Molecular identification of SARS-CoV-2 variants of concern at urban wastewater treatment plants across South Africa

Mukhlid Yousif[1,2,#], Said Rachida[1], Setshaba Taukobong[1], Nkosenhle Ndlovu[1], Chinwe Iwu-Jaja[1], Wayne Howard[1], Shelina Moonsamy[1], Nompilo Mhlambi[1], Sipho Gwala[1], Joshua I. Levy[3], Kristian G. Andersen[3], Cathrine Scheepers[4], Anne von Gottberg[5,6], Nicole Wolter[5,6], Arshad Ismail[7,8], Melinda Suchard[9], Kerrigan McCarthy[1,10] for the SACCESS network.

[1]Centre for Vaccines and Immunology, National Institute for Communicable Diseases, a division of the National Health Laboratory Service, South Africa

[2]Department of Virology, School of Pathology, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg

[3]Department of Immunology and Microbiology, The Scripps Research Institute, La Jolla, CA 92037, USA

[4]SAMRC Antibody Immunity Research Unit, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa.

[5]Centre for Respiratory Diseases and Meningitis, National Institute for Communicable Diseases, a division of the National Health Laboratory Service, South Africa

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

[7]Sequencing Core Facility, National Institute for Communicable Diseases, a division of the National Health Laboratory Service, South Africa.

[8]Department of Biochemistry and Microbiology, Faculty of Science, Engineering and Agriculture, University of Venda, Thohoyandou, South Africa.

[9]Department of Chemical Pathology, School of Pathology, University of the Witwatersrand, Johannesburg.

[10]School of Public Health, University of the Witwatersrand, Johannesburg

[#] **Corresponding author**

Dr Mukhlid Yousif

Centre for Vaccines and Immunology

National Institute for Communicable Diseases

Researcher, Virology department

Faculty of Health Sciences

University of the Witwatersrand

Tel: +27 11 386 6461

Email: mukhlidy@nicd.ac.za, Mukhlid.yousif@gmail.com

Keywords: environmental, wastewater, mutations, sequence, whole genome sequencing, SARS-CoV-2, COVID-19, surveillance, minor, spike, alpha, beta, delta, omicron, C1.2

Abstract

The use of wastewater for SARS-CoV-2 surveillance is a useful complementary tool to clinical surveillance. The aims of this study were to characterize SARS-CoV-2 from wastewater samples, and to identify variants of concern present in samples collected from wastewater treatment plants in South African urban metros from April 2021 to January 2022. A total of 325 samples were collected from 15 wastewater treatment plants. Nucleic acids were extracted from concentrated samples, and subjected to amplicon-based whole genome sequencing. To identify variants of concerns and lineages, we used the Freyja tool (https://github.com/andersen-lab/Freyja), which assigns each sample with the prevalence of each variant present. We also used signature mutation analysis to identify variants in each wastewater treatment site. A heatmap was generated to identify patterns of emerging mutations in the spike gene using Excel conditional formatting. Using the Freyja tool, the Beta variant was detected and became predominate from April to June 2021 followed by the Delta variant and lastly the Omicron variant. Our heatmap approach was able to identify a pattern during the changes of predominate variant in wastewater with the emergence of mutations and the loss of others. In conclusion, sequencing of SARS-CoV-2 from wastewater largely corresponded with sequencing from clinical specimens. Our heatmap has the potential to detect new variants prior to emergence in clinical samples and this may be particularly useful during times of low disease incidence between waves, when few numbers of positive

clinical samples are collected and submitted for testing. A limitation of wastewater sequencing is that it is not possible to identify new variants, as variants are classified based on known mutations in clinical strains.

Background

As SARS-CoV-2 is shed into stool and urine, and is detectable in wastewater [1], quantification and sequencing of SARS-CoV-2 in wastewater has the potential to overcome inherent limitations in clinically-based epidemiological approaches. Over the pandemic, clinical surveillance has relied on testing and sequencing of samples from infected individuals. However, when clinical testing forms the basis for surveillance, population health seeking behaviour, test accessibility and testing practices of attending clinicians limit the generalisability of data. In particular, only symptomatic patients approach the health system for testing and testing practices vary by location and over time [2], leading to an incomplete representation of local virus spread and diversity. Testing of wastewater for SARS-CoV-2 levels overcomes these limitations by allowing population levels of SARS-CoV-2 to be monitored over time and by location, adding key information to our understanding of SARS-CoV-2 transmission dynamics. The value of wastewater monitoring of SARS-CoV-2 is attested to by the fact that over 70 countries now provide monitoring and public reporting of geographical and temporal trends in wastewater levels [3 4]

Wastewater genomic surveillance offers an opportunity to monitor circulating variants present in the community. Whole genome sequencing [5], and other methods such as real-time PCR [6] enable detection and characterisation of SARS-CoV-2 variants, and have now been applied to wastewater samples. Recent work has shown the potential for recovery of complete virus genomes from wastewater[7], shown comparable results of wastewater and clinical surveillance, and identified novel mutations and lineages before appearance in clinical samples [8 2]. To date, wastewater sequencing of SARS-CoV-2 has not been widely applied in low or middle income countries.

South Africa is a middle-income country with a population of over 55 million persons, most of whom live in urban centres located in five of the country's nine provinces. South Africa has over 1,000 wastewater treatment plants [9] and the majority of South Africans (84%) have access to piped sanitation (flush toilets connected to a public sewerage system or a septic tank) [10]. Wastewater testing for SARS-CoV-2 was first described in South Africa in June 2020 [11]. The South African Collaborative

COVID-19 Environmental Surveillance System (SACCESS) arose to monitor trends in SARS-CoV-2 levels in wastewater [12].

Most laboratory testing is provided to the public through an extensive network of laboratories including the National Health Laboratory Service (NHLS) that covers over 80% of the population. South Africa identified its first case of COVID-19 on the 5th of March 2020, [13] and four waves of COVID-19 occurred within the first 24 months of virus introduction into the country. Following the initial SARS-CoV-2 wave , the Beta variant [14] was discovered and was predominant from November 2020 to February 2021 (second wave). The third wave (May to September 2021) was characterized by the dominance of the Delta variant [1516] and the fourth wave (November 2021 to January 2022) by the Omicron BA.1 variant [17]. The National Institute for Communicable Diseases (NICD), a division of the NHLS, provides SARS-CoV-2 epidemiological surveillance data through collation of SARS-CoV-2 PCR results from public and private laboratories. The Network for Genomics Surveillance in South Africa (NGS-SA) monitors the epidemiology of SARS-CoV-2 variants in PCR-confirmed cases in South Africa and reports weekly on findings [18]. The NGS-SA provided the first global reports of the emergence of Beta and Omicron variants of concern (VOC) [1417].

Here, we show that wastewater can be used to effectively characterize SARS-CoV-2 virus dynamics in the population and identify variants of concern present using samples from sentinel wastewater treatment plants in urban metros collected from April, 2021 to the end of the fourth wave in January, 2022, demonstrating the utility of wastewater genomic surveillance to complement clinical surveillance efforts in a middle income setting. We identify the potential strengths and limitations of genomic wastewater surveillance for SARS-CoV-2 in the South African context.

Methods

Wastewater sites

A total of 325 samples were collected from 15 wastewater treatment plants (WWTP) in metropolitan areas also undergoing quantitative analysis of SARS-CoV-2 by the NICD [19]. Sites were situated in Gauteng, Eastern Cape, Western Cape, Free State, and KwaZulu- Natal provinces. Supplementary table

S1 shows population sizes draining to each WWTP, and number of samples collected and sequenced. The samples were collected between April 2021 to January 2022 during the third and fourth waves of SARS-CoV-2 infections in South Africa.

Sample collection, RNA extraction, amplification and sequencing

One liter of grab sewage samples were collected and transported to NICD at 4°C. Viruses were concentrated from the sample by ultrafiltration [20], and RNA was extracted using the QIAamp Viral RNA kit (Qiagen, GmbH, Germany). SARS-CoV-2 was detected by RT-PCR using the Allplex™ 2019- nCoV Assay from Seegene (Seoul, Korea). RNA was re-extracted from SARS-CoV-2 positive concentrates and subjected to amplicon-based whole genome sequencing using the Sinai protocol with some modifications as described [21] [22]. Paired-end libraries were prepared using Illumina COVIDSeq Kit as previously described [23] followed by sequencing (2x 150 pb) on NextSeq 1000/2000 platofrm (Illumina Inc, USA).

Sequence analysis

Quality control checks

FastQ files were trimmed, filtered based on sequence quality, assembled and mapped to the reference genome (NC_045512.2) according to published criteria [24] using Exatype web-based bioinformatics tool (https://sars-cov-2.exatype.com/). Quality control indicators such as number of reads, number of mapped reads, and sequence coverage were recorded. Samples that passed the internal quality control were processed for mutation analysis using ARTIC protocol (https://artic.network/ncov-2019/ncov2019-bioinformatics-sop.html) in Galaxy (https://usegalaxy.eu/) [25]. Again, reads were trimmed and filtered, assembled and mapped. At least 10 reads were required at each nucleotide position for downstream analysis. Amino acid mutations present at 5% of reads or less were removed from the analysis. Table 1 illustrates an example of amino acid variation analysis output.

Sequence analysis using amino acid mutations

Amino acids variation analysis

As SARS-CoV-2 RNA in wastewater is fragmented, and fragments originate from multiple individuals (generally infected with genetically distinct viruses), the generation of consensus sequences from wastewater samples is not meaningful. Rather, we inferred the presence of variants by using amino acid mutations uniquely associated with each VOC, as follows: Using the amino acid variation data file generated by the Galaxy pipeline above, we used STATA software (v 17.1) (https://www.stata.com/) to collate spike-gene mutations in a matrix such that the columns represented the amino acid positions of the spike protein and each row recorded mutations identified from a single wastewater sample. We included all mutations, and recorded the proportion of reads where that mutation was detected (the 'read frequency') as a percentage of total reads. For each VOC or Variant of Interest (VOI), we identified signature single amino acid mutations by comparing the new variant/ lineage with the Wuhan reference sequence in a public database [26] (Table 2). Using this list of unique mutations for each VOC and VOI in the spike protein region (Table 2) we interrogated our matrix for the presence or absence of known signature mutations in each sample using STATA software. As new variants/lineages were detected and identified in clinical specimens, we added signature mutations to the STATA code, allowing us to identify the presence of new variants both retrospectively and prospectively.

Heatmap and dot blot

Using the amino acid variations data output file from Galaxy and the generated excel file that contains all amino acid variations and their respective read frequency, a heatmap was generated to identify patterns of emerging mutations in the spike gene using Excel conditional formatting. To identify the events and times at which uncommon mutations were identified in our samples, an in-house R script (R v.4.2.0) was used to generate a mutational dot plot.

Analysis using the Freyja tool

To capture the dynamics of virus evolution and spread, we used Freyja [27] , a tool to estimate the relative abundance of virus lineages present in wastewater. Freyja uses a "barcode" library of lineage-defining mutations to uniquely define all known SARS-CoV-2 lineages and solves for lineage abundance using a depth-weighted, least absolute deviation regression approach. Freyja is free to use and available at https://github.com/andersen-lab/Freyja.

Results

Quality control

A total of 325 wastewater samples from sites listed in Table S1 were amplified and sequenced. The median number of sequence reads was $1.72x10^6$ (inter quartile range [IQR] $2.53X10^5$-$2.67X10^6$) with a range of $6.5x10^4$ to $8.16x10^6$reads (Figure S1). A total of 229 (70.5%) samples had > 1 million reads, and 237 (72.9%) samples had >50% reads mapped to the reference sequence (range 0-89%). Regarding sequence coverage in 10x depth, 183 (56.3%) samples had >50% sequence coverage of the whole genome (Figure S2a), and 177 (54.5%) samples had >50% sequence coverage in the spike region (Figure S2b).

Detection of SARS-CoV-2 variants from wastewater samples using signature mutation analysis

Signature mutations (Table 2) were identified in 170 samples (52.3%), 79 samples from Gauteng, 32 from KwaZulu-Natal, 32 Free State, 12 from Western Cape, and 15 from the Eastern Cape provinces respectively. Figure 1 illustrates the signature mutations for VOCs that were identified in samples from each wastewater treatment plant. In most of wastewater treatment plants a transition change in VOCs was observed in all wastewater treatment plants in the country, from Beta to Delta to Omicron during post second wave, third wave of fourth wave of infections. C.1.2 was also detected during Delta wave.

Detection of SARS-CoV-2 variants from wastewater samples using Freyja tool

Out of the 325 samples sequenced, 168 (51.7%) samples had a sequence coverage of >50% at 10x depth, and were successfully assigned a SARS-CoV-2 variant using the Freyja tool (Figure 2). In these samples, the Beta variant was detected and became predominate from April to June 2021. Delta variant emerge in May 2021 and predominates in June until October 2021. Omicron variant was detected in November 2021 and immediately dominate until January 2022. Alpha variant was seen in small proportion in June and July 2021 and lineage A and variant Kappa was detected in June 2021 in a very small proportion (Figure 2a). The proportions of lineages were shown in Figure 2b, a total of 68 lineages were detected of which 53 lineages were also reported in clinical cases and 15 lineages were not reported in clinical cases.

Characterization of amino acid mutations in the spike region

A total of 411 amino acid mutations were observed in the spike protein amongst all sequenced samples. Alignment by amino acid position in a heatmap (Figure 3) demonstrated a characteristic pattern of mutations in each epidemiological wave of COVID-19. The transition from Delta variant to Omicron was characterized by a loss of mutations in the N-terminal domain (NTD) region (E156del, F157del, and R158G), and new mutations in the receptor binding (RBD) domain (G339D, S371L, 373, N440K, S477N, E484A, Q493R, G496S, Q498R), and fusion peptide (FP) region (N764K, D796Y), and the heptad repeat 1 (HR1) region (Q954H, N969K, L9811F). Between the third and fourth wave of infection low sequence coverage of spike was observed, likely due to low caseload, and few mutations were detected. Of the 411 substitutions/ deletions detected, 78 were present at >1% prevalence. When we compared those mutations to known published mutations at GISAID (https://gisaid.org/), 58 were commonly reported (Table S2), and 10 were uncommonly reported (Figure 4).

Discussion:

In this study, our sequencing methodology and bioinformatics pipeline facilitated detection and genomic characterization of SARS-CoV-2 lineages present in South African wastewater treatment plants in five provinces during the period April 2021 to January 2022. Our results showed the presence of all VOCs also detected amongst clinical samples (Beta, Delta, and Omicron). We detected amino acid mutations that were well-described and available publically in the GISAID database, and also amino acid mutations that were uncommon or rarely reported in clinical samples. We were also able to show the emergence of new mutations and loss of other mutations during the time of wave changes from Beta to Delta and from Delta to Omicron. Collectively, these findings illustrate how sequence analysis of SARS-CoV-2 in wastewater complements epidemiological findings based on clinical sequencing.

Wastewater is a complex matrix containing highly fragmented virus genomic material. Our methodology generated a high number of reads with good quality and depth when compared to other studies that have sequenced SARS-CoV-2 from wastewater[28]. In addition, the highest proportion of mapped reads was as 89% which was in the range of what have been achieved in other studies [2]. The

highest sequence coverage in a depth of 10x reads in this study was 99% for the whole genome and spike protein.

Our approach to interpretation of sequence data generated from wastewater samples is unlike analysis of sequences derived from clinical specimens. Because wastewater contains a mix of RNA fragments from viral particles originating from many infected individuals, the generation of consensus sequences is not meaningful. Our approach allows for identification of previously described variants in wastewater samples and also for detection of new patterns of mutations suggesting previously undescribed variants. Similarly, this approach has also been used Crits-Chritoph and colleague [8].

Regarding detection of previously described variants in our wastewater samples, both approaches we used (mutational analysis and the Freyja tool) successfully identified lineages in wastewater that corresponded to lineages identified in clinical specimens [18]. The Beta variant (first described in clinical specimens in South Africa in December 2020 [14]) was consistently observed from the start of wastewater surveillance until the end of the third wave in epidemiological week 19 (May 2021). Similarly, the Delta variant was first seen in clinical specimens during epidemiological week 21 of 2021 [16], and was first detected in wastewater samples the same week. Lineage C.1.2, described first in South Africa [29] was successfully detected in sequences from wastewater samples during week 22 to 45 in 2021 whilst clinical detections of this lineage appeared from week 16 to 46. In epidemiological week 46 in 2021, the Omicron variant was identified in clinical samples whilst sequences from wastewater samples also identified mutations specific to Omicron in the same week. The Freyja tool [30] complemented our mutational profile analysis, identifying the proportions of each variant/lineage in wastewater at specific time points. Results from Freyja comparable to the prevalence of VOCs of SARS-CoV-2 reported from clinical specimens and additionally indicate the presence of lineages in our wastewater samples that were absent amongst sequences from clinical cases. This most likely arises through the sampling bias inherent in clinical surveillance, in which only symptomatic patients are tested and of whom only a fraction was sequenced.

Regarding detection of previously undescribed variants, our use of a spike-protein heatmap illustrated how each variant had a distinct mutational profile of RNA sequences of the spike gene, and that this changed in each wave. Through observation of the spike protein heatmap, samples with changing profiles may be identified before the new variant is sequenced from clinical isolates. This is clearly evident in the transition from the Delta to the Omicron variant, where a constellation of mutations in the NTD fell away and RBD, FP, and HR1 regions had new mutations (Figure 3). The shifting mutational profile correlated with the increased transmissibility of the Omicron variant that led to the fourth wave of infection in South Arfrica[17].

Our mutational analysis identified multiple instances of rare mutations in the population (Figure 4). These mutations were found at a prevalence of >1.0% in wastewater samples, but were detected at <0.001% in clinical cases based on the data from GISAID. Although they were uncommon, some mutations were reported previously. Mutation S50L was found to be associated with reduced protein stability [31]. Mutation Q498H has reportedly caused increased binding affinity of RBD to ACE2 [32]. The presence of uncommon mutations could be explained as “cryptic lineages” as described by Smyth and colleagues [33]. In their paper they attributed the cryptic lineages to either un-sampled cases or spillover of SARS-CoV-2 from an unidentified animal reservoir. Other wastewater studies have demonstrated cryptic variants, for example Yaniv and colleagues described a cryptic Delta variant in wastewater samples [34]. Unusual substitutions /deletions are useful tools that support epidemiological tracking of variants and may support hypothesis generation regarding the origins of SARS-CoV-2.

Sequencing of SARS-Cov-2 in wastewater currently has a number of limitations. Refining methodological approaches is essential in order to prevent inhibition from substances within the wastewater matrix. Where the incidence of SARS-CoV-2 is very low, virus concentration in wastewater may proceed below the level of detection, which makes it difficult to amplify and sequence the viral genome. Further, emergence of a new variants, such as Omicron sub-variants BA.1 and BA.2, can lead to poor primer binding and lower coverage rates, particularly in the spike protein because of S gene dropout [35]. Both of these scenarios renders sequencing of SARS-CoV-2 from wastewater challenging. In addition, bioinformatics methods for wastewater, including mutational

analysis and the Freyja tool, are currently limited by their reliance on lineage assignment based on prior clinical sequencing and publicly available sequences.

Conclusion

Sequencing of SARS-CoV-2 from wastewater largely corresponded with sequencing from clinical specimens. The prevalence of VOCs and lineages in clinical specimens was shown to be detectable in wastewater during the same times, which enabled us to provide comprehensive details on VOCs and lineages in the population. Despite inherent limitations of SARS-CoV-2 sampling in wastewater, we have generated a database spanning three SARS-CoV-2 waves, that document variant and lineage changes with time and geographical location and which correspond to clinically identified variants. We have illustrated how sequences not found in clinical specimens may be identified in populations through wastewater. Our heatmap has the potential to detect new variants prior to emergence in clinical samples and this may be particularly useful during times of low disease incidence between waves, when few numbers of positive clinical samples are collected and submitted for testing.

Figure legends

Figure 1 Signature mutation analysis of Variants of concern (VOC) and Variants of interest (VOI). The Y axis shows the read frequencies of each mutations, and X axis shows the epidemiological week when the samples were collected. The red line with dots shows the time-point analysed. Different shapes and colours of signature mutations were shown in the key. A-E: Gauteng province sites. E-G: KwaZulu-Natal province. H-I: Free State province. J-K: Western Cape province. L-O: Eastern Cape province. P: key for the different shapes and colours for each mutation and variants.

Figure 2 The proportion of variants and lineages by month from wastewater samples, from April 2021 to January 2022 using the Freyja tool. The X-axis shows the month and the number of samples sequenced. The Y axis shows the proportion of each variant present in the specific month. Only samples with sequence coverage of >50 were included. “Other” represents unconfirmed lineages found in the wastewater samples. **A.** SARS-CoV-2 variants. **B.** SARS-CoV-2 lineages

Figure 3 Heatmap of amino acid mutations distributed across the SARS-CoV2 spike protein in comparison with the Wuhan reference strain, arranged vertically in chronological order. Each row represents a sample, organized by the date of sample collection. Each column represents an amino acid position of the spike protein. Regions with no mutations or low occurrences are represented in grey and light yellow (0-34%). Regions with mutations that have a 50% read frequency are represented in bright yellow. Regions with mutations with a read frequency between 60-80% are represented in orange and very high occurring mutations (89-100%) are represented in red, as per the key.

Figure 4 Dot blot showing the uncommon mutations of SARS-CoV-2 detected in wastewater during the period April, 2021 – January, 2022 and their prevalence. The x-axis represents the uncommon mutations and the y-axis represents the times at which the mutations were first and last observed (as represented by continuous line between the dates). The size of the dot of each mutation describes the number of times the mutation was observed (prevalence) in the collected samples.

Declaration

Ethics approval and consent to participant

The study did not involve any human participants. An application for ethics waiver was made to the Human Research Ethics Committee of the University of the Witwatersrand and was approved (number R14/49).

Consent for publication

Not applicable

Availability of data and materials

Supplementary figures

Figure S1 Stacked graph illustrating the total number of sequence reads from wastewater samples (N = 325) between April, 2021 – January, 2022. Mapped reads are represented in blue and unmapped reads are represented in orange. A horizontal red line inserted at the mark one million reads. Samples had sequence reads of more than one million were to the left of the red vertical line and samples had

less than one million reads to the right of the red vertical line. The lower quartile value ($2.53x10^5$), median ($1.72x10^6$) and third quartile ($2.6x10^6$) values are represented in blue horizontal broken lines.

Figure S2. A graph showing the percentage coverage of SARS-CoV-2 in 10x depth of sequence reads. Y-axis is the percentage of coverage, and x-axis is the total number of samples. **A.** sequence coverage for whole genome. **B.** sequence coverage for spike protein.

Supplementary tables

Table S1, table S2,

Competing interests

All authors declare no competing interests.

Funding

We acknowledge the financial support from the National Institute for Communicable Diseases (NICD) of South Africa, the Water Research Commission (WRC) of South Africa, the German Society for International Cooperation (GIZ) and Bill and Melinda Gates foundation (BMGF), and Africa CDC.

Authors contributions

MY: co- conceptualized study, co- performed analysis, wrote, edit, and reviewed manuscript. SR: co- conceptualized study, edit, and reviewed manuscript. ST: co- performed analysis, edit, and reviewed manuscript. NN: co- performed analysis, edit, and reviewed manuscript. CI: edit, and reviewed manuscript. WH: edit, and reviewed manuscript, SM: edit, and reviewed manuscript, SG: edit, and reviewed manuscript, JIL: co- performed analysis, edit, and reviewed manuscript, KGA: edit, and reviewed manuscript, CS: edit, and reviewed manuscript, AvG: edit, and reviewed manuscript, NW: edit, and reviewed manuscript AI: edit, and reviewed manuscript, MS: co- conceptualized, edit, and reviewed manuscript. KM: co- conceptualized, edit, and reviewed manuscript.

Acknowledgements

The authors would like to thank the local government and wastewater treatment staff for sample collection and transport. We also thank the staff of the NICD Centre for Vaccines and Immunology and

the Centre for Respiratory Disease and Meningitis. special thanks to: Josie Everatt, Boitshoko Mahlangu, Anele Mnguni, Noxolo Ntuli, Gerald Motsatsi for their assistance in setting up and troubleshooting PCR testing, and ongoing supportive collaboration. We thank the team at Hyrax Biosciences for the use of their tool exatype. We would like to acknowledge the contribution from the SACCESS network. We also acknowledge the funding from the NICD, Bill and Melinda Gates Foundation, Water Research Commission of South Africa, and Africa CDC.

References

1. Peng, L. *et al.* SARS-CoV-2 can be detected in urine, blood, anal swabs, and oropharyngeal swabs specimens. *J. Med. Virol.* **92**, 1676–1680 (2020).

2. Fontenele, R. S. *et al.* High-throughput sequencing of SARS-CoV-2 in wastewater provides insights into circulating variants. *Water Res.* **205**, 117710 (2021).

3. The economist. How covid-19 spurred governments to snoop on sewage. (2022).

4. COVID19WBEC. COVID-19 WBE Collaborative. https://www.covid19wbec.org/ (2020).

5. Zhu, N. *et al.* A novel coronavirus from patients with pneumonia in China, 2019. *N. Engl. J. Med.* (2020).

6. Wang, W. *et al.* Detection of SARS-CoV-2 in Different Types of Clinical Specimens. *JAMA* **323**, 1843–1844 (2020).

7. Lara, R. W. I. *et al.* Monitoring SARS-CoV-2 circulation and diversity through community wastewater sequencing. *medRxiv* (2020).

8. Crits-Christoph, A. *et al.* Genome sequencing of sewage detects regionally prevalent SARS-CoV-2 variants. *MBio* **12**, e02703-20 (2021).

9. DWS. NATIONAL INTEGRATED WATER INFORMATION SYSTEM. https://www.dws.gov.za/niwis2?AspxAutoDetectCookieSupport=1 (2022).

10. STATS SA. Stats SA General Household Survey 2021. https://www.statssa.gov.za/?p=15482

(2021).

11. Johnson, R. *et al.* Qualitative and quantitative detection of SARS-CoV-2 RNA from untreated wastewater in the Western Cape, South Africa. *Methodol. Detect. SARS COV 2 RNA lN WASTEWATER SURVElLLANCE* 52 (2020).

12. NICD. WASTEWATER-BASED EPIDEMIOLOGY FOR SARS-COV-2 IN SOUTH AFRICA INCLUDING WASTEWATER GENOMICS. https://www.nicd.ac.za/diseases-a-z-index/disease-index-covid-19/surveillance-reports/weekly-reports/wastewater-based-epidemiology-for-sars-cov-2-in-south-africa/ (2022).

13. NICD. FIRST CASE OF COVID-19 CORONAVIRUS REPORTED IN SA. https://www.nicd.ac.za/first-case-of-covid-19-coronavirus-reported-in-sa/ (2020).

14. Tegally, H. *et al.* Emergence and rapid spread of a new severe acute respiratory syndrome-related coronavirus 2 (SARS-CoV-2) lineage with multiple spike mutations in South Africa. *MedRxiv* (2020).

15. ECDC. Brief, Threat Assessment: Emergence of SARS-CoV-2 B. 1.617 variants in India and situation in the EU/EEA. (2021).

16. Tegally, H. *et al.* Rapid replacement of the Beta variant by the Delta variant in South Africa. *MedRxiv* (2021).

17. Karim, S. S. A. & Karim, Q. A. Omicron SARS-CoV-2 variant: a new chapter in the COVID-19 pandemic. *Lancet* **398**, 2126–2128 (2021).

18. NICD. Network for Genomic Surveillance in South Africa (NGS-SA). SARS-CoV-2 Sequencing Update 19 August 2022. (2022).

19. Iwu-Jaja, C., Dlovu, KL., Said, R., Yousif, M., Taukobong, S., Macheka, M., Mhlanga, L., van Schalkwyk, C., Pulliam, J., Moultrie, T., Suchard, M., McCarthy, K. Cumulative evidence from the South African COVID Collaborative Environmental Surveillance System (SACCESS) supports reliance on wastewater-based epidemiological approaches for COVID

pandemic monitoring in a middle income African country. In preparation.

20. Ikner, L. A., Soto-Beltran, M. & Bright, K. R. New method using a positively charged microporous filter and ultrafiltration for concentration of viruses from tap water. *Appl. Environ. Microbiol.* **77**, 3500–3506 (2011).

21. Rachida, S. *Validating the COVIDSeq and Sinai protocols for wastewater-based sequencing of SARS-CoV-2. In preparation.*

22. Gonzalez-Reiche, A. S. *et al.* Introductions and early spread of SARS-CoV-2 in the New York City area. *Science (80-.).* **369**, 297–301 (2020).

23. Bhoyar, R. C. *et al.* High throughput detection and genetic epidemiology of SARS-CoV-2 using COVIDSeq next-generation sequencing. *PLoS One* **16**, e0247115 (2021).

24. Khailany, R. A., Safdar, M. & Ozaslan, M. Genomic characterization of a novel SARS-CoV-2. *Gene reports* **19**, 100682 (2020).

25. Afgan, E. *et al.* The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. *Nucleic Acids Res.* **44**, W3–W10 (2016).

26. Gangavarapu, K. *et al.* Outbreak. info genomic reports: scalable and dynamic surveillance of SARS-CoV-2 variants and mutations. *medRxiv* (2022).

27. Karthikeyan, S. *et al.* Wastewater sequencing reveals early cryptic SARS-CoV-2 variant transmission. *Nature* **609**, 101–108 (2022).

28. Lou, E. G. *et al.* Direct comparison of RT-ddPCR and targeted amplicon sequencing for SARS-CoV-2 mutation monitoring in wastewater. *Sci. Total Environ.* **833**, 155059 (2022).

29. Scheepers, C. *et al.* Emergence and phenotypic characterization of C. 1.2, a globally detected lineage that rapidly accumulated mutations of concern. (2021).

30. Karthikeyan, S. *et al.* Wastewater sequencing uncovers early, cryptic SARS-CoV-2 variant transmission. *medRxiv* (2021).

31. Laha, S. *et al.* Characterizations of SARS-CoV-2 mutational profile, spike protein stability and viral transmission. *Infect. Genet. Evol.* **85**, 104445 (2020).

32. Bate, N. *et al.* In vitro evolution predicts emerging SARS-CoV-2 mutations with high affinity for ACE2 and cross-species binding. *PLoS Pathog.* **18**, e1010733 (2022).

33. Smyth, D. S. *et al.* Tracking cryptic SARS-CoV-2 lineages detected in NYC wastewater. *Nat. Commun.* **13**, 1–9 (2022).

34. Yaniv, K. *et al.* Managing an evolving pandemic: Cryptic circulation of the Delta variant during the Omicron rise. *Sci. Total Environ.* **836**, 155599 (2022).

35. WHO. *Classification of Omicron (B.1.1.529): SARS-CoV-2 Variant of Concern*. https://www.who.int/news/item/26-11-2021-classification-of-omicron-(b.1.1.529)-sars-cov-2-variant-of-concern (2021).

433 Table 1. Illustration of amino acid variations in samples, produced by the Galaxy pipeline.

A	B	C	D	E	F	G
Sample	**QC filter**	**Number of reads**	**Mutation effect**	**GENE**	**Mutation**	**Mutation frequency**
21-0128	PASS	4073	NON_SYNONYMOUS_CODING	ORF1ab	E41K	92
21-0128	PASS	12106	NON_SYNONYMOUS_CODING	S	L1203F	96
21-0128	min_af_0.1Xmin_dp_5Xmin_dp_alt_10	12234	FRAME_SHIFT	S	V1228	1
21-0128	PASS	13569	NON_SYNONYMOUS_CODING	ORF3a	T190I	88
21-0128	PASS	1536	NON_SYNONYMOUS_CODING	E	P71L	97
21-0128	min_af_0.1Xmin_dp_5Xmin_dp_alt_10	1339	NON_SYNONYMOUS_CODING	M	I8S	1

434 **A** shows sample ID. **B** is the quality indicator. Filter value 'PASS' are samples with a minimum of 10
435 reads and read frequency greater than 10. **C** is the number of reads produced for each sample. **D** is
436 the effect of the mutation detected in the gene. **E** is the name of the gene where mutation occurred.
437 **F** is the mutation detected. **G** is the proportion of the mutant cells in a sample or population.

438

439 Table 2: List of signature mutations which was used to identify VOC and VOI present in wastewater
440 samples (Gangavarapu *et al.*, 2022)

441

Omicron	Alpha	Beta	Delta	C.1.2	Gamma	Lambda	Mu
G339D	A570D	D80A	T19R	P9L	T20N	G75V	Y144S
S371L	S982A		R145H	P25L	P26S	T76I	Y145N
S375F	D1118H		E156del	C136F	T1027I	D253N	
Q493R			R158G	Y449H		L452Q	
G496S			A222V			F490S	
Y505H							
T547K							
N856K							
Q954H							
N969K							
L981F							

442

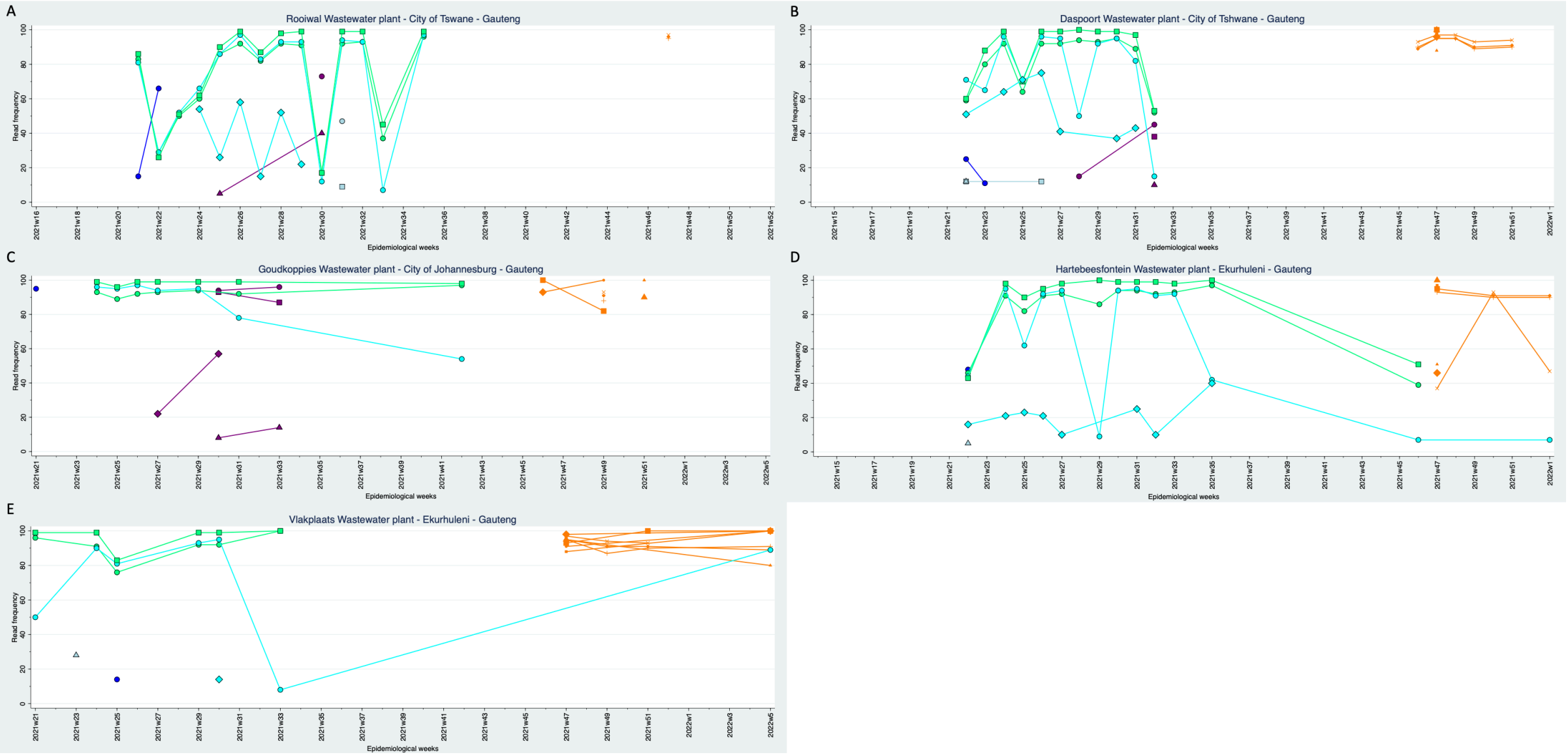
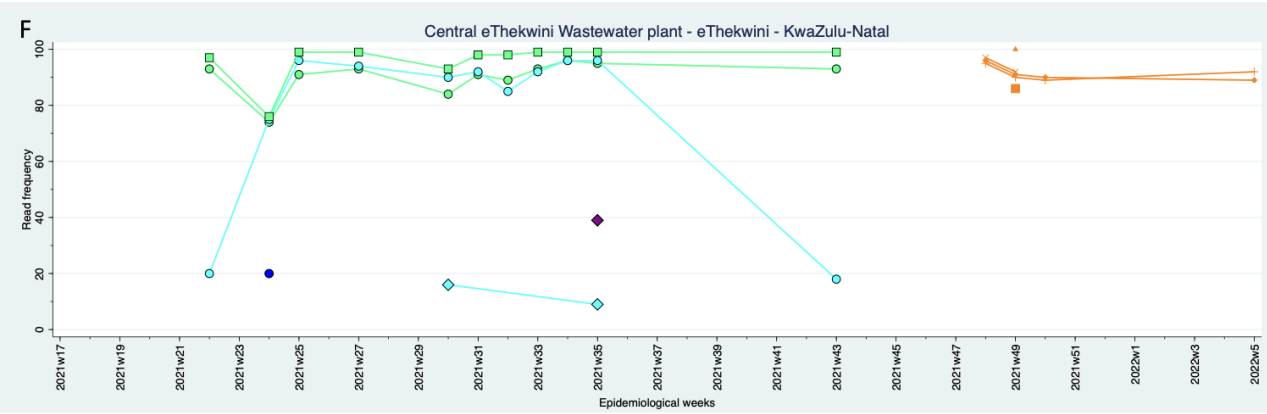

A
Rooiwal Wastewater plant - City of Tswane - Gauteng
B
Daspoort Wastewater plant - City of Tshwane - Gauteng
C
Goudkoppies Wastewater plant - City of Johannesburg - Gauteng
D
Hartebeesfontein Wastewater plant - Ekurhuleni - Gauteng
E
Vlakplaats Wastewater plant - Ekurhuleni - Gauteng
Read frequency
Epidemiological weeks

Figure 1 A-E


F
Central eThekwini Wastewater plant - eThekwini - KwaZulu-Natal
Read frequency
0
20
40
60
80
100
2021w17
2021w19
2021w21
2021w23
2021w25
2021w27
2021w29
2021w31
2021w33
2021w35
2021w37
2021w39
2021w41
2021w43
2021w45
2021w47
2021w49
2021w51
2022w1
2022w3
2022w5
Epidemiological weeks

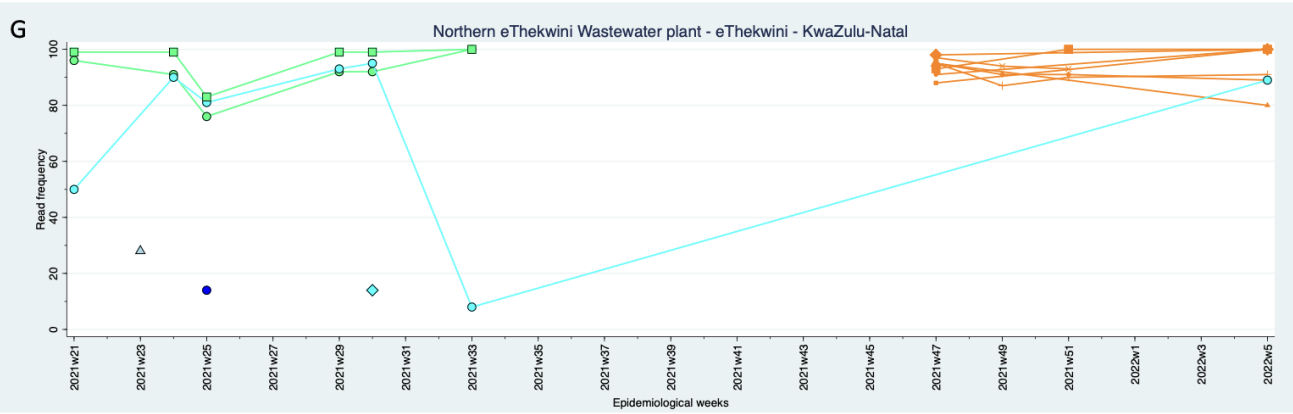

G
Northern eThekwini Wastewater plant - eThekwini - KwaZulu-Natal
Read frequency
0
20
40
60
80
100
2021w21
2021w23
2021w25
2021w27
2021w29
2021w31
2021w33
2021w35
2021w37
2021w39
2021w41
2021w43
2021w45
2021w47
2021w49
2021w51
2022w1
2022w3
2022w5
Epidemiological weeks

Figure 1 F-G

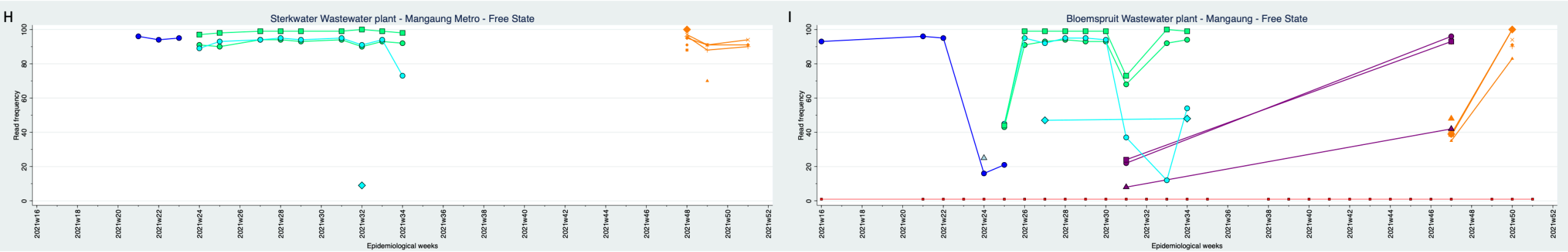
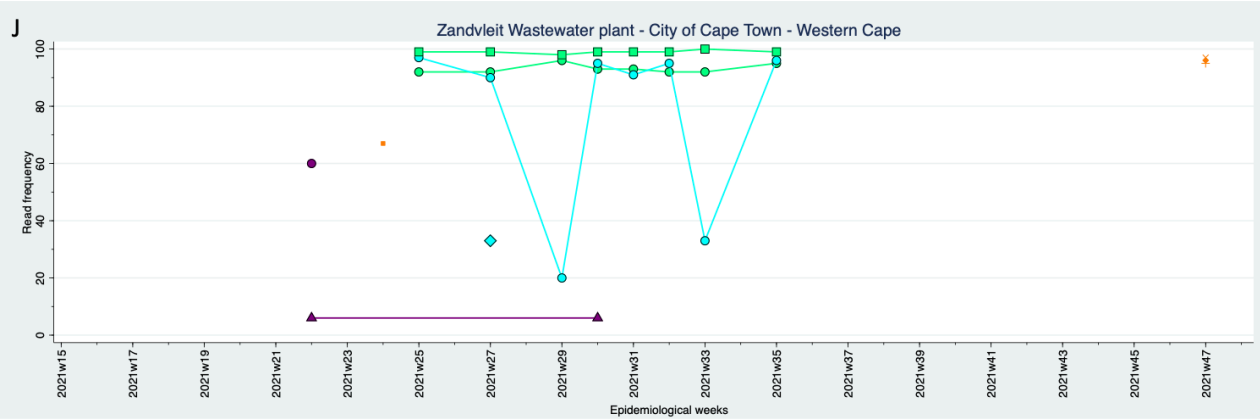

H
Sterkwater Wastewater plant - Mangaung Metro - Free State
Read frequency
0
20
40
60
80
100
2021w16
2021w18
2021w20
2021w22
2021w24
2021w26
2021w28
2021w30
2021w32
2021w34
2021w36
2021w38
2021w40
2021w42
2021w44
2021w46
2021w48
2021w50
2021w52
Epidemiological weeks
I
Bloemspruit Wastewater plant - Mangaung - Free State
Read frequency
0
20
40
60
80
100
2021w16
2021w18
2021w20
2021w22
2021w24
2021w26
2021w28
2021w30
2021w32
2021w34
2021w36
2021w38
2021w40
2021w42
2021w44
2021w46
2021w48
2021w50
2021w52
Epidemiological weeks

Figure 1 H-I


J
Zandvleit Wastewater plant - City of Cape Town - Western Cape
Read frequency
0
20
40
60
80
100
2021w15
2021w17
2021w19
2021w21
2021w23
2021w25
2021w27
2021w29
2021w31
2021w33
2021w35
2021w37
2021w39
2021w41
2021w43
2021w45
2021w47
Epidemiological weeks

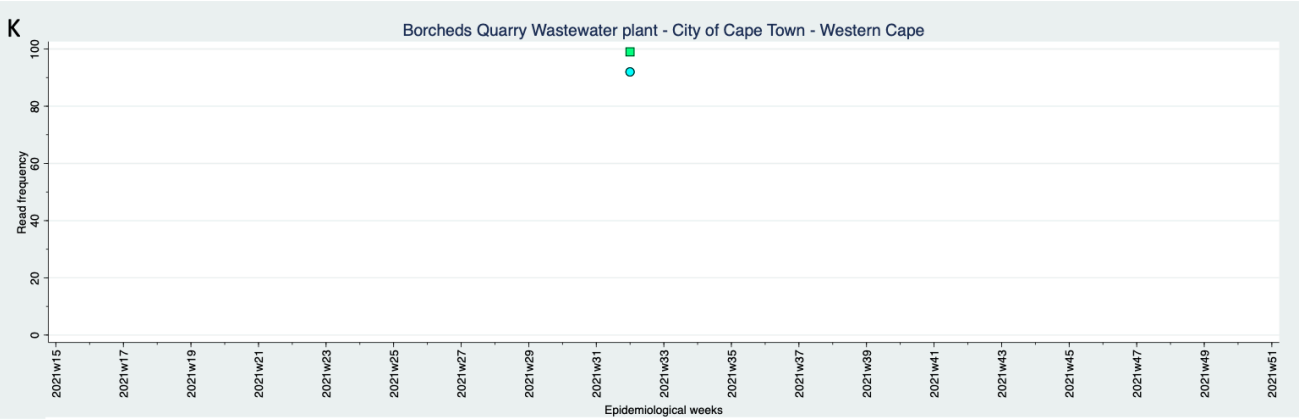

K
Borcheds Quarry Wastewater plant - City of Cape Town - Western Cape
Read frequency
0
20
40
60
80
100
2021w15
2021w17
2021w19
2021w21
2021w23
2021w25
2021w27
2021w29
2021w31
2021w33
2021w35
2021w37
2021w39
2021w41
2021w43
2021w45
2021w47
2021w49
2021w51
Epidemiological weeks

Figure 1 J-K

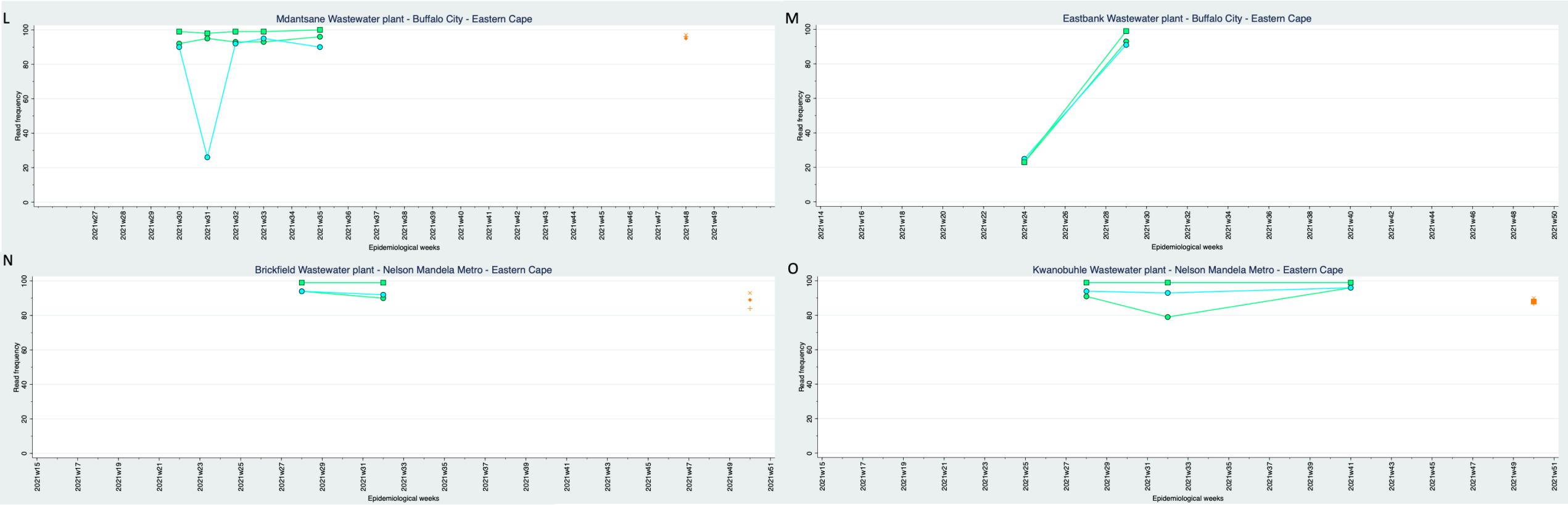

L
Mdantsane Wastewater plant - Buffalo City - Eastern Cape
Read frequency
Epidemiological weeks
M
Eastbank Wastewater plant - Buffalo City - Eastern Cape
Read frequency
Epidemiological weeks
N
Brickfield Wastewater plant - Nelson Mandela Metro - Eastern Cape
Read frequency
Epidemiological weeks
O
Kwanobuhle Wastewater plant - Nelson Mandela Metro - Eastern Cape
Read frequency
Epidemiological weeks

Figure 1 L-O


Om_G339D
Om_S371L
Om_S375F
Om_G496S
Om_Q493R
Om_Y505H
Om_T547K
Om_Q954H
Om_N969K
Om_L981F
C12_P9L
C12_P25L
C12_C136F
C12_Y449H
a_A570D
a_S982A
a_D1118H
b_D80A
d_T19R
dvar_G142D
dAY42_R145H
d_E156del
d_R158G
d_AY42_A222V
l_G75V
l_T76I
l_D253N
l_L452Q
l_F490S
m_Y144S
m_Y145N
sample collection

Figure 1 P

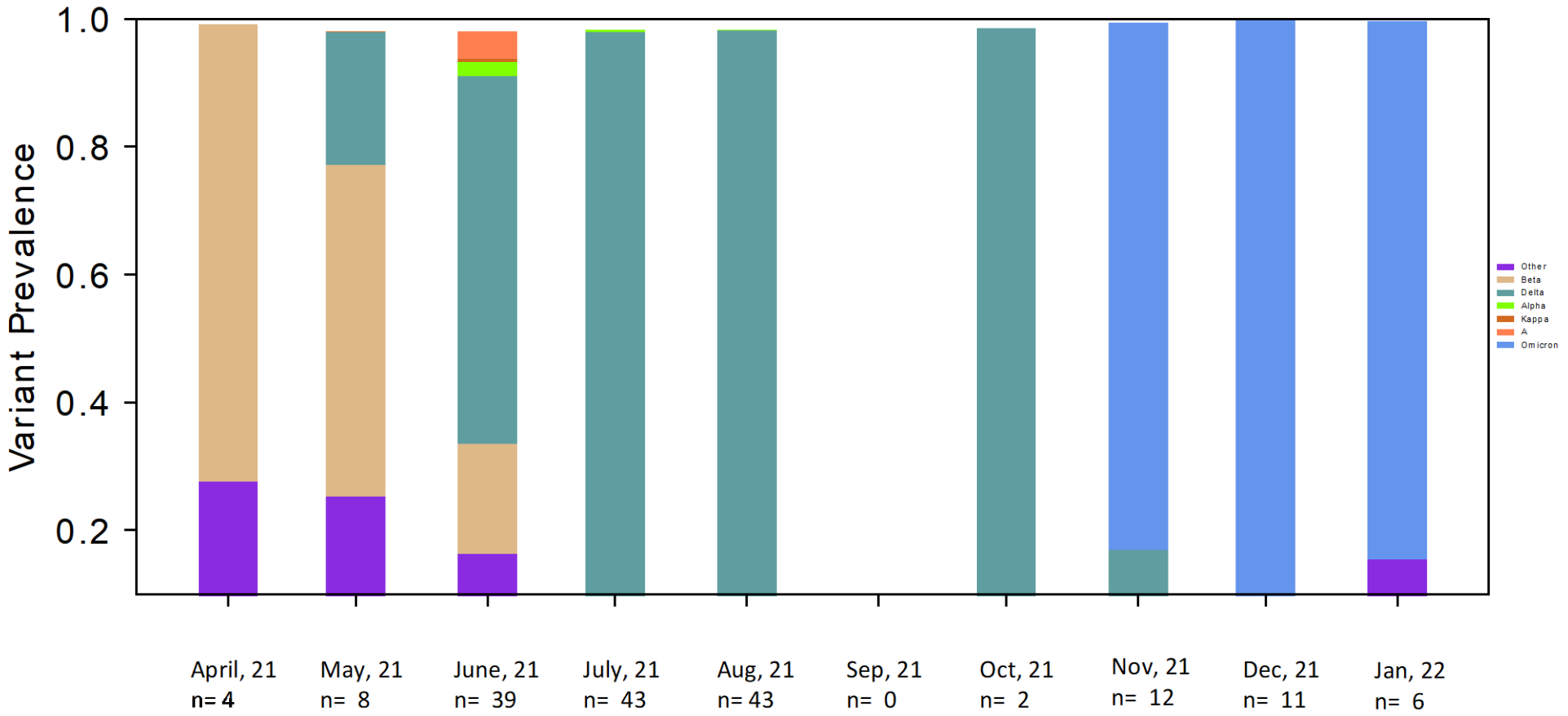

1.0
0.8
0.6
0.4
0.2
Variant Prevalence
Other
Beta
Delta
Alpha
Kappa
A
Omicron
April, 21
n= 4
May, 21
n= 8
June, 21
n= 39
July, 21
n= 43
Aug, 21
n= 43
Sep, 21
n= 0
Oct, 21
n= 2
Nov, 21
n= 12
Dec, 21
n= 11
Jan, 22
n= 6

Figure 2A

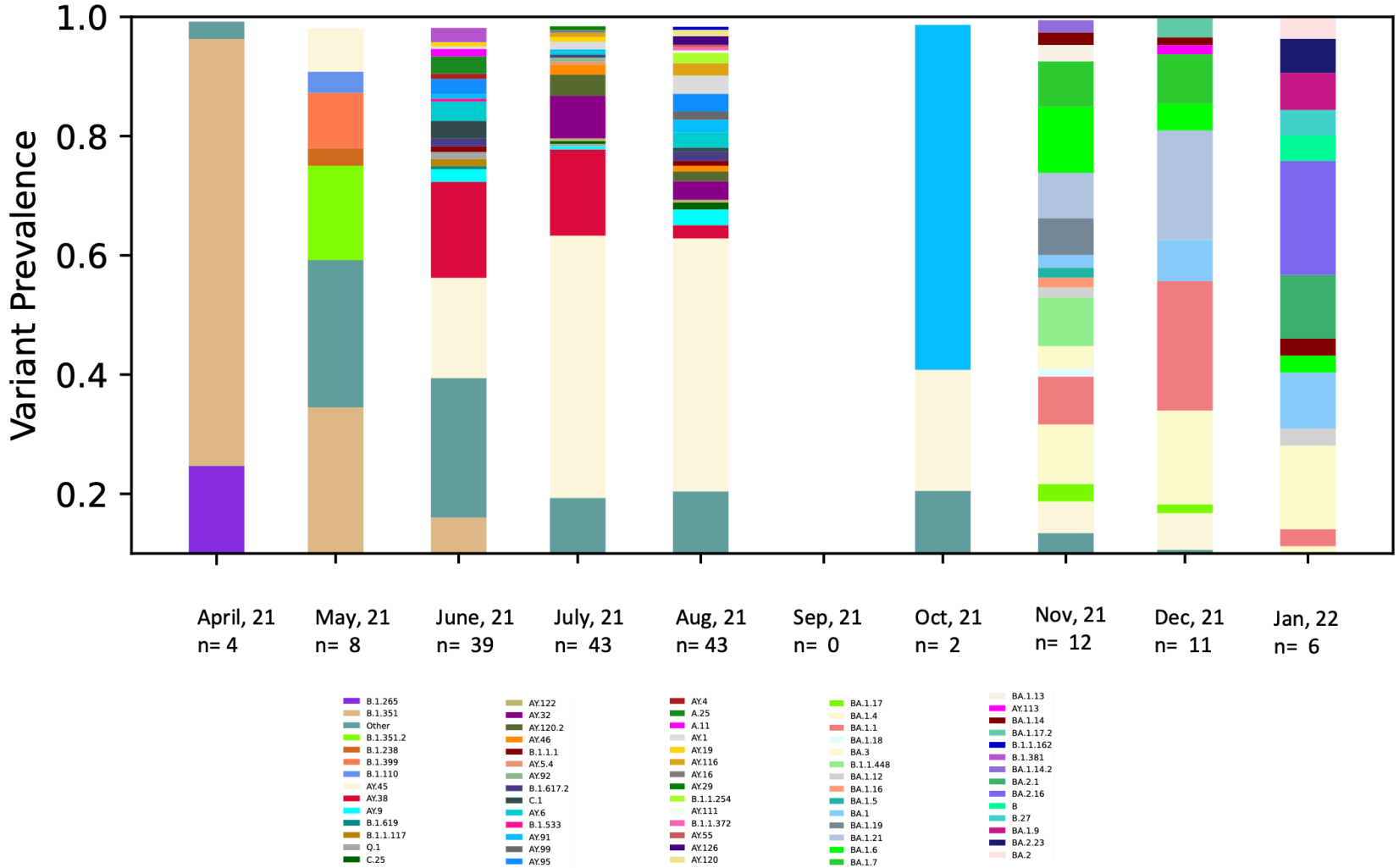

1.0
0.8
0.6
0.4
0.2
Variant Prevalence
April, 21
n= 4
May, 21
n= 8
June, 21
n= 39
July, 21
n= 43
Aug, 21
n= 43
Sep, 21
n= 0
Oct, 21
n= 2
Nov, 21
n= 12
Dec, 21
n= 11
Jan, 22
n= 6
B.1.265
B.1.351
Other
B.1.351.2
B.1.238
B.1.399
B.1.110
AY.45
AY.38
AY.9
B.1.619
B.1.1.117
Q.1
C.25
AY.122
AY.32
AY.120.2
AY.46
B.1.1.1
AY.5.4
AY.92
B.1.617.2
C.1
AY.6
B.1.533
AY.91
AY.99
AY.95
AY.4
A.25
A.11
AY.1
AY.19
AY.116
AY.16
AY.29
B.1.1.254
AY.111
B.1.1.372
AY.55
AY.126
AY.120
BA.1.17
BA.1.4
BA.1.1
BA.1.18
BA.3
B.1.1.448
BA.1.12
BA.1.16
BA.1.5
BA.1
BA.1.19
BA.1.21
BA.1.6
BA.1.7
BA.1.13
AY.113
BA.1.14
BA.1.17.2
B.1.1.162
B.1.381
BA.1.14.2
BA.2.1
BA.2.16
B
B.27
BA.1.9
BA.2.23
BA.2

Figure 2B

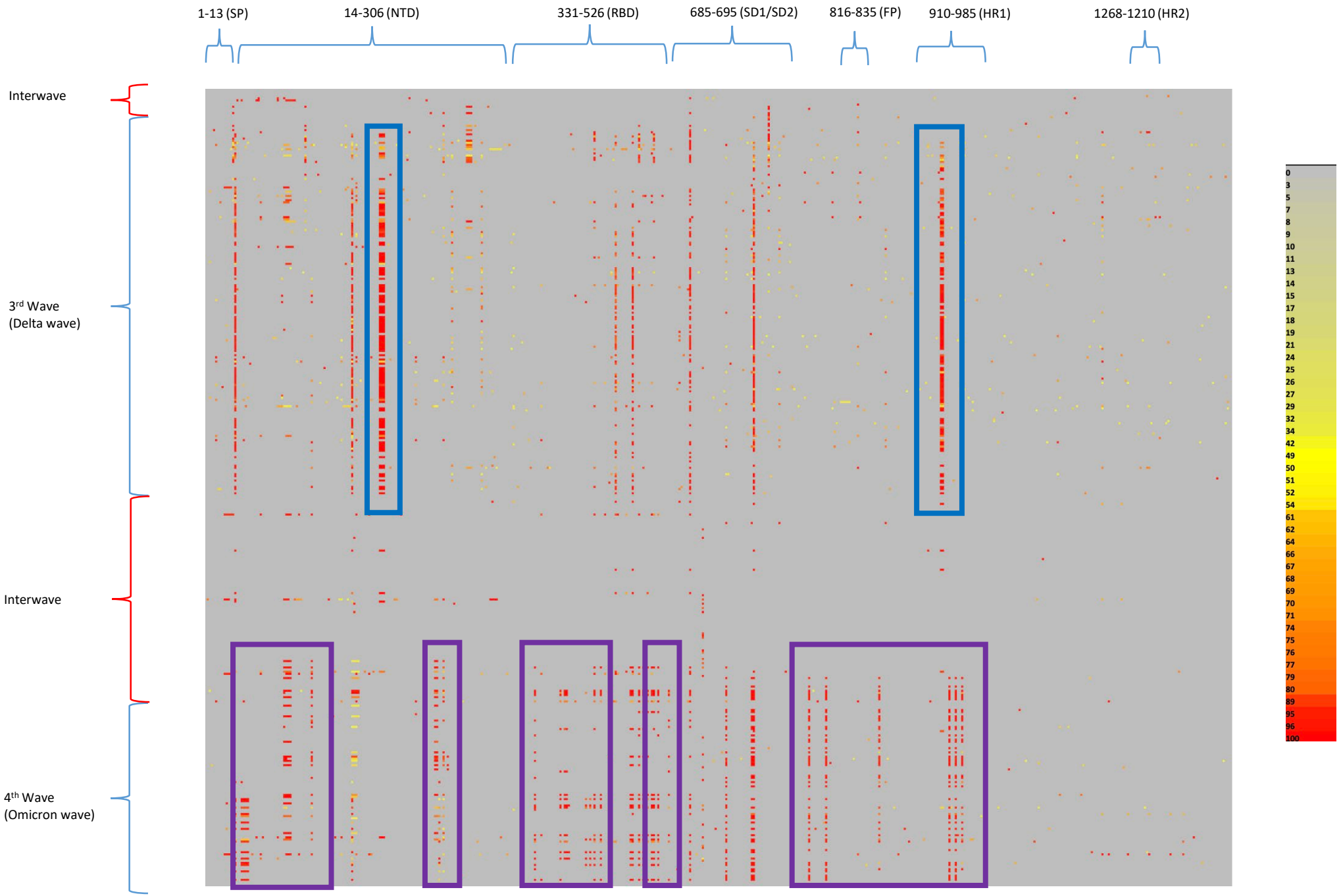
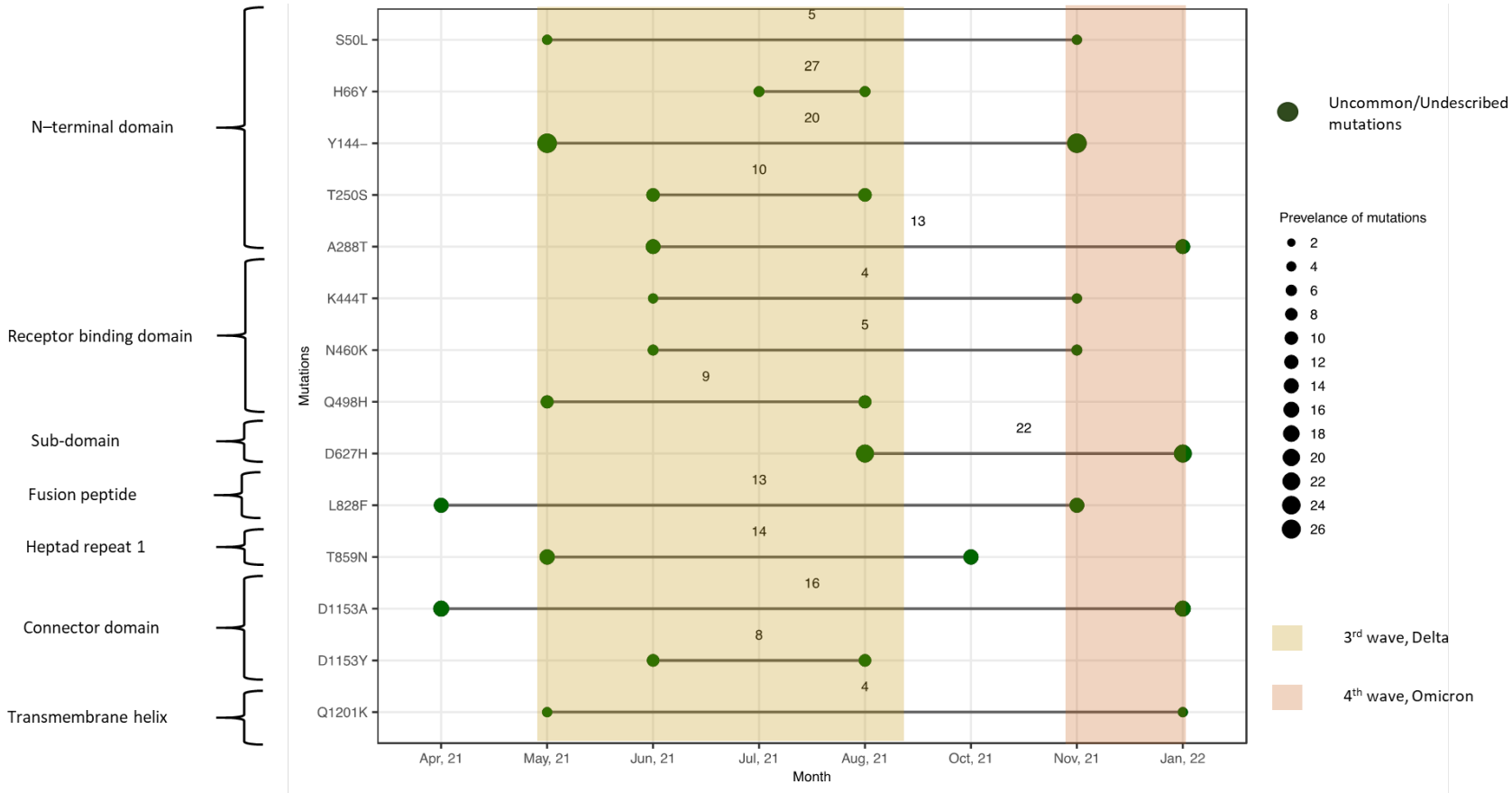

1-13 (SP)
14-306 (NTD)
331-526 (RBD)
685-695 (SD1/SD2)
816-835 (FP)
910-985 (HR1)
1268-1210 (HR2)
Interwave
3rd Wave
(Delta wave)
Interwave
4th Wave
(Omicron wave)
0
3
5
7
8
9
10
11
13
14
15
17
18
19
21
24
25
26
27
29
32
34
42
49
50
51
52
54
61
62
64
66
67
68
69
70
71
74
75
76
77
79
80
89
95
96
100

Figure 3



Mutation	Number of samples detected in Wastewater	Prevalence in wastewater South Africa (N = 325)
S50L	5	1,54
H66Y	27	8,31
Y144-	20	6,15
T250S	10	3,08
A288T	13	4,00
K444T	4	1,23
N460K	5	1,54
Q498H	9	2,77
D627H	22	6,77
L828F	13	4,00
T859N	14	4,31
D1153A	16	4,92
D1153Y	8	2,46
Q1201	4	1,23

Figure 4