

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

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

7 <sup>1</sup>Centre for Vaccines and Immunology, National Institute for Communicable Diseases, a division of the  
8 National Health Laboratory Service, South Africa

9 <sup>2</sup>Department of Virology, School of Pathology, Faculty of Health Sciences, University of the  
10 Witwatersrand, Johannesburg

11 <sup>3</sup>Department of Immunology and Microbiology, The Scripps Research Institute, La Jolla, CA 92037,  
12 USA

13 <sup>4</sup>SAMRC Antibody Immunity Research Unit, Faculty of Health Sciences, University of the  
14 Witwatersrand, Johannesburg, South Africa.

15 <sup>5</sup>Centre for Respiratory Diseases and Meningitis, National Institute for Communicable Diseases, a  
16 division of the National Health Laboratory Service, South Africa

17 <sup>6</sup>School of Pathology, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg,  
18 South Africa.

19 <sup>7</sup>Sequencing Core Facility, National Institute for Communicable Diseases, a division of the National  
20 Health Laboratory Service, South Africa.

21 <sup>8</sup>Department of Biochemistry and Microbiology, Faculty of Science, Engineering and Agriculture,  
22 University of Venda, Thohoyandou, South Africa.

23 <sup>9</sup>Department of Chemical Pathology, School of Pathology, University of the Witwatersrand,  
24 Johannesburg.

25 <sup>10</sup>School of Public Health, University of the Witwatersrand, Johannesburg

26

**NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.**

27 **# Corresponding author**

28 Dr Mukhlid Yousif

29 Centre for Vaccines and Immunology

30 National Institute for Communicable Diseases

31 Researcher, Virology department

32 Faculty of Health Sciences

33 University of the Witwatersrand

34 Tel: +27 11 386 6461

35 Email: [mukhlidy@nicd.ac.za](mailto:mukhlidy@nicd.ac.za), [Mukhlid.yousif@gmail.com](mailto:Mukhlid.yousif@gmail.com)

36

37 **Keywords:** environmental, wastewater, mutations, sequence, whole genome sequencing, SARS-CoV-  
38 2, COVID-19, surveillance, minor, spike, alpha, beta, delta, omicron, C1.2

39

40 **Abstract**

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

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

## 60 **Background**

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

74 Wastewater genomic surveillance offers an opportunity to monitor circulating variants present in the  
75 community. Whole genome sequencing <sup>5</sup>, and other methods such as real-time PCR <sup>6</sup> enable detection  
76 and characterisation of SARS-CoV-2 variants, and have now been applied to wastewater samples.

77 Recent work has shown the potential for recovery of complete virus genomes from wastewater<sup>7</sup>,  
78 shown comparable results of wastewater and clinical surveillance, and identified novel mutations and  
79 lineages before appearance in clinical samples <sup>8 2</sup>. To date, wastewater sequencing of SARS-CoV-2  
80 has not been widely applied in low or middle income countries.

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

86 COVID-19 Environmental Surveillance System (SACCESS) arose to monitor trends in SARS-CoV-2  
87 levels in wastewater <sup>12</sup>.

88 Most laboratory testing is provided to the public through an extensive network of laboratories  
89 including the National Health Laboratory Service (NHLS) that covers over 80% of the population.  
90 South Africa identified its first case of COVID-19 on the 5<sup>th</sup> of March 2020, <sup>13</sup> and four waves of  
91 COVID-19 occurred within the first 24 months of virus introduction into the country. Following the  
92 initial SARS-CoV-2 wave , the Beta variant <sup>14</sup> was discovered and was predominant from November  
93 2020 to February 2021 (second wave). The third wave (May to September 2021) was characterized by  
94 the dominance of the Delta variant <sup>1516</sup> and the fourth wave (November 2021 to January 2022) by the  
95 Omicron BA.1 variant <sup>17</sup>. The National Institute for Communicable Diseases (NICD), a division of the  
96 NHLS, provides SARS-CoV-2 epidemiological surveillance data through collation of SARS-CoV-2  
97 PCR results from public and private laboratories. The Network for Genomics Surveillance in South  
98 Africa (NGS-SA) monitors the epidemiology of SARS-CoV-2 variants in PCR-confirmed cases in  
99 South Africa and reports weekly on findings <sup>18</sup>. The NGS-SA provided the first global reports of the  
100 emergence of Beta and Omicron variants of concern (VOC) <sup>1417</sup>.

101 Here, we show that wastewater can be used to effectively characterize SARS-CoV-2 virus dynamics in  
102 the population and identify variants of concern present using samples from sentinel wastewater  
103 treatment plants in urban metros collected from April, 2021 to the end of the fourth wave in January,  
104 2022, demonstrating the utility of wastewater genomic surveillance to complement clinical surveillance  
105 efforts in a middle income setting. We identify the potential strengths and limitations of genomic  
106 wastewater surveillance for SARS-CoV-2 in the South African context.

## 107 **Methods**

### 108 Wastewater sites

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

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

#### 115 Sample collection, RNA extraction, amplification and sequencing

116 One liter of grab sewage samples were collected and transported to NICD at 4°C. Viruses were  
117 concentrated from the sample by ultrafiltration <sup>20</sup>, and RNA was extracted using the QIAamp Viral  
118 RNA kit (Qiagen, GmbH, Germany). SARS-CoV-2 was detected by RT-PCR using the Allplex™  
119 2019-nCoV Assay from Seegene (Seoul, Korea). RNA was re-extracted from SARS-CoV-2 positive  
120 concentrates and subjected to amplicon-based whole genome sequencing using the Sinai protocol with  
121 some modifications as described <sup>21 22</sup>. Paired-end libraries were prepared using Illumina COVIDSeq  
122 Kit as previously described <sup>23</sup> followed by sequencing (2x 150 pb) on NextSeq 1000/2000 platform  
123 (Illumina Inc, USA).

#### 124 Sequence analysis

##### 125 Quality control checks

126 FastQ files were trimmed, filtered based on sequence quality, assembled and mapped to the reference  
127 genome (NC\_045512.2) according to published criteria <sup>24</sup> using Exatype web-based bioinformatics tool  
128 (<https://sars-cov-2.exatype.com/>). Quality control indicators such as number of reads, number of  
129 mapped reads, and sequence coverage were recorded. Samples that passed the internal quality control  
130 were processed for mutation analysis using ARTIC protocol ([https://artic.network/ncov-](https://artic.network/ncov-2019/ncov2019-bioinformatics-sop.html)  
131 [2019/ncov2019-bioinformatics-sop.html](https://artic.network/ncov-2019/ncov2019-bioinformatics-sop.html)) in Galaxy (<https://usegalaxy.eu/> ) <sup>25</sup>. Again, reads were  
132 trimmed and filtered, assembled and mapped. At least 10 reads were required at each nucleotide position  
133 for downstream analysis. Amino acid mutations present at 5% of reads or less were removed from the  
134 analysis. Table 1 illustrates an example of amino acid variation analysis output.

##### 135 Sequence analysis using amino acid mutations

##### 136 Amino acids variation analysis

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

#### 152 Heatmap and dot blot

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

#### 158 Analysis using the Freyja tool

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

164 **Results**

165 Quality control

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

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

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

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

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

190



191 *Characterization of amino acid mutations in the spike region*

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

203 **Discussion:**

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

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

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

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

226 Regarding detection of previously described variants in our wastewater samples, both approaches we  
227 used (mutational analysis and the Freyja tool) successfully identified lineages in wastewater that  
228 corresponded to lineages identified in clinical specimens <sup>18</sup>. The Beta variant (first described in  
229 clinical specimens in South Africa in December 2020 <sup>14</sup>) was consistently observed from the start of  
230 wastewater surveillance until the end of the third wave in epidemiological week 19 (May 2021).  
231 Similarly, the Delta variant was first seen in clinical specimens during epidemiological week 21 of  
232 2021 <sup>16</sup>, and was first detected in wastewater samples the same week. Lineage C.1.2, described first  
233 in South Africa <sup>29</sup> was successfully detected in sequences from wastewater samples during week 22 to  
234 45 in 2021 whilst clinical detections of this lineage appeared from week 16 to 46. In epidemiological  
235 week 46 in 2021, the Omicron variant was identified in clinical samples whilst sequences from  
236 wastewater samples also identified mutations specific to Omicron in the same week. The Freyja tool <sup>30</sup>  
237 complemented our mutational profile analysis, identifying the proportions of each variant/lineage in  
238 wastewater at specific time points. Results from Freyja comparable to the prevalence of VOCs of  
239 SARS-CoV-2 reported from clinical specimens and additionally indicate the presence of lineages in  
240 our wastewater samples that were absent amongst sequences from clinical cases. This most likely  
241 arises through the sampling bias inherent in clinical surveillance, in which only symptomatic patients  
242 are tested and of whom only a fraction was sequenced.

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

251 Our mutational analysis identified multiple instances of rare mutations in the population (Figure 4).  
252 These mutations were found at a prevalence of >1.0% in wastewater samples, but were detected at  
253 <0.001% in clinical cases based on the data from GISAID. Although they were uncommon, some  
254 mutations were reported previously. Mutation S50L was found to be associated with reduced protein  
255 stability<sup>31</sup>. Mutation Q498H has reportedly caused increased binding affinity of RBD to ACE2<sup>32</sup>. The  
256 presence of uncommon mutations could be explained as “cryptic lineages” as described by Smyth and  
257 colleagues<sup>33</sup>. In their paper they attributed the cryptic lineages to either un-sampled cases or spillover  
258 of SARS-CoV-2 from an unidentified animal reservoir. Other wastewater studies have demonstrated  
259 cryptic variants, for example Yaniv and colleagues described a cryptic Delta variant in wastewater  
260 samples<sup>34</sup>. Unusual substitutions /deletions are useful tools that support epidemiological tracking of  
261 variants and may support hypothesis generation regarding the origins of SARS-CoV-2.

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

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

## 272 **Conclusion**

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

## 283 **Figure legends**

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

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

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

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

#### 308 **Declaration**

#### 309 **Ethics approval and consent to participant**

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

#### 313 **Consent for publication**

314 Not applicable

#### 315 **Availability of data and materials**

#### 316 **Supplementary figures**

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

321 less than one million reads to the right of the red vertical line. The lower quartile value ( $2.53 \times 10^5$ ),  
322 median ( $1.72 \times 10^6$ ) and third quartile ( $2.6 \times 10^6$ ) values are represented in blue horizontal broken lines.

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

## 326 **Supplementary tables**

327 **Table S1, table S2,**

## 328 **Competing interests**

329 All authors declare no competing interests.

## 330 **Funding**

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

## 335 **Authors contributions**

336 MY: co- conceptualized study, co- performed analysis, wrote, edit, and reviewed manuscript. SR: co-  
337 conceptualized study, edit, and reviewed manuscript. ST: co- performed analysis, edit, and reviewed  
338 manuscript. NN: co- performed analysis, edit, and reviewed manuscript. CI: edit, and reviewed  
339 manuscript. WH: edit, and reviewed manuscript, SM: edit, and reviewed manuscript, SG: edit, and  
340 reviewed manuscript, JIL: co- performed analysis, edit, and reviewed manuscript, KGA: edit, and  
341 reviewed manuscript, CS: edit, and reviewed manuscript, AvG: edit, and reviewed manuscript, NW:  
342 edit, and reviewed manuscript AI: edit, and reviewed manuscript, MS: co- conceptualized, edit, and  
343 reviewed manuscript. KM: co- conceptualized, edit, and reviewed manuscript.

## 344 **Acknowledgements**

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

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

### 353 **References**

- 354 1. Peng, L. *et al.* SARS-CoV-2 can be detected in urine, blood, anal swabs, and oropharyngeal  
355 swabs specimens. *J. Med. Virol.* **92**, 1676–1680 (2020).
- 356 2. Fontenele, R. S. *et al.* High-throughput sequencing of SARS-CoV-2 in wastewater provides  
357 insights into circulating variants. *Water Res.* **205**, 117710 (2021).
- 358 3. The economist. How covid-19 spurred governments to snoop on sewage. (2022).
- 359 4. COVID19WBEC. COVID-19 WBE Collaborative. <https://www.covid19wbec.org/> (2020).
- 360 5. Zhu, N. *et al.* A novel coronavirus from patients with pneumonia in China, 2019. *N. Engl. J.*  
361 *Med.* (2020).
- 362 6. Wang, W. *et al.* Detection of SARS-CoV-2 in Different Types of Clinical Specimens. *JAMA*  
363 **323**, 1843–1844 (2020).
- 364 7. Lara, R. W. I. *et al.* Monitoring SARS-CoV-2 circulation and diversity through community  
365 wastewater sequencing. *medRxiv* (2020).
- 366 8. Crits-Christoph, A. *et al.* Genome sequencing of sewage detects regionally prevalent SARS-  
367 CoV-2 variants. *MBio* **12**, e02703-20 (2021).
- 368 9. DWS. NATIONAL INTEGRATED WATER INFORMATION SYSTEM.  
369 <https://www.dws.gov.za/niwis2?AspxAutoDetectCookieSupport=1> (2022).
- 370 10. STATS SA. Stats SA General Household Survey 2021. <https://www.statssa.gov.za/?p=15482>

- 371 (2021).
- 372 11. Johnson, R. *et al.* Qualitative and quantitative detection of SARS-CoV-2 RNA from untreated  
373 wastewater in the Western Cape, South Africa. *Methodol. Detect. SARS COV 2 RNA IN*  
374 *WASTEWATER SURVEILLANCE* 52 (2020).
- 375 12. NICD. WASTEWATER-BASED EPIDEMIOLOGY FOR SARS-COV-2 IN SOUTH  
376 AFRICA INCLUDING WASTEWATER GENOMICS. [https://www.nicd.ac.za/diseases-a-z-](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/)  
377 [index/disease-index-covid-19/surveillance-reports/weekly-reports/wastewater-based-](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/)  
378 [epidemiology-for-sars-cov-2-in-south-africa/](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).
- 379 13. NICD. FIRST CASE OF COVID-19 CORONAVIRUS REPORTED IN SA.  
380 <https://www.nicd.ac.za/first-case-of-covid-19-coronavirus-reported-in-sa/> (2020).
- 381 14. Tegally, H. *et al.* Emergence and rapid spread of a new severe acute respiratory syndrome-  
382 related coronavirus 2 (SARS-CoV-2) lineage with multiple spike mutations in South Africa.  
383 *MedRxiv* (2020).
- 384 15. ECDC. Brief, Threat Assessment: Emergence of SARS-CoV-2 B. 1.617 variants in India and  
385 situation in the EU/EEA. (2021).
- 386 16. Tegally, H. *et al.* Rapid replacement of the Beta variant by the Delta variant in South Africa.  
387 *MedRxiv* (2021).
- 388 17. Karim, S. S. A. & Karim, Q. A. Omicron SARS-CoV-2 variant: a new chapter in the COVID-  
389 19 pandemic. *Lancet* **398**, 2126–2128 (2021).
- 390 18. NICD. Network for Genomic Surveillance in South Africa (NGS-SA). SARS-CoV-2  
391 Sequencing Update 19 August 2022. (2022).
- 392 19. Iwu-Jaja, C., Dlovu, KL., Said, R., Yousif, M., Taukobong, S., Macheke, M., Mhlanga, L., van  
393 Schalkwyk, C., Pulliam, J., Moultrie, T., Suchard, M., McCarthy, K. Cumulative evidence  
394 from the South African COVID Collaborative Environmental Surveillance System  
395 (SACCESS) supports reliance on wastewater-based epidemiological approaches for COVID



- 396 pandemic monitoring in a middle income African country. In preparation.
- 397 20. Ikner, L. A., Soto-Beltran, M. & Bright, K. R. New method using a positively charged  
398 microporous filter and ultrafiltration for concentration of viruses from tap water. *Appl.*  
399 *Environ. Microbiol.* **77**, 3500–3506 (2011).
- 400 21. Rachida, S. *Validating the COVIDSeq and Sinai protocols for wastewater-based sequencing of*  
401 *SARS-CoV-2. In preparation.*
- 402 22. Gonzalez-Reiche, A. S. *et al.* Introductions and early spread of SARS-CoV-2 in the New York  
403 City area. *Science (80-. )*. **369**, 297–301 (2020).
- 404 23. Bhojar, R. C. *et al.* High throughput detection and genetic epidemiology of SARS-CoV-2  
405 using COVIDSeq next-generation sequencing. *PLoS One* **16**, e0247115 (2021).
- 406 24. Khailany, R. A., Safdar, M. & Ozaslan, M. Genomic characterization of a novel SARS-CoV-2.  
407 *Gene reports* **19**, 100682 (2020).
- 408 25. Afgan, E. *et al.* The Galaxy platform for accessible, reproducible and collaborative biomedical  
409 analyses: 2016 update. *Nucleic Acids Res.* **44**, W3–W10 (2016).
- 410 26. Gangavarapu, K. *et al.* Outbreak. info genomic reports: scalable and dynamic surveillance of  
411 SARS-CoV-2 variants and mutations. *medRxiv* (2022).
- 412 27. Karthikeyan, S. *et al.* Wastewater sequencing reveals early cryptic SARS-CoV-2 variant  
413 transmission. *Nature* **609**, 101–108 (2022).
- 414 28. Lou, E. G. *et al.* Direct comparison of RT-ddPCR and targeted amplicon sequencing for  
415 SARS-CoV-2 mutation monitoring in wastewater. *Sci. Total Environ.* **833**, 155059 (2022).
- 416 29. Scheepers, C. *et al.* Emergence and phenotypic characterization of C. 1.2, a globally detected  
417 lineage that rapidly accumulated mutations of concern. (2021).
- 418 30. Karthikeyan, S. *et al.* Wastewater sequencing uncovers early, cryptic SARS-CoV-2 variant  
419 transmission. *medRxiv* (2021).

- 420 31. Laha, S. *et al.* Characterizations of SARS-CoV-2 mutational profile, spike protein stability and  
421 viral transmission. *Infect. Genet. Evol.* **85**, 104445 (2020).
- 422 32. Bate, N. *et al.* In vitro evolution predicts emerging SARS-CoV-2 mutations with high affinity  
423 for ACE2 and cross-species binding. *PLoS Pathog.* **18**, e1010733 (2022).
- 424 33. Smyth, D. S. *et al.* Tracking cryptic SARS-CoV-2 lineages detected in NYC wastewater. *Nat.*  
425 *Commun.* **13**, 1–9 (2022).
- 426 34. Yaniv, K. *et al.* Managing an evolving pandemic: Cryptic circulation of the Delta variant  
427 during the Omicron rise. *Sci. Total Environ.* **836**, 155599 (2022).
- 428 35. WHO. *Classification of Omicron (B.1.1.529): SARS-CoV-2 Variant of Concern.*  
429 [https://www.who.int/news/item/26-11-2021-classification-of-omicron-\(b.1.1.529\)-sars-cov-2-](https://www.who.int/news/item/26-11-2021-classification-of-omicron-(b.1.1.529)-sars-cov-2-variant-of-concern)  
430 [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).

431

432

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

A	B	C	D	E	F	G
Sampl e	QC filter	Numb er of reads	Mutation effect	GENE	Mutatio n	Mutatio n frequen cy
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

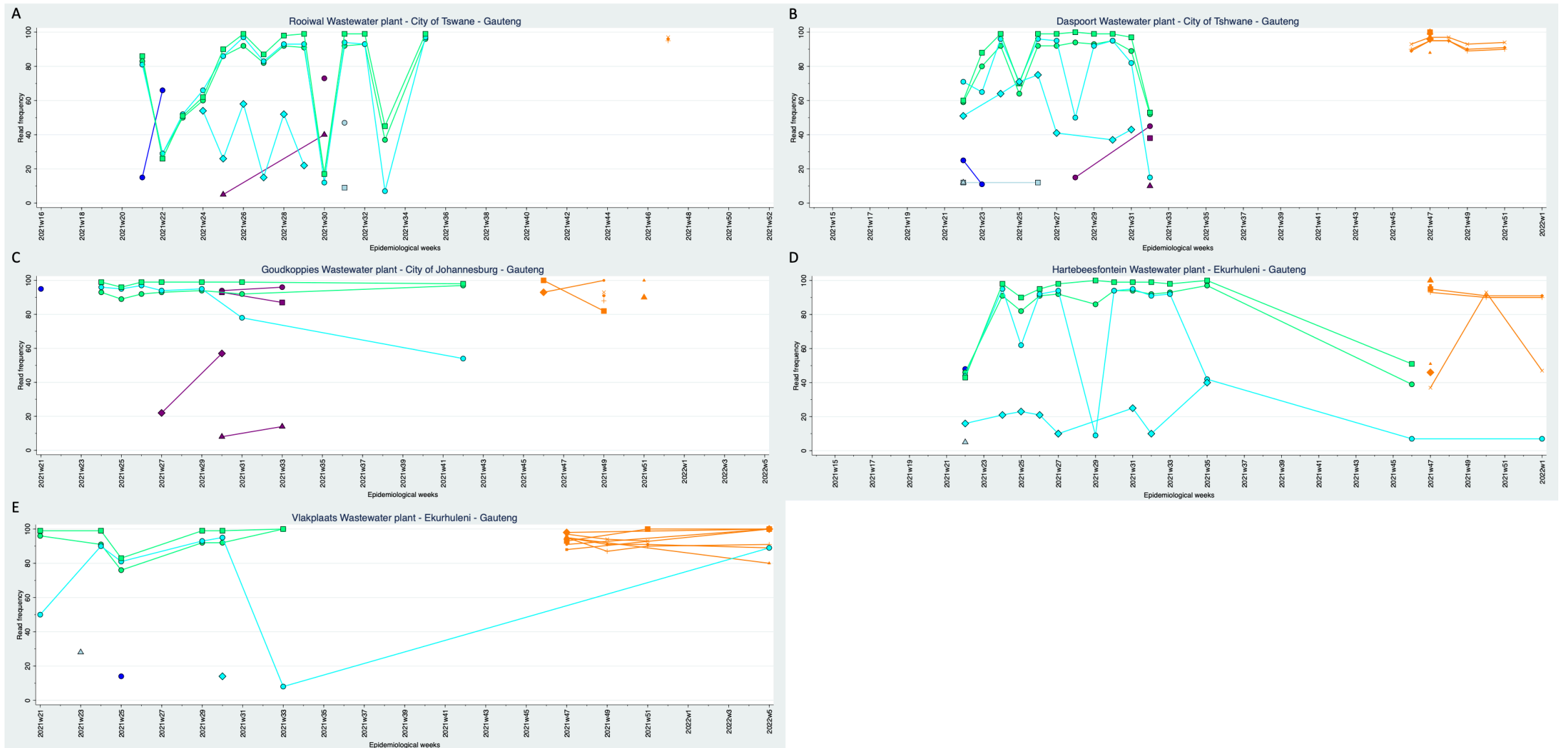


Figure 1 A-E

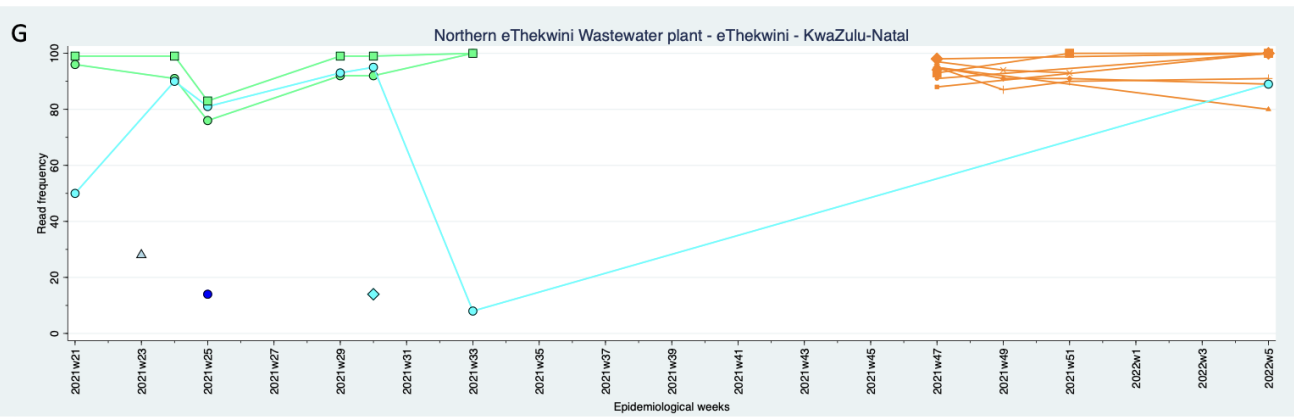
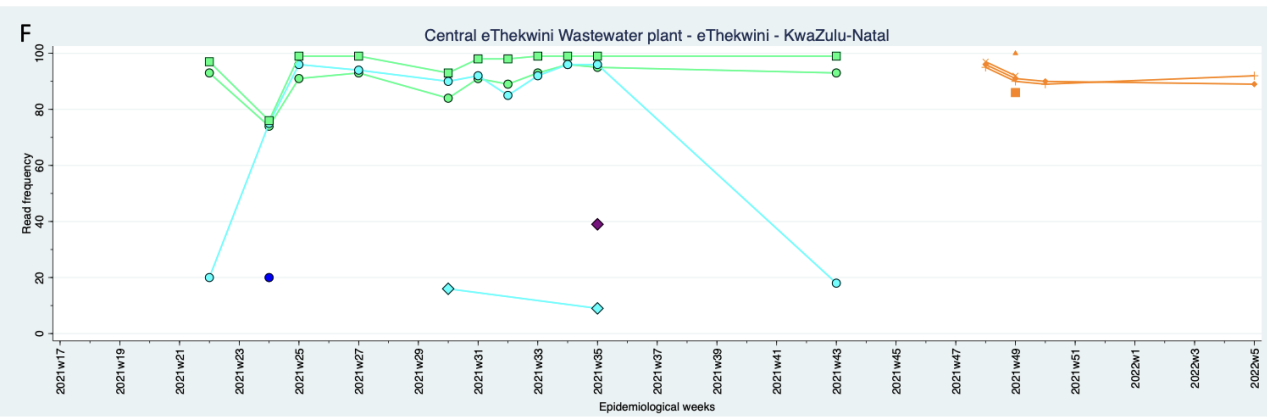


Figure 1 F-G

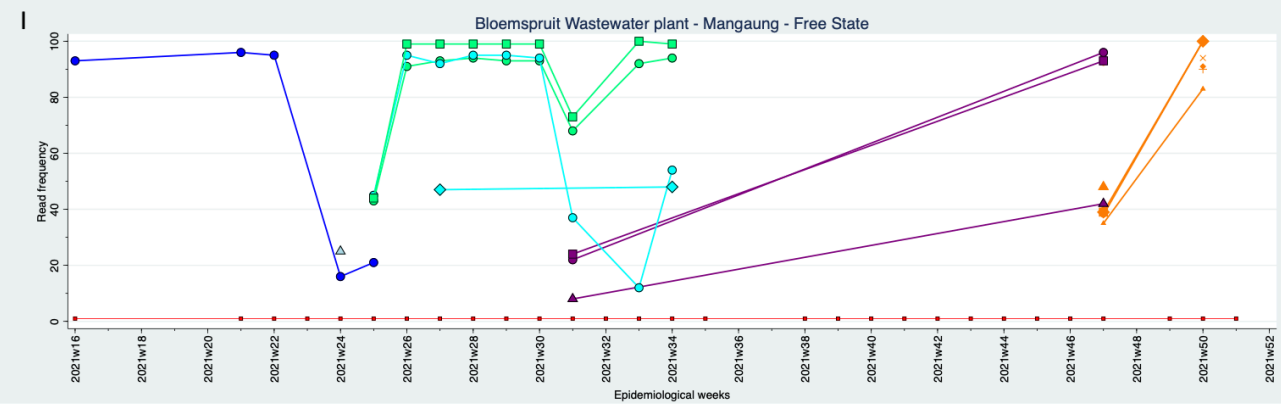
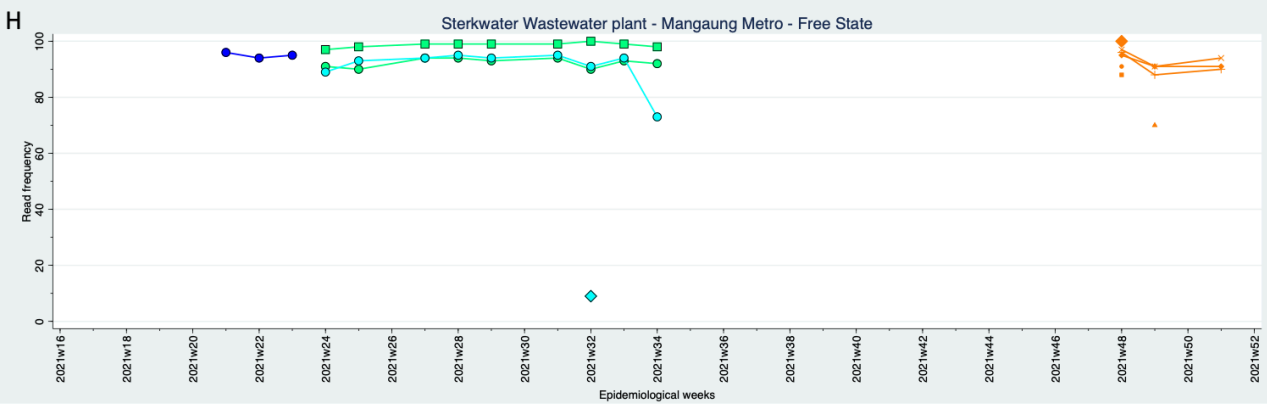


Figure 1 H-I

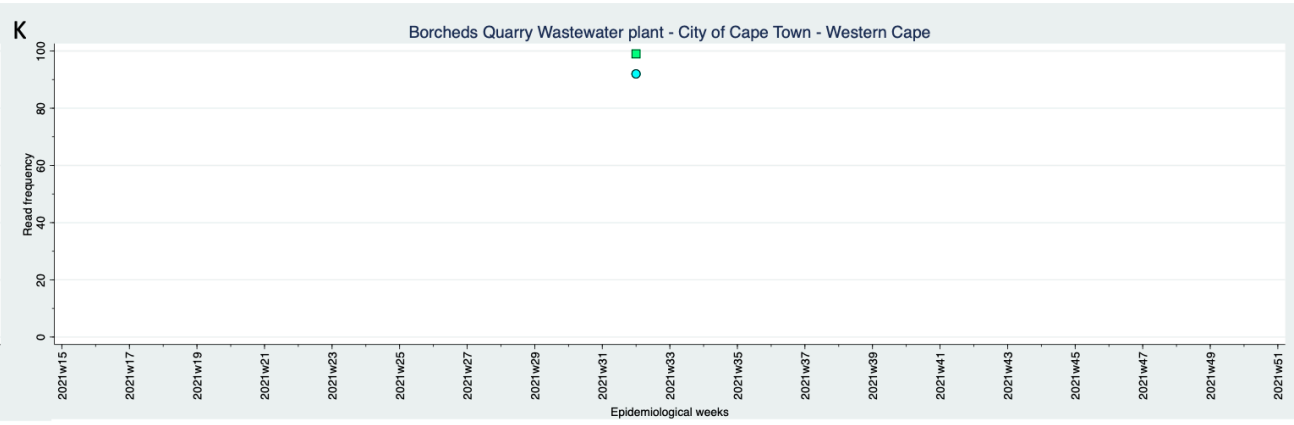
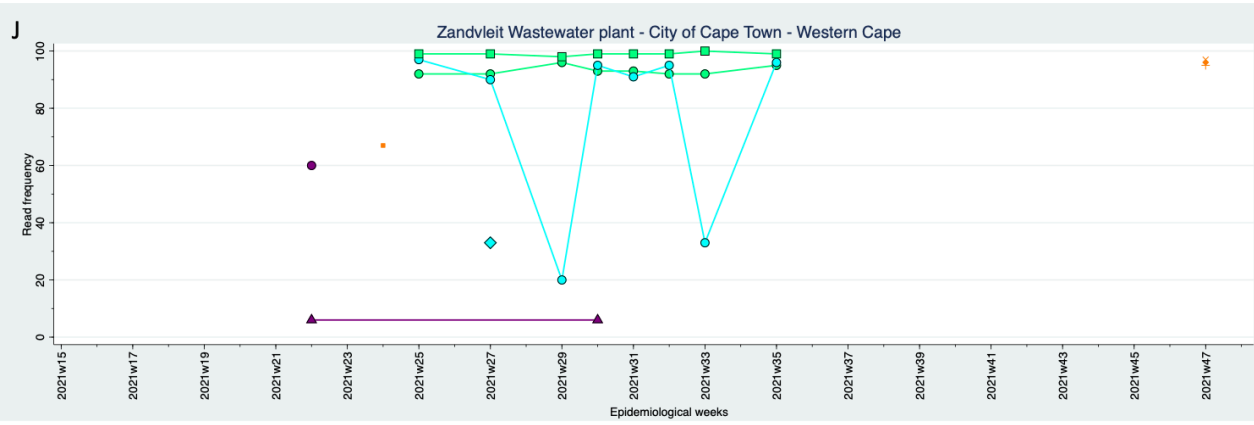


Figure 1 J-K



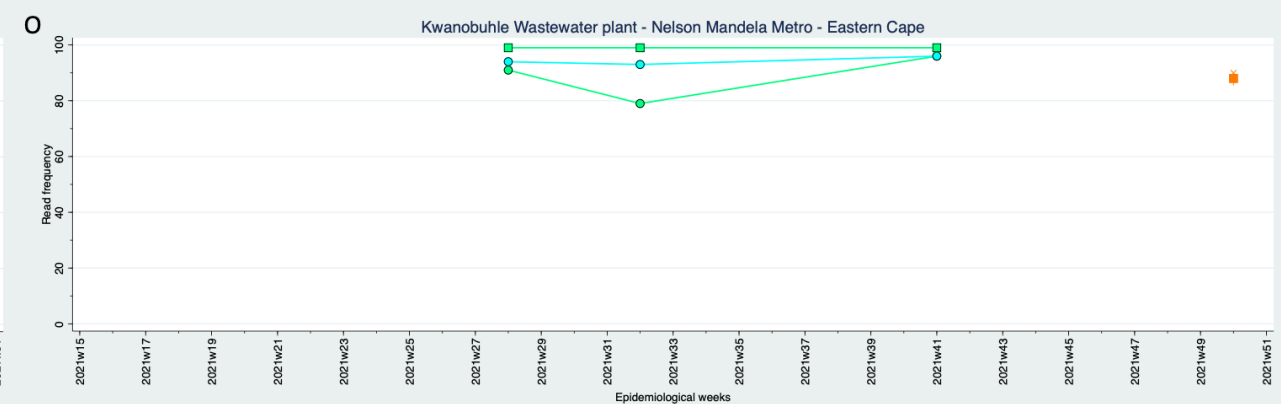
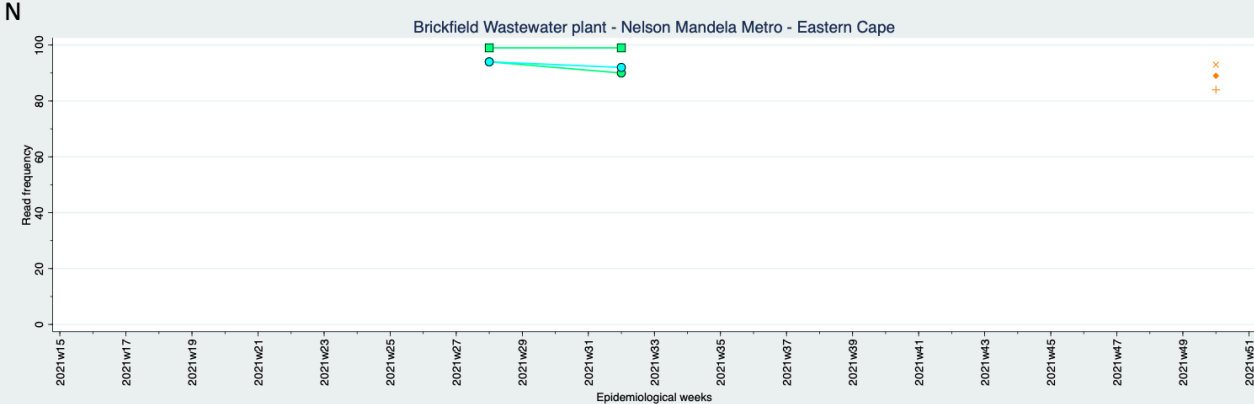
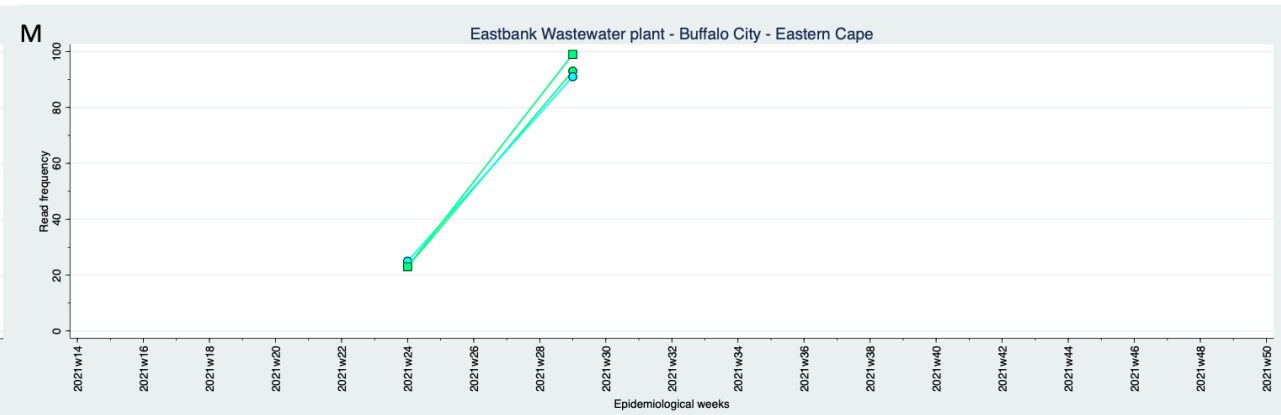
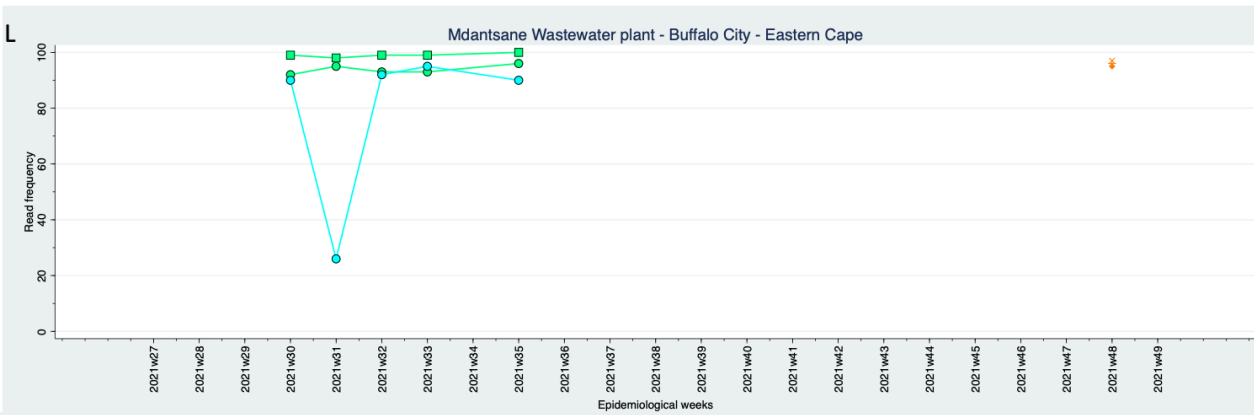


Figure 1 L-O

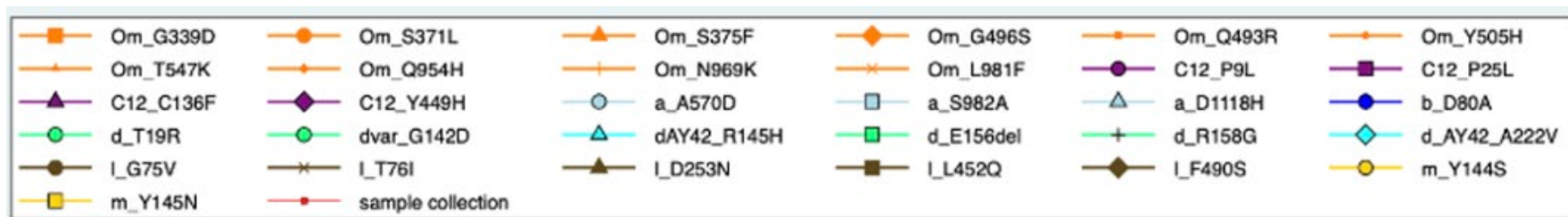


Figure 1 P

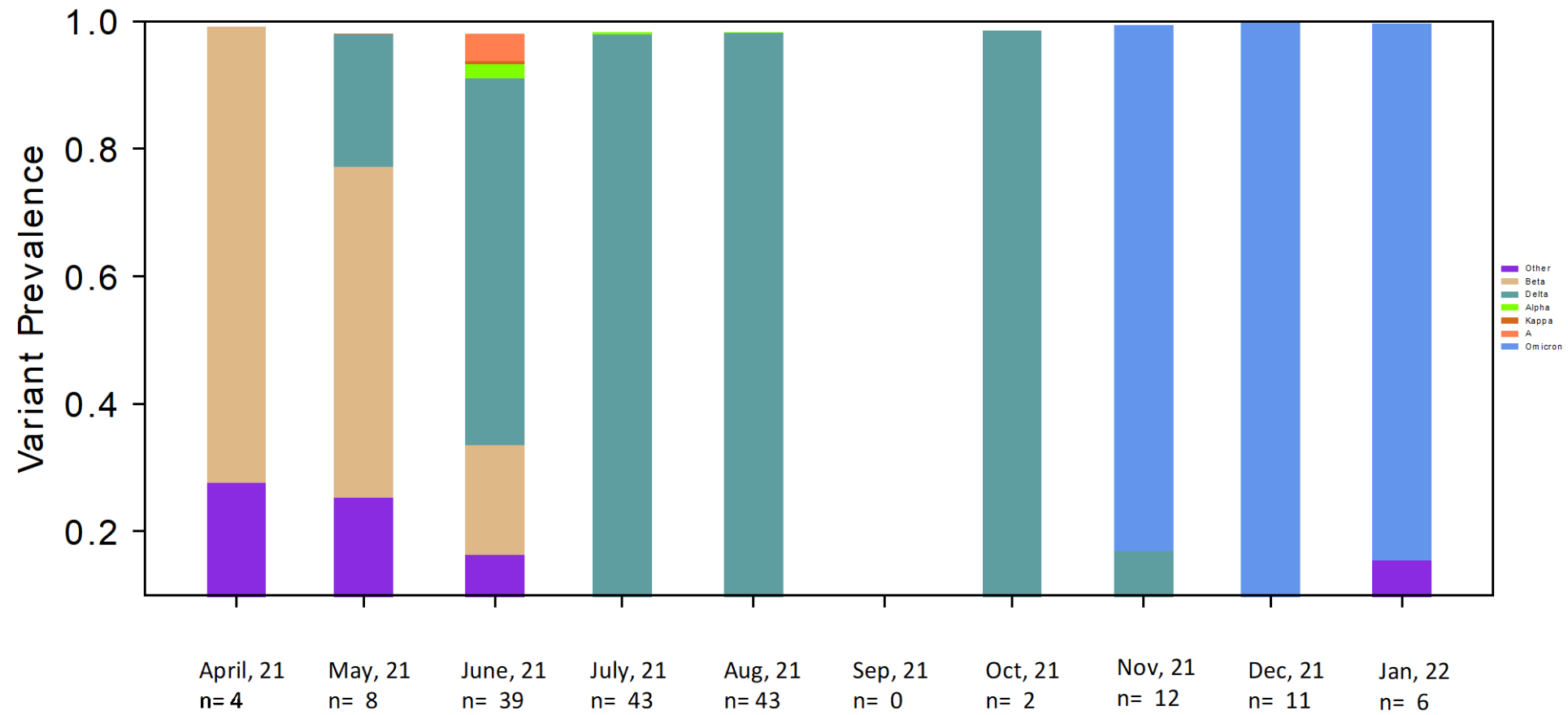


Figure 2A

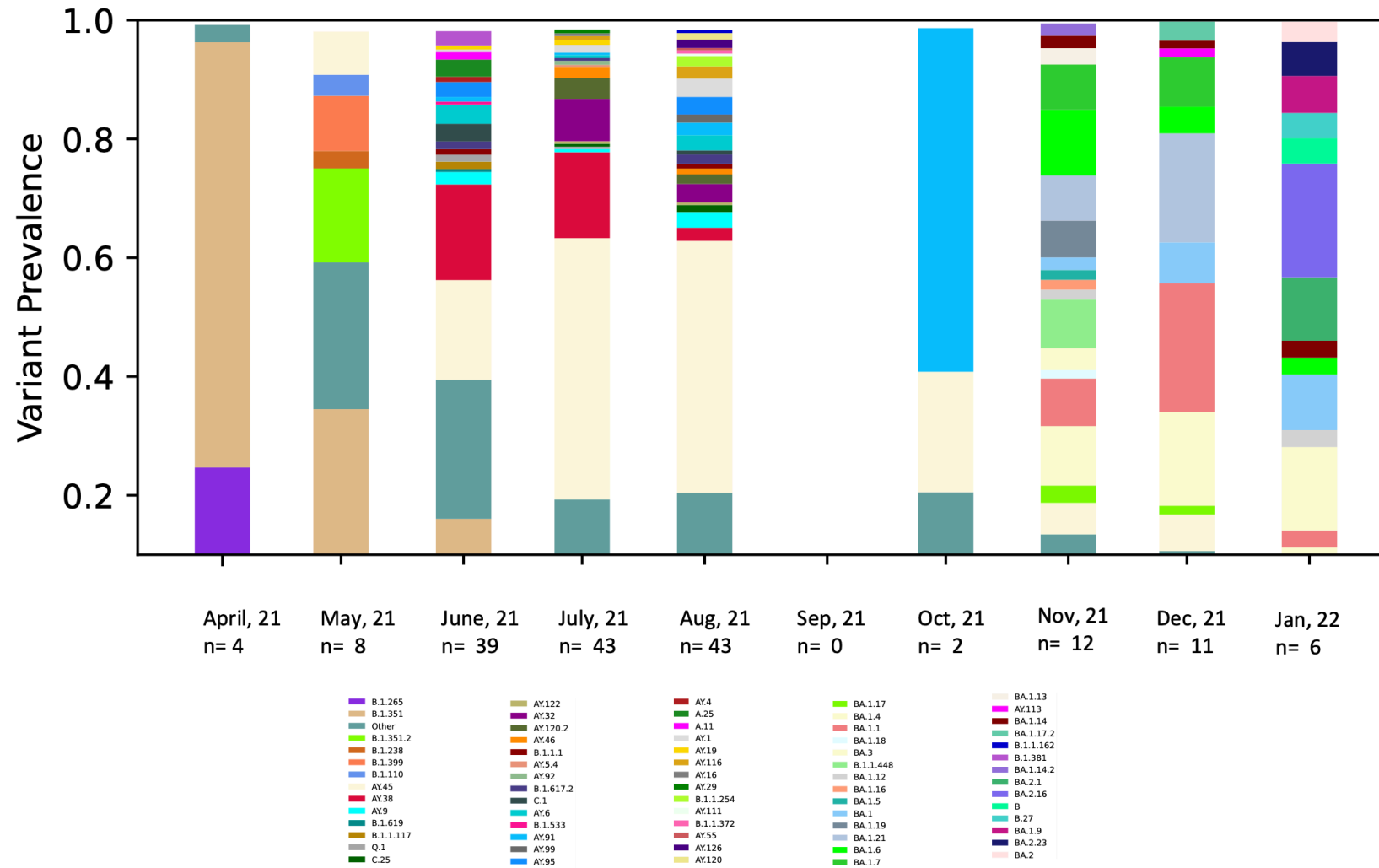


Figure 2B

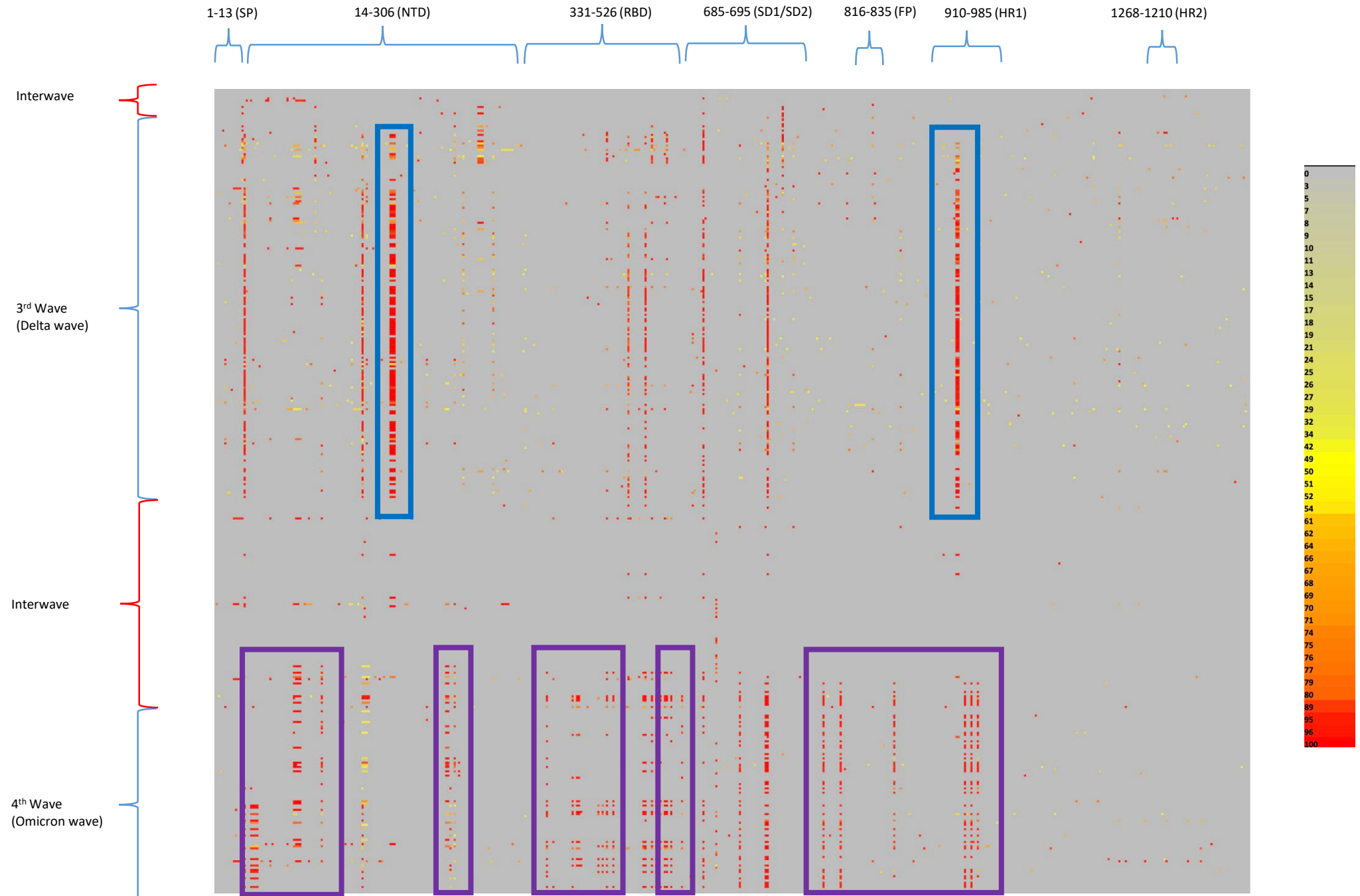
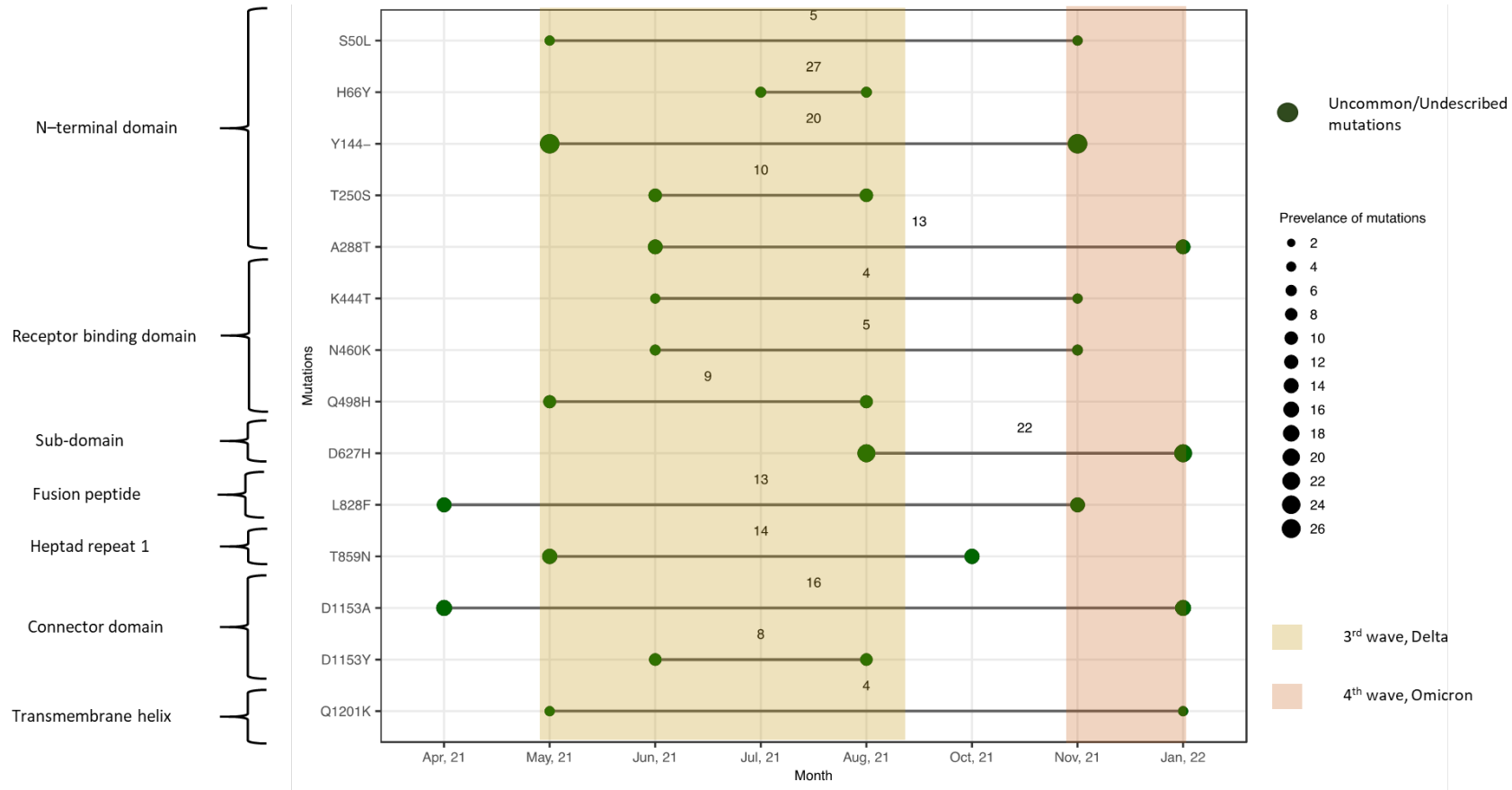


Figure 3



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
Q1201K	4	1,23

Figure 4