observed escape of SARS-CoV-2 BA.2 variant mutation S371L/F. bioRxiv Prepr Serv Biol [Internet]. 2022;1–5. Available from: http://www.ncbi.nlm.nih.gov/pubmed/35262083%0Ahttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC8902888
125. Department: Statistics South Africa. Mid-Year population estimates 2021 [Internet]. 2021. Available from: www.statssa.gov.za,info@statssa.gov.za,Tel+27123108911%0Awww.statssa.gov.za/?page_id=1854&PPN=P0302&SCH=72983
126. Meiring S, Tempia S, Bhiman JN, Buys A, Kleynhans J, Makhasi M, et al. Prolonged Shedding of Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) at High Viral Loads Among Hospitalized Immunocompromised Persons Living With Human Immunodeficiency Virus (HIV), South Africa. Clin Infect Dis. 2022;2(Xx).
127. Leung WF, Chorlton S, Tyson J, Al-Rawahi GN, Jassem AN, Prystajecky N, et al. COVID-19 in an immunocompromised host: persistent shedding of viable SARS-CoV-2 and emergence of multiple mutations: a case report. Int J Infect Dis.

2022;114:178–82.

128. Tang X, Sun R, Ge W, Mao T, Qian L, Huang C, et al. Enhanced in fl ammation and suppressed adaptive immunity in COVID-19 with prolonged RNA shedding. Cell Discov. 2022;8:1–18.
129. Long H, Zhao J, Zeng HL, Lu Q Bin, Fang LQ, Wang Q, et al. Prolonged viral shedding of SARS-CoV-2 and related factors in symptomatic COVID-19 patients: a prospective study. BMC Infect Dis [Internet]. 2021;21:1–10. Available from: https://doi.org/10.1186/s12879-021-07002-w
130. Tarcsai KR, Corolciuc O, Tordai A, Ongrádi J. SARS-CoV-2 infection in HIV-infected patients: potential role in the high mutational load of the Omicron variant emerging in South Africa. GeroScience [Internet]. 2022;1–9. Available from: https://doi.org/10.1007/s11357-022-00603-6
131. Jergović M, Uhrlaub JL, Watanabe M, Bradshaw CM, White LM, LaFleur BJ, et al. Competent immune responses to SARS-CoV-2 variants in older adults following two doses of mRNA vaccination. Nat Commun. 2022;13:1–8.
132. Thakur V, Bhola S, Thakur P, Patel SKS, Kulshrestha S, Ratho RK, et al. Waves and variants of SARS-CoV-2: understanding the causes and effect of the COVID-19 catastrophe. Infection [Internet]. 2022;50:309–25. Available from: https://doi.org/10.1007/s15010-021-01734-2
133. Li J, Du P, Yang L, Zhang J, Song C, Chen D, et al. Two-step fitness selection for intra-host variations in SARS-CoV-2. Cell Rep [Internet]. 2022;38:1–14. Available from: https://doi.org/10.1016/j.celrep.2021.110205
134. Karim F, Moosa M, Gosnell B, Cele S, Giandhari J, Pillay S, et al. Persistent SARS-CoV-2 infection and intra-host evolution in association with advanced HIV infection. medRxiv. 2021;1–13.
135. Planas D, Veyer D, Baidaliuk A, Staropoli I, Guivel-Benhassine F, Rajah MM, et al. Reduced sensitivity of SARS-CoV-2 variant Delta to antibody neutralization. Nature [Internet]. 2021;596:276–80. Available from: http://dx.doi.org/10.1038/s41586-021-03777-9
136. Zhou D, Dejnirattisai W, Supasa P, Liu C, Mentzer AJ, Ginn HM, et al. Evidence of escape of SARS-CoV-2 variant B.1.351 from natural and vaccine-induced sera. Cell [Internet]. 2021;184:2348–61. Available from: https://doi.org/10.1016/j.cell.2021.02.037

137. Dhar MS, Marwal R, Radhakrishnan VS, Ponnusamy K, Jolly B, Bhoyar RC, et al. Genomic characterization and epidemiology of an emerging SARS-CoV-2 variant in Delhi, India. Science (80-). 2021;374:995–9.
138. Tegally H, Moir M, Everatt J, Giovanetti M, Scheepers C, Wilkinson E, et al. Emergence of SARS-CoV-2 Omicron lineages BA.4 and BA.5 in South Africa. Nat Med [Internet]. 2022;28:1785–9. Available from: doi: https://doi.org/10.1101/2022.05.01.22274406
139. Department of Health Republic of South Africa. COVID-19 policies [Internet]. Available from: https://www.health.gov.za/covid19/downloads.html?type=policies%0D%0Aaccessed 10 May 2023%0D%0A
140. SAPHRA. South african health products regulatory authority. https://www.sahpra.org.za/news-and-updates/sahpra-authorises-covishield-an-adenovirus-vectored-vaccine-for- the-prevention-of-covid-19/.
141. World Health Organisation. World Health Organisation [Internet]. Available from: https://covid19.who.int/ (accessed 10 May 2023)
142. GISAID. GISAID [Internet]. Available from: https://www.epicov.org/epi3/frontend#30823 (accessed 10 May 2023)%0D%0A
143. Scheepers C, Everatt J, Amoako DG, Tegally H, Wibmer CK, Mnguni A, et al. Emergence and phenotypic characterization of the global SARS-CoV-2 C.1.2 lineage. Nat Commun [Internet]. 2022;13:1–9. Available from: https://doi.org/10.1038/s41467-022-29579-9
144. Brandt C, Spott R, Hölzer M, Kühnert D, Fuchs S, Lohde M, et al. Molecular epidemiology of SARS-CoV-2 - a regional to global perspective. medRxiv [Internet]. 2021;1–9. Available from: https://doi.org/10.1101/2021.01.25.21250447
145. E G, M L, M C. On the effect of age on the transmission of SARS-CoV-2 in households, schools and the community. medRxiv Prepr. 2020;1–16.
146. Cécile TK, Bosetti P, Paireau J, Crépey P, Salje H, Lefrancq N, et al. SARS-CoV-2 transmission across age groups in France and implications for control. Nat Commun. 2021;12:1–12.
147. Phaswana-Mafuya N, Shisana O, Jassat W, Baral SD, Makofane K, Phalane E, et al. Understanding the differential impacts of COVID-19 among hospitalised patients in South Africa for equitable response. South African Med J. 2021;111(11):1084–91.
148. Monod M, Blenkinsop A, Xi X, Hebert D, Bershan S, Tietze S, et al. Age groups that

sustain resurging COVID-19 epidemics in the United States. Science (80-). 2021;371:1–12.

149. National Institute for Communicable Diseases. SARS-CoV-2 genomic surveillance update June 2021 [Internet]. 2021. Available from: https://www.nicd.ac.za/diseases-a-z-index/disease-index-covid-19/sars-cov-2-genomic-surveillance-update/

150. NGS-SA. Network for Genomic Surveillance in South Africa (NGS-SA) SARS-CoV-2 Sequencing Update 7 July 2021 [Internet]. 2021. Available from: https://www.nicd.ac.za/wp-content/uploads/2021/12/Update-of-SA-sequencing-data-from-GISAID-1-Dec-Final.pdf

151. Alteri C, Cento V, Piralla A, Costabile V, Tallarita M, Colagrossi L, et al. Genomic epidemiology of SARS-CoV-2 reveals multiple lineages and early spread of SARS-CoV-2 infections in Lombardy, Italy. Nat Commun [Internet]. 2021;12:1–13. Available from: http://dx.doi.org/10.1038/s41467-020-20688-x

152. Andrés C, Piñana M, Borràs-Bermejo B, González-Sánchez A, García-Cehic D, Esperalba J, et al. A year living with SARS-CoV-2: an epidemiological overview of viral lineage circulation by whole-genome sequencing in Barcelona city (Catalonia, Spain). Emerg Microbes Infect [Internet]. 2022;11:172–81. Available from: https://doi.org/10.1080/22221751.2021.2011617

153. Silva CLJ, Rivero R, Douglas J, Bouckaert R, Arenas CJV, Atkins K, et al. Genomic epidemiology of circulating SARS-CoV-2 variants during first two years of the pandemic in Colombia. medRxiv [Internet]. 2022;1–24. Available from: doi: https://doi.org/10.1101/2022.06.20.22275744

154. Quispe-Ricalde MA, Castelán-Sánchez HG, Meza-Rodríguez PM, Dávila-Ramos S, Sierra JL, Batista-Garcia R, et al. Evidence of natural selection and dominance of SARS-CoV-2 variant Lambda (C.37) over variants of concern in Cusco, Peru. Arch Virol [Internet]. 2023;168:4–23. Available from: https://doi.org/10.1007/s00705-022-05645-x

155. Oróstica KY, Contreras S, Mohr SB, Dehning J, Ulloa S, Castillo AE, et al. Mutational signatures and transmissibility of SARS-CoV-2 Gamma and Lambda variants. arXiv. 2021;1–28.

156. Zambrana Montaño R, Culasso ACA, Fernández F, Marquez N, Debat H, Salmerón M, et al. Evolution of SARS-CoV-2 during the first year of the COVID-19 pandemic in Northwestern Argentina. Virus Res. 2023;323.

157. Ruiz-Sternberg ÁM, Chaparro-Solano HM, Albornóz LL, Pinzón-Rondón ÁM, Pardo-Oviedo JM, Molano-González N, et al. Genomic characterization of SARS-CoV-2 and its association with clinical outcomes: a 1-year longitudinal study of the pandemic in Colombia. Int J Infect Dis. 2022;116:91–100.
158. Mwangi P, Okendo J, Mogotsi M, Ogunbayo A, Adelabu O, Sondlane H, et al. SARS-CoV-2 variants from COVID-19 positive cases in the Free State province, South Africa from July 2020 to December 2021. Front Virol. 2022;2:1–9.
159. Mou K, Abdalla M, Qing D, Tahir M. Emerging mutations in envelope protein of SARS-CoV-2 and their effect on thermodynamic properties. Informatics Med Unlocked. 2020;25:1–10.
160. Liu S, Shen J, Fang S, Li K, Liu J, Yang L, et al. Genetic Spectrum and Distinct Evolution Patterns of SARS-CoV-2. Front Microbiol. 2020;11:1–14.
161. Shen L, Bard JD, Triche TJ, Judkins AR, Biegel JA, Gai X. Emerging variants of concern in SARS-CoV-2 membrane protein: a highly conserved target with potential pathological and therapeutic implications. Emerg Microbes Infect. 2021;10:885–93.
162. Marques-Pereira C, Pires MN, Gouveia RP, Pereira NN, Caniceiro AB, Rosário-Ferreira N, et al. SARS-CoV-2 Membrane Protein: From Genomic Data to Structural New Insights. Int J Mol Sci. 2022;23(6).
163. Xia S, Wang L, Zhu Y, Lu L, Jiang S. Origin, virological features, immune evasion and intervention of SARS-CoV-2 Omicron sublineages. Signal Transduct Target Ther. 2022;7:1–7.
164. Alsuwairi FA, Alsaleh A, Alsanea MS, Al-qahtani AA. Association of SARS-CoV-2 Nucleocapsid Protein Mutations with Patient Demographic and Clinical Characteristics during the Delta and Omicron Waves. medRxiv Prepr. 2023;1–35.
165. Sant'Anna FH, Muterle Varela AP, Prichula J, Comerlato J, Comerlato CB, Roglio VS, et al. Emergence of the novel SARS-CoV-2 lineage VUI-NP13L and massive spread of P.2 in South Brazil. Emerg Microbes Infect. 2021;10:1431–40.
166. Ghaleh SS, Rahimian K, Mahmanzar M, Mahdavi B, Tokhanbigli S, Sisakht MM, et al. SARS-CoV-2 Non-structural protein 1(NSP1) mutation virulence and natural selection: Evolutionary trends in the six continents. Virus Res [Internet]. 2023;323:1–10. Available from: https://doi.org/10.1016/j.virusres.2022.199016
167. Moen LV, Vollan HS, Bråte J, Hungnes O, Bragstad K. Molecular Epidemiology of the Norwegian SARS-CoV-2 Delta Lineage AY.63. Viruses. 2022;14:1–11.

168. Azad GK, Khan PK. Variations in Orf3a protein of SARS-CoV-2 alter its structure and function. Biochem Biophys Reports [Internet]. 2021;26:1–6. Available from: https://doi.org/10.1016/j.bbrep.2021.100933
169. Eskier D, Karakülah G, Suner A, Oktay Y. RdRp mutations are associated with SARS-CoV-2 genome evolution. PeerJ. 2020;8:1–14.
170. Pachetti M, Marini B, Benedetti F, Giudici F, Mauro E, Storici P, et al. Emerging SARS-CoV-2 mutation hot spots include a novel RNA-dependent-RNA polymerase variant. J Transl Med [Internet]. 2020;18:1–9. Available from: https://doi.org/10.1186/s12967-020-02344-6
171. Meng B, Kemp SA, Papa G, Datir R, Ferreira IATM, Marelli S, et al. Recurrent emergence of SARS-CoV-2 spike deletion H69/V70 and its role in the Alpha variant B.1.1.7. Cell Rep. 2021;35:1–24.
172. Takano K, Watanabe Y, Hariu M, Seki M. Detection of representative mutant strains and a case of prolonged infection by SARS-CoV-2 with spike 69/70 deletion in Japan. Infect Drug Resist. 2021;14:2579–81.
173. Chen KWK, Tsung-Ning Huang D, Huang LM. SARS-CoV-2 variants – Evolution, spike protein, and vaccines. Biomed J [Internet]. 2022;45:573–9. Available from: https://doi.org/10.1016/j.bj.2022.04.006
174. Jackson CB, Zhang L, Farzan M, Choe H. Functional importance of the D614G mutation in the SARS-CoV-2 spike protein. Biochem Biophys Res Commun [Internet]. 2021;538:108–15. Available from: https://doi.org/10.1016/j.bbrc.2020.11.026
175. Abavisani M, Rahimian K, Mahdavi B, Tokhanbigli S, Mollapour Siasakht M, Farhadi A, et al. Mutations in SARS-CoV-2 structural proteins: a global analysis. Virol J. 2022;19:1–19.
176. Candido KL, Eich CR, de Fariña LO, Kadowaki MK, da Conceição Silva JL, Maller A, et al. Spike protein of SARS-CoV-2 variants: a brief review and practical implications. Brazilian J Microbiol. 2022;53:1133–57.
177. Asif A, Ilyas I, Abdullah M, Sarfraz S, Mustafa M, Mahmood A. The Comparison of Mutational Progression in SARS-CoV-2: A Short Updated Overview. J Mol Pathol. 2022;3:201–18.
178. World Health Organization. COVID-19 Weekly Epidemiological Update Edition 70 [Internet]. Edition 70, published 14 December 2021. 2021. Available from:

https://www.who.int/publications/m/item/covid-19-weekly-epidemiological-update
179. Mohapatra RK, Kandi V, Tuli HS, Chakraborty C, Dhama K. The recombinant variants of SARS-CoV-2: Concerns continues amid COVID-19 pandemic. J Med Virol. 2022;94:3506–8.
180. World Health Organization. COVID-19 Weekly Epidemiological Update Edition 115 published 26 October 2022. 2022.
181. Focosi D, Quiroga R, McConnell S, Johnson MC, Casadevall A. Convergent Evolution in SARS-CoV-2 Spike Creates a Variant Soup from Which New COVID-19 Waves Emerge. Int J Mol Sci. 2023;24:1–17.
182. Caniels TG, Bontjer I, van der Straten K, Poniman M, Burger JA, Appelman B, et al. Emerging SARS-CoV-2 variants of concern evade humoral immune responses from infection and vaccination. Sci Adv. 2021;7(1–11).
183. Brown B, Ojha V, Fricke I, Al-Sheboul SA, Imarogbe C, Gravier T, et al. Innate and Adaptive Immunity during SARS-CoV-2 Infection: Biomolecular Cellular Markers and Mechanisms. Vaccines. 2023;11:1–60.
184. Walls AC, Park YJ, Tortorici MA, Wall A, McGuire AT, Veesler D. Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. Cell [Internet]. 2020;180:281–92. Available from: http://dx.doi.org/10.1016/j.cell.2020.02.058
185. Silva MJA, Ribeiro LR, Lima KVB, Lima LNGC. Adaptive immunity to SARS-CoV-2 infection: A systematic review. Front Immunol. 2022;13:1–25.
186. Ali F, Kasry A, Amin M. The new SARS-CoV-2 strain shows a stronger binding affinity to ACE2 due to N501Y mutant. Med Drug Discov. 2021;10:1–4.
187. Davies NG, Abbott S, Barnard RC, Jarvis CI, Kucharski AJ, Munday JD, et al. Estimated transmissibility and impact of SARS-CoV-2 lineage B.1.1.7 in England. Science (80-). 2021;372:1–9.
188. Wibmer CK, Ayres F, Hermanus T, Madzivhandila M. SARS-CoV-2 501Y . V2 escapes neutralization by South African COVID-19 donor plasma. Nat Med [Internet]. 2021;27:622–5. Available from: https://www.nature.com/articles/s41591-021-01285-x
189. Moyo-gwete T, Madzivhandila M, Makhado Z, Ayres F, Burgers WA. Moyo-Gwete 2021 SARS-CoV-2 501Y.V2 (B.1.351) elicits cross-reactive neutralizing antibodies Biorxiv submission. 2021;2.
190. Riou C, Keeton R, Moyo-Gwete T, Hermanus T, Kgagudi P, Baguma R, et al. Escape from recognition of SARS-CoV-2 variant spike epitopes but overall preservation of T

cell immunity. Sci Transl Med. 2022;14:1–13.
191. Chakraborty C, Sharma AR, Bhattacharya M, Lee SS. A Detailed Overview of Immune Escape, Antibody Escape, Partial Vaccine Escape of SARS-CoV-2 and Their Emerging Variants With Escape Mutations. Front Immunol. 2022;13:1–26.
192. Geers D, Shamier MC, Bogers S, den Hartog G, Gommers L, Nieuwkoop NN, et al. SARS-CoV-2 variants of concern partially escape humoral but not T-cell responses in COVID-19 convalescent donors and vaccinees. Sci Immunol. 2021;6:1–14.
193. Changrob S, Fu Y, Guthmiller JJ, Halfmann PJ, Li L, Stamper CT, et al. Cross-Neutralization of Emerging SARS-CoV-2 Variants of Concern by Antibodies Targeting Distinct Epitopes on Spike. Am Soc Microbiol. 2021;12:1–15.
194. Mengist HM, Arnaud J, Kombe K, Jin T. Immune Evasion by the Highly Mutated SARS-CoV-2 Omicron Variant. Infect Drug Resist. 2022;15:4013–27.
195. Zhou Y, Zhi H, Teng Y. The outbreak of SARS-CoV-2 Omicron lineages, immune escape, and vaccine effectivity. J Med Virol. 2022;95:1–9.
196. Cele S, Jackson L, Khoury DS, Khan K, Moyo-Gwete T, Tegally H, et al. Omicron extensively but incompletely escapes Pfizer BNT162b2 neutralization. Nature. 2021;602:654–6.
197. Thurmond J, Yoon H, Kuiken C, Yusim K, Perkins S, Theiler J, et al. Web-based design and evaluation of T-cell vaccine candidates. Bioinformatics. 2008;24:1639–40.
198. Gupta S, Kapoor P, Chaudhary K, Gautam A, Kumar R, Raghava GPS. In Silico Approach for Predicting Toxicity of Peptides and Proteins. PLoS One. 2013;8:1–16.
199. Martínez-Flores D, Zepeda-Cervantes J, Cruz-Reséndiz A, Aguirre-Sampieri S, Sampieri A, Vaca L. SARS-CoV-2 Vaccines Based on the Spike Glycoprotein and Implications of New Viral Variants. Front Immunol. 2021;12:1–21.
200. Duijn J Van, Stieh D, Fernandez N, King D, Gilmour J, Tolboom J, et al. Mosaic HIV-1 vaccination induces anti-viral CD8+ T cell functionality in the phase 1/2a clinical trial APPROACH. J Virol. 2023;0:1–18.
201. Hou J, Shrivastava S, Fraser CC, Loo HL, Wong LH, Ho V, et al. Dengue mosaic vaccines enhance cellular immunity and expand the breadth of neutralizing antibody against all four serotypes of dengue viruses in mice. Front Immunol. 2019;10:1–10.
202. Ysrafil Y, Imran AK, Wicita PS, Kamba V, Mohamad F, Ismail I, et al. Mosaic vaccine design targeting mutational spike protein of SARS-CoV-2: an Immunoinformatics Approaches. BioImpacts. 2023;1–33.

203. Aloor A, Aradhya R, Venugopal P, Nair BG. Glycosylation in SARS-CoV-2 variants: A path to infection and recovery. Biochem Pharmacol. 2022;206:1–14.
204. Gong Y, Qin S, Dai L. The glycosylation in SARS-CoV-2 and its receptor ACE2. Signal Transduct Target Ther. 2021;6:1–23.
205. Peter AS, Grüner E, Socher E, Fraedrich K, Richel E, Mueller-Schmucker S, et al. Characterization of SARS-CoV-2 Escape Mutants to a Pair of Neutralizing Antibodies Targeting the RBD and the NTD. Int J Mol Sci. 2022;23:1–16.
206. Wu Y, Wang S, Zhang Y, Yuan L, Zheng Q, Wei M, et al. Lineage-mosaic and mutation-patched spike proteins for broad-spectrum COVID-19 vaccine. Cell Host Microbe [Internet]. 2022;30:1732–44. Available from: https://doi.org/10.1016/j.chom.2022.10.011

Appendix A

Received: 9 March 2022 | Accepted: 9 April 2022

DOI: 10.1002/jmv.27797

RESEARCH ARTICLE

JOURNAL OF MEDICAL VIROLOGY WILEY

Identification of SARS-CoV-2 Omicron variant using spike gene target failure and genotyping assays, Gauteng, South Africa, 2021

Kathleen Subramoney[1,2] | Nkhensani Mtileni[1] | Avani Bharuthram[1,2] | Ashlyn Davis[1] | Beauty Kalenga[1] | Mikateko Rikhotso[1] | Mpho Maphahlele[1] | Jennifer Giandhari[3] | Yeshnee Naidoo[3] | Sureshnee Pillay[3] | Upasana Ramphal[3] | Yajna Ramphal[3] | Houriiyah Tegally[3] | Eduan Wilkinson[4] | Thabo Mohale[5] | Arshad Ismail[5] | Bonolo Mashishi[1,2] | Nonhlanhla Mbenenge[1,2] | Tulio de Oliveira[3,4] | Zinhle Makatini[1,2] | Burtram C. Fielding[6] | Florette K. Treurnicht[1,2] | Network for Genomics Surveillance in South Africa

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

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

[3]KwaZulu-Natal Research Innovation and Sequencing Platform (KRISP), Nelson R Mandela School of Medicine, University of KwaZulu-Natal, Durban, South Africa

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

[5]Sequencing Core Facility, National Institute for Communicable Diseases, Johannesburg, South Africa

[6]Department of Medical Biosciences, Molecular Biology and Virology Research Laboratory, University of the Western Cape, Bellville, South Africa

Correspondence
Kathleen Subramoney, Department of Virology, National Health Laboratory Service, Charlotte Maxeke Johannesburg Academic Hospital, Private Bag X39, Parktown, Johannesburg, Gauteng 2193, South Africa.
Email: kathleen.subramoney@nhls.ac.za

Florette K. Treurnicht, School of Pathology, Faculty of Health Sciences, University of the Witwatersrand, Private Bag 3, Parktown, Johannesburg, Gauteng 2193, South Africa.
Email: florette.treurnicht@nhls.ac.za

Funding Information
Medical Research Council; National Research Foundation; National Health Laboratory Service

Abstract

The circulation of Omicron BA.1 led to the rapid increase in severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) cases in South Africa in November 2021, which warranted the use of more rapid detection methods. We, therefore, assessed the ability to detect Omicron BA.1 using genotyping assays to identify specific mutations in SARS-CoV-2 positive samples, Gauteng province, South Africa. The TaqPath™ COVID-19 real-time polymerase chain reaction assay was performed on all samples selected to identify spike gene target failure (SGTF). SARS-CoV-2 genotyping assays were used for the detection of del69/70 and K417N mutation. Whole-genome sequencing was performed on a subset of genotyped samples to confirm these findings. Of the positive samples received, 11.0% (175/1589) were randomly selected to assess if SGTF and genotyping assays, that detect del69/70 and K417N mutations, could identify Omicron BA.1. We identified SGTF in 98.9% (173/175) of samples, of which 88.0% (154/175) had both the del69/70 and K417N mutation. The genotyped samples (45.7%; 80/175) that were sequenced confirmed Omicron BA.1 (97.5%; 78/80). Our data show that genotyping for the detection of the del69/70 and K417N coupled with SGTF is efficient to exclude

Appendix B

PLOS ONE

RESEARCH ARTICLE

SARS-CoV-2 spike protein diversity at an intra-host level, among SARS-CoV-2 infected individuals in South Africa, 2020 to 2022

Kathleen Subramoney[1,2]*, Nkhensani Mtileni[2], Ashlyn Davis[1], Jennifer Giandhari[3], Houriiyah Tegally[3,4], Eduan Wilkinson[4], Yeshnee Naidoo[4], Yajna Ramphal[3], Sureshnee Pillay[3], Upasana Ramphal[3], Andiswa Simane[2], Bhaveshan Reddy[1,2], Bonolo Mashishi[1,2], Nonhlanhla Mbenenge[1,2], Tulio de Oliveira[3,4], Burtram C. Fielding[5], Florette K. Treurnicht[1,2]

1 School of Pathology, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa, **2** Department of Virology, National Health Laboratory Service, Charlotte Maxeke Johannesburg Academic Hospital, Johannesburg, South Africa, **3** KwaZulu-Natal Research Innovation and Sequencing Platform (KRISP), Nelson R Mandela School of Medicine, University of KwaZulu-Natal, Durban, South Africa, **4** Centre for Epidemic Response and Innovation (CERI), School of Data Science and Computational Thinking, Stellenbosch University, Stellenbosch, South Africa, **5** Molecular Biology and Virology Research Laboratory, Department of Medical BioSciences, University of the Western Cape, Cape Town, South Africa

* kathleen.subramoney@nhls.ac.za

Citation: Subramoney K, Mtileni N, Davis A, Giandhari J, Tegally H, Wilkinson E, et al. (2023) SARS-CoV-2 spike protein diversity at an intra-host level, among SARS-CoV-2 infected individuals in South Africa, 2020 to 2022. PLoS ONE 18(5): e0286373. https://doi.org/10.1371/journal.pone.0286373

Editor: Nagarajan Raju, Emory University, UNITED STATES

Received: December 8, 2022

Accepted: May 15, 2023

Published: May 30, 2023

Data Availability Statement: Sequence data for heterogeneous SARS-CoV-2 samples and associated metadata in this dataset are published in GISAID's EpiCoV database (16). Data can be accessed with the GISAID Identifier: EPI_SET_230423vk (doi: 10.55876/gis8.230423vk) with the linked fastq files, and using accession IDs from S4 Table to download data from NCBI SRA (https://www.ncbi.nlm.nih.gov/sra).

Abstract

Intra-host diversity studies are used to characterise the mutational heterogeneity of SARS-CoV-2 infections in order to understand the impact of virus-host adaptations. This study investigated the frequency and diversity of the spike (S) protein mutations within SARS-CoV-2 infected South African individuals. The study included SARS-CoV-2 respiratory samples, from individuals of all ages, received at the National Health Laboratory Service at Charlotte Maxeke Johannesburg Academic hospital, Gauteng, South Africa, from June 2020 to May 2022. Single nucleotide polymorphism (SNP) assays and whole genome sequencing were performed on a random selection of SARS-CoV-2 positive samples. The allele frequency (AF) was determined using TaqMan Genotyper software for SNP PCR analysis and galaxy.eu for analysis of FASTQ reads from sequencing. The SNP assays identified 5.3% (50/948) of Delta cases with heterogeneity at delY144 (4%; 2/50), E484Q (6%; 3/50), N501Y (2%; 1/50) and P681H (88%; 44/50), however only heterogeneity for E484Q and delY144 were confirmed by sequencing. From sequencing we identified 9% (210/2381) of cases with Beta, Delta, Omicron BA.1, BA.2.15, and BA.4 lineages that had heterogeneity in the S protein. Heterogeneity was primarily identified at positions 19 (1.4%) with T19IR (AF 0.2–0.7), 371 (92.3%) with S371FP (AF 0.1–1.0), and 484 (1.9%) with E484AK (0.2–0.7), E484AQ (AF 0.4–0.5) and E484KQ (AF 0.1–0.4). Mutations at heterozygous amino acid positions 19, 371 and 484 are known antibody escape mutations, however the impact of the combination of multiple substitutions identified at the same position is unknown. Therefore, we hypothesise that intra-host SARS-CoV-2 quasispecies with heterogeneity in the S protein facilitate competitive advantage of variants that can completely/partially evade host's natural and vaccine-induced immune responses.

Appendix C

viruses

MDPI

Article

Molecular Epidemiology of SARS-CoV-2 during Five COVID-19 Waves and the Significance of Low-Frequency Lineages

Kathleen Subramoney [1,2,3,*], Nkhensani Mtileni [2], Jennifer Giandhari [4], Yeshnee Naidoo [5], Yajna Ramphal [4], Sureshnee Pillay [4], Upasana Ramphal [4], Akhil Maharaj [4], Derek Tshiabuila [5], Houriiyah Tegally [4], Eduan Wilkinson [5], Tulio de Oliveira [4,5], Burtram C. Fielding [6] and Florette K. Treurnicht [1,2]

[1] School of Pathology, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg 2193, South Africa; florette.treurnicht@nhls.ac.za
[2] Department of Virology, National Health Laboratory Service, Charlotte Maxeke Johannesburg Academic Hospital, Johannesburg 2193, South Africa
[3] Centre for Vaccines and Immunology, National Institute for Communicable Diseases, Johannesburg 2131, South Africa
[4] KwaZulu-Natal Research Innovation and Sequencing Platform (KRISP), Nelson R Mandela School of Medicine, University of KwaZulu-Natal, Durban 4001, South Africa
[5] Centre for Epidemic Response and Innovation (CERI), School of Data Science and Computational Thinking, Stellenbosch University, Stellenbosch 7600, South Africa
[6] Molecular Biology and Virology Research Laboratory, Department of Medical BioSciences, University of the Western Cape, Cape Town 7535, South Africa
* Correspondence: kathleens@nicd.ac.za

Abstract: SARS-CoV-2 lineages and variants of concern (VOC) have gained more efficient transmission and immune evasion properties with time. We describe the circulation of VOCs in South Africa and the potential role of low-frequency lineages on the emergence of future lineages. Whole genome sequencing was performed on SARS-CoV-2 samples from South Africa. Sequences were analysed with Nextstrain pangolin tools and Stanford University Coronavirus Antiviral & Resistance Database. In 2020, 24 lineages were detected, with B.1 (3%; 8/278), B.1.1 (16%; 45/278), B.1.1.348 (3%; 8/278), B.1.1.52 (5%; 13/278), C.1 (13%; 37/278) and C.2 (2%; 6/278) circulating during the first wave. Beta emerged late in 2020, dominating the second wave of infection. B.1 and B.1.1 continued to circulate at low frequencies in 2021 and B.1.1 re-emerged in 2022. Beta was outcompeted by Delta in 2021, which was thereafter outcompeted by Omicron sub-lineages during the 4th and 5th waves in 2022. Several significant mutations identified in VOCs were also detected in low-frequency lineages, including S68F (E protein); I82T (M protein); P13L, R203K and G204R/K (N protein); R126S (ORF3a); P323L (RdRp); and N501Y, E484K, D614G, H655Y and N679K (S protein). Low-frequency variants, together with VOCs circulating, may lead to convergence and the emergence of future lineages that may increase transmissibility, infectivity and escape vaccine-induced or natural host immunity.

Keywords: SARS-CoV-2; low frequency; lineages; molecular epidemiology

check for updates

Citation: Subramoney, K.; Mtileni, N.; Giandhari, J.; Naidoo, Y.; Ramphal, Y.; Pillay, S.; Ramphal, U.; Maharaj, A.; Tshiabuila, D.; Tegally, H.; et al. Molecular Epidemiology of SARS-CoV-2 during Five COVID-19 Waves and the Significance of Low-Frequency Lineages. *Viruses* **2023**, *15*, 1194. https://doi.org/10.3390/v15051194

Academic Editor: Ulrich Desselberger

Received: 26 April 2023
Revised: 12 May 2023
Accepted: 15 May 2023
Published: 18 May 2023
Corrected: 4 July 2023

1. Introduction

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is continually evolving, resulting in the emergence of several lineages and variants of concern (VOC). The first VOC detected was Alpha (B.1.1.7) which originated in the UK [1], followed by Beta (B.1.351), which originated in South Africa (SA) in August 2020 [2,3]. Delta (B.1.617.2) was initially detected in India in April 2021 but spread rapidly to other countries, including SA, and led to the emergence of several highly transmissible sub-lineages [4]. The most recent VOC identified was Omicron (B.1.1.529), which is divided into BA.1, BA.2, BA.4 and BA.5 lineages and several sub-lineages [5]. Omicron BA.1 initially circulated in November