

1 **Early Detection of Novel SARS-CoV-2 Variants from Urban and Rural Wastewater**
2 **through Genome Sequencing and Machine Learning**

3
4 Xiaowei Zhuang^{1,2,3#}, Van Vo^{1#}, Michael A. Moshi^{1,2}, Ketan Dhede^{1,2}, Nabih Ghani¹,
5 Shakraiz Akbar, Ching-Lan Chang^{1,2}, Angelia K. Young⁴, Erin Buttery⁴, William Bendik⁴,
6 Hong Zhang⁴, Salman Afzal⁴, Duane Moser⁵, Dietmar Cordes³, Cassius Lockett^{4*},
7 Daniel Gerrity^{6*}, Horng-Yuan Kan^{4*}, Edwin C. Oh^{1,2,7,8*}

8
9 ¹Laboratory of Neurogenetics and Precision Medicine, College of Sciences,

10 ²Neuroscience Interdisciplinary Ph.D. program, ⁷Department of Brain Health,

11 ⁸Department of Internal Medicine, Kirk Kerkorian School of Medicine at UNLV, University

12 of Nevada Las Vegas, Las Vegas, NV 89154; ³Cleveland Clinic Lou Ruvo Center for

13 Brain Health, Las Vegas, NV. ⁴Southern Nevada Health District, Las Vegas NV, 89106;

14 ⁵Division of Hydrologic Sciences, Desert Research Institute, Las Vegas, NV 89119;

15 ⁶Southern Nevada Water Authority, P.O. Box 99954, Las Vegas NV, 89193.

16
17 #These first authors contributed equally to this article.

18 *To whom correspondence should be addressed: Cassius Lockett: lockett@snhd.org,

19 Daniel Gerrity: daniel.gerrity@snwa.com, Horng-Yuan Kan: kan@snhd.org, Edwin Oh:

20 edwin.oh@unlv.edu.

21
22 **Key words:** SARS-CoV-2; COVID-19; Variants; Wastewater; Machine Learning; Novel
23 Mutations; Independent Component Analysis; Co-varying Mutation Patterns.

24

25 **Abstract**

26 Genome sequencing from wastewater has emerged as an accurate and cost-effective
27 tool for identifying SARS-CoV-2 variants. However, existing methods for analyzing
28 wastewater sequencing data are not designed to detect novel variants that have not been
29 characterized in humans. Here, we present an unsupervised learning approach that
30 clusters co-varying and time-evolving mutation patterns leading to the identification of
31 SARS-CoV-2 variants. To build our model, we sequenced 3,659 wastewater samples
32 collected over a span of more than two years from urban and rural locations in Southern
33 Nevada. We then developed a multivariate independent component analysis (ICA)-based
34 pipeline to transform mutation frequencies into independent sources with co-varying and
35 time-evolving patterns and compared variant predictions to >5,000 SARS-CoV-2 clinical
36 genomes isolated from Nevadans. Using the source patterns as data-driven reference
37 “barcodes”, we demonstrated the model’s accuracy by successfully detecting the Delta
38 variant in late 2021, Omicron variants in 2022, and emerging recombinant XBB variants
39 in 2023. Our approach revealed the spatial and temporal dynamics of variants in both
40 urban and rural regions; achieved earlier detection of most variants compared to other
41 computational tools; and uncovered unique co-varying mutation patterns not associated
42 with any known variant. The multivariate nature of our pipeline boosts statistical power
43 and can support accurate and early detection of SARS-CoV-2 variants. This feature offers
44 a unique opportunity for novel variant and pathogen detection, even in the absence of
45 clinical testing.

46

47 **Introduction**

48 Public health testing and the sequencing of SARS-CoV-2 genomes have been pivotal in
49 advancing the development of precise vaccine therapeutics and facilitating
50 comprehensive surveillance of circulating variants^{1,2}. However, clinical surveillance of
51 COVID-19 transmission often imposes substantial demands on laboratory resources and
52 relies on individuals actively seeking testing^{3,4}. These challenges underscore the need for
53 complementary, proactive, and cost-effective methods to monitor the emergence and
54 spread of novel SARS-CoV-2 variants.

55
56 In the United States, there has been substantial spatial and temporal variation
57 reported for the dynamics of the COVID-19 pandemic in urban and rural counties^{5,6}.
58 Studies have described disease incidence being initially high in urban locations, followed
59 by a rapid surge in infections from rural areas^{5,6}. Notably, the overall disease incidence is
60 reported to be lower in rural compared to urban regions^{5,7}. Given that rural communities
61 have fewer healthcare resources and an established reluctance to seek medical care, an
62 overarching infection prevention and control effort could benefit significantly from new
63 interventions designed to track disease incidence.

64
65 Wastewater-based epidemiology (WBE) has emerged as a valuable alternative for
66 tracking changes in SARS-CoV-2 viral levels and variants within a community⁸⁻¹². The
67 SARS-CoV-2 virus is a single-stranded RNA virus that can be shed in wastewater through
68 human waste such as feces, saliva, and urine^{13,14}. In comparison to clinical testing,
69 wastewater analyses provide a less-biased approach to viral monitoring, particularly in

70 areas with limited healthcare resources and unclear testing hesitancy rates^{15–19}.
71 Throughout the COVID-19 pandemic, WBE served as a pivotal and cost-effective tool to
72 monitor and characterize the emergence and spread of SARS-CoV-2 variants of concerns
73 (VoCs), offering early detection for potential outbreaks^{8,18,20–23}.

74
75 The complexity of wastewater matrices presents significant challenges in obtaining
76 high-quality nucleic acid sequences and detecting SARS-CoV-2 variants. To overcome
77 this shortfall, targeted hybridization and amplicon-based sequencing methods have been
78 implemented to characterize the viral composition within a sample^{24–31}. Complementing
79 these sequencing approaches, bioinformatic pipelines have then been developed with
80 unique computational considerations that support the identification and quantification of
81 SARS-CoV-2 variants^{18,21,32–36}. For example, the COJAC pipeline³³ utilizes a variant-
82 specific mutation pattern and counts aligned read pairs to detect the presence of a VoC
83 in wastewater samples, even with a relatively low viral load. In addition, the Vpipe³⁷,
84 LoFreq³⁸, or iVar³⁹ pipeline can be used to define single nucleotide polymorphisms (SNPs)
85 with alternative allele frequencies within the SARS-CoV-2 genome. To determine the
86 presence and abundance of VoCs, these pooled SNP alternative allele frequencies are
87 generally modeled as a linear combination of predefined VoCs using GISAID or UshER-
88 curated reference barcodes^{20,21,34–36}. Among various linear regression models and
89 optimization methods to estimate the abundance of variants, a pipeline called Freyja³⁴ is
90 frequently employed due to its simplicity in modeling and interpretation. However, despite
91 the widespread use of these pipelines, a shared bias toward pre-defined reference
92 barcodes exists, potentially leading to incorrect variant predictions with metadata errors.

93 This bias can be exacerbated when either 1) the variants included in the reference
94 barcodes do not match the current circulating VoCs in the communities, or 2) a new VoC
95 may be circulating within the community, but has not been identified through clinical
96 sequencing. Furthermore, since wastewater samples represent a composite of multiple
97 clinical genomes, there may be limited statistical power to detect emerging VoCs with low
98 abundance in a single sample.

99

100 Here, we introduce a multivariate method designed to analyze SARS-CoV-2
101 wastewater sequencing data and identify circulating variants. We hypothesize that our
102 independent component analysis (ICA)-based pipeline called ICA-Var (Independent
103 Component Analysis of Variants) can leverage multiple sequencing datasets to amplify
104 statistical power and thereby enable early and precise detection of variants within the
105 community. To validate our approach, we compare the results obtained using our pipeline
106 with those generated by the state-of-the-art tool Freyja³⁴. Our approach also identifies
107 emerging co-varying mutation patterns, which may belong to more recent VoCs or have
108 not been reported. Collectively, our findings demonstrate the effectiveness of this new
109 pipeline, even in the absence of clinical data. These results underscore the potential for
110 ICA-Var to identify mutation patterns within the SARS-CoV-2 genome that could give rise
111 to novel circulating variants.

112

113 **Results**

114 **Large-scale genome sequencing and the development of a computational pipeline**

115 Analyzing wastewater SARS-CoV-2 genomes presents inherent complexities resulting
116 from various factors. For example, the use of short sequencing reads introduces
117 challenges in accurately phasing genomes and the degradation of viral genomes in
118 environmental samples contributes to uneven genome coverage and sequencing read
119 depth. To address these challenges, we sequenced 3,659 wastewater samples using an
120 amplicon-targeted approach and employed stringent quality control measures. Using a
121 minimum threshold of 80% genome coverage at more than 50X sequencing depth, we
122 selected 1,385 of these samples, covering 59,422 locations/mutations on the genome,
123 for further analysis (**Supplementary Figure 1**). Leveraging this extensive dataset, we
124 developed a data-driven approach named ICA-Var. This method transforms mutation
125 frequencies in wastewater samples into independent sources with co-varying mutation
126 patterns and utilizes a dual-regression method to re-associate the independent sources
127 back to the original samples (**Figure 1A and Methods**). We hypothesized that the ICA
128 sources could effectively capture the evolving dominant SARS-CoV-2 variants over time,
129 with each being characterized by distinct determinant mutation patterns. To evaluate the
130 performance of our tool, we conducted a comparative analysis of variant detection against
131 the state-of-the-art tool known as Freyja (**Figure 1A-B**).

132

133 In late 2021, both Freyja and ICA-Var reliably identified B.1.617.2 (Delta) and BA.1
134 (Omicron) VoCs in wastewater samples (**Figure 1C**, yellow in first two rows), reflecting
135 the prevalence of both variants during this period. In 2022, ICA-Var demonstrated the

136 ability to detect BA.2, BA.4, BA.5, BF.7, BQ.1, XBB.1, and XBB.1.5 variants one or several
137 weeks before Freyja (**Figure 1C**, green). Consistent detection of Omicron variants was
138 obtained by Freyja and ICA-Var throughout 2022 (**Figure 1C**, yellow). In 2023, both Freyja
139 and ICA-Var successfully identified XBB.1.16 in late March, a month prior to the first
140 sequenced clinical sample in Southern Nevada (**Figure 1C**, yellow). For more emerging
141 VoCs in 2023, such as EG.5, ICA-Var detected this variant in early June, coinciding with
142 the week of the first reported clinical sequence in Southern Nevada (first red box in **Figure**
143 **1C**). In contrast, Freyja reliably identified the EG.5 signal only once the VoC became more
144 prevalent in early July. Similarly, for HV.1, and BA.2.86 VoCs, ICA-Var detected the
145 presence of these variants in wastewater several weeks before Freyja (**Figure 1C**, red
146 boxes).

147
148 To explore the earlier detection of the emerging VoCs EG.5, HV.1, and BA.2.86 by
149 ICA-Var compared to Freyja, we generated a heatmap illustrating alternative allele
150 frequencies at the dominant mutation sites for these variants. This heatmap represents
151 samples from the first week of detection by each method (**Figure 2**). Specifically, for EG.5,
152 ICA-Var initially identified the variant during the week of 06/05/2023, and two wastewater
153 samples in this week exhibited reliable mutation frequencies at three out of eight EG.5
154 dominant mutation sites (**Figure 2**, top row in panel EG.5). Freyja reported abundances
155 of 0.66% and 0.53% for EG.5 in these two samples, respectively. Furthermore, an
156 additional wastewater sample from the same week showed reliable mutation frequencies
157 at two EG.5 dominant mutation sites, but Freyja did not identify EG.5 in this particular
158 sample. As a multivariate method, ICA-Var leveraged all these samples with reliable, yet

159 relatively low prevalence of EG.5 mutation sites, thereby enhancing statistical power and
160 consequently enabling an earlier detection of EG.5 in this particular week. Conversely,
161 Freyja reported a 23.08% abundance of one wastewater sample in the week of
162 07/10/2023. As shown in **Figure 2** (bottom row in panel EG.5), this sample demonstrated
163 reliable alternative allele frequencies at five out of eight EG.5 determinant mutation sites
164 (**Figure 2**, red box). The increased prevalence of EG.5 mutation sites in this sample
165 contributed to the detection of EG.5 in Freyja. Similarly, for HV.1 (**Figure 2**, HV.1 panel)
166 and BA.2.86 (**Figure 2**, BA.2.86 panel), ICA-Var capitalized on multiple wastewater
167 samples with reliable yet relatively low prevalence of dominant mutation sites, thereby
168 enhancing statistical power and achieving an earlier detection. In contrast, Freyja
169 mandated at least one individual sample to exhibit the presence of dominant mutation
170 sites for detection (red boxes in **Figure 2**).

171
172 We further evaluated the earlier detection of VoCs in 2022 for ICA-Var and Freyja
173 (**Supplementary Figure 2**). In addition to enhancing statistical power, the inclusion of
174 deletions as additional sites in the proposed pipeline (indicated by orange boxes in
175 **Supplementary Figure 2**) played a significant role in the earlier detections of BA.2, BA.4,
176 and BA.5 variants compared to Freyja. This advantage arises from the fact that no
177 deletions were utilized in the inference process within the default settings of Freyja, a
178 result previously discussed in another computational pipeline designed to analyze
179 wastewater sequencing data³⁵.

180

181 **Detection of VoCs in urban and rural samples**

182 From the beginning of 2022, we sequenced and analyzed wastewater samples from rural
183 areas in Southern Nevada (**Supplementary Table 1**). The number of urban (orange curve)
184 and rural (yellow curve) samples for each week were plotted in **Supplementary Figure**
185 **1D**. We conducted a comprehensive urban-rural epidemiological comparison in our
186 wastewater analyses (**Methods**), with samples categorized as urban and rural analyzed
187 separately for each week. We present a summary of the detection of 18 VoCs utilizing
188 both the established Freyja pipeline (**Figure 3A**) and our proposed ICA-Var pipeline.
189 (**Figure 3B**).

190
191 ICA-Var and Freyja both identified 16 out of the 18 VoCs in urban wastewater samples
192 prior to detecting these VoCs in wastewater samples from rural locations (**Figure 3A-B**).
193 This data suggest that new SARS-CoV-2 variants typically enter urban areas first before
194 spreading into rural areas. Interestingly, XBB.1 and FL.1.5.1 were first detected in rural
195 wastewater samples by either Freyja or ICA-Var (black boxes in **Figure 3A-B**). More
196 specifically, Freyja first identified the XBB.1 variant in a rural sample in the week of
197 11/07/2022 (dashed red box in **Supplementary Figure 2**, panel XBB.1), but ICA-Var was
198 able to detect XBB.1 one week prior to Freyja in urban samples (**Figure 3C**). Both ICA-
199 Var and Freyja first detected FL.1.5.1 in rural samples on 07/10/2023 (**Figure 3C**).
200 Detailed inspections showed that one rural sample on 07/12/2023 showed an
201 overwhelming presence of FL.1.5.1 dominant mutations (dashed red box in
202 **Supplementary Figure 2**, panel FL.1.5.1), which contributed to this earlier detection in
203 rural areas. In contrast, urban samples demonstrated a much lower alternative allele
204 frequencies and prevalence at FL.1.5.1 mutations.

205

206 **Identification of mutation sites with significant time-evolving contributions**

207 Out of 59,422 mutation sites included, and following the analyses pipeline in **Figure 1A**,
208 a total of 730 mutation sites demonstrated significant contributions during the multivariate
209 group ICA (**Supplementary Figure 1B**). Among them, a subset of 177 mutations showed
210 a significant time-evolving contribution from August 2021 to November 2023 (**Methods**).
211 As a proof of concept, we cross-referenced these 177 mutations with dominant mutation
212 sites in B.1.617.2, BA. 1 and XBB.1 variants, and plotted their weekly contributions in
213 **Figure 4**.

214

215 Significant fluctuating contributions were observed in late 2021 for 16 out of 25
216 dominant mutation sites in B.1.617.2 (**Figure 4**, panel B.1.617.2). These contributions
217 gradually declined through 2022 and diminished further in 2023. For the BA.1 variant,
218 there was a noticeable increase in contributions related to the associated mutations in
219 late 2021, peaking in early 2022 (**Figure 4**, panel BA.1, orange box). For several BA.1
220 mutation sites, time-evolving contributions continued to fluctuate in 2023, and their
221 involvement in other Omicron sub-lineages (e.g., XBB.1) was reported at nextstrain.org.
222 In addition, 22 out of 25 dominant mutations in XBB.1 displayed significant time-evolving
223 contributions, with a substantial impact after September 2022. Similar fluctuation patterns
224 were observed for several mutation sites (**Figure 4**, panel XBB.1, orange box), indicating
225 that these mutation sites co-vary together and demonstrate a recombinant nature for
226 XBB.1. Collectively, our data (**Figure 4**) demonstrate that time-evolving contributions for
227 mutation sites identified by ICA-Var were consistent with the clinical emergence of Delta,

228 Omicron, and XBB.1 variants. These results further solidify the foundation for the
229 proposed pipeline, indicating its potential to identify novel mutation patterns that may lead
230 to the emergence of new variants.

231

232 **Discovery of potential novel variants**

233 Upon cross-referencing with dominant mutation sites in 15 VoCs (18 VoCs in **Figure 1B**,
234 excluding emerging variants EG.5, HV.1, and BA.2.86), a set of 113 mutations sites
235 emerged as potential novel mutations. Using a hierarchical clustering algorithm with ward
236 distance, six clusters were obtained at a cut-off ward distance of 18 (**Figure 5A, Methods**).
237 Among these clusters, cluster 2, 3, 4, and 5 showed overlapping mutation sites with
238 emerging variants in late 2023 (bottom table in **Figure 5A**). Using cluster 3 as an example,
239 we observed two sets of co-varying patterns after 06/2023 (dashed orange boxes in
240 **Figure 5B**), both were overlapping with dominant mutations of EG.5 and HV.1.
241 Furthermore, there were no overlapping mutations between cluster 1 or 6 with known
242 mutation sites in emerging variants in late 2023 (bottom table in **Figure 5A**). Co-varying
243 patterns after 2023/08 were evident for mutation sites in cluster 1 (**Figure 5D**). For these
244 eight mutations, we verified the presence of these sites in clinical sequencing data from
245 GISAID. Our analysis revealed that these mutations had been infrequently reported in
246 any clinical samples (**Supplementary Figure 3**). Hence, these mutations could
247 potentially lead to the emergence of novel SARS-CoV-2 variants and warrant close
248 monitoring, pending clinical testing.

249

250 **Discussion**

251 Wastewater-based epidemiology (WBE) offers a unique opportunity to monitor the
252 emergence and spread of SARS-CoV-2 variants at the population level. Our proposed
253 pipeline demonstrates early detection of the SARS-CoV-2 VoCs in wastewater preceding
254 identification in clinical data. We further show the spatial and temporal dynamics of most
255 emerging SARS-CoV-2 VoCs transitioning from urban to rural areas. Leveraging the data-
256 driven nature of our proposed pipeline, ICA-Var identifies modules of mutations in the
257 SARS-CoV-2 genome that are consistent with parallel time-changing patterns, and
258 consequently gave rise to VoCs from August 2021 to November 2023. The proposed
259 method offers an opportunity to identify mutation sites that are occurring simultaneously
260 and could lead to potential novel variants, even in the absence of clinical data. Importantly,
261 ICA-Var can also take advantage of the dual regression feature to associate an identified
262 group source with a limited number of recently collected samples.

263

264 Enhanced sensitivity and specificity in SARS-CoV-2 VoC detection. Wastewater samples
265 are a composite of multiple clinical genomes spanning a local community at a given time
266 point⁴⁰. COVID-19 clinical testing and reports indicate that certain VoCs were dominant
267 at specific time points from August 2021 to November 2023^{41–43}. Our method, ICA-Var
268 enables the separation of multiple genomic signals into independent sources⁴⁴. Utilizing
269 a significant number of wastewater samples spanning this timeframe, retrospective ICA
270 can uncover the original mutation profile, each representing an individual or a set of
271 similar VoCs spanning communities at different time points. Notable, ICA-Var can handle
272 non-Gaussian and non-linearly mixed signals, operates without the need for prior
273 knowledge, and performs blind source separation⁴⁴. These inherent properties make ICA

274 robust to our real-world application of de-mixing wastewater samples, enabling the
275 identification of clinically-relevant VoCs.

276

277 Following group ICA, we performed the dual-regression analysis to re-associate the
278 original source with weekly samples to investigate and characterize the signals from ICA
279 within each week. Previous studies have applied the dual-regression method in functional
280 magnetic resonance imaging data analysis to associate group networks (i.e., sources)
281 identified by ICA with individual brain maps⁴⁵⁻⁴⁸. Following the same concept, we adapted
282 the dual-regression method for our approach to project group sources back onto weekly
283 samples. This step allows us to enhance the specificity of the group signals and provides
284 accurate localization of mutation patterns in weekly samples. It also leads to enhanced
285 interpretability when comparing against dominant mutations from each VoC derived from
286 clinical sequencing data.

287

288 Collectively, as compared to the state-of-the-art Freyja tool that analyzes each
289 individual wastewater sample with a univariate approach, the proposed pipeline boosts
290 the statistical power and enables the earlier and more accurate detection of each VoC
291 (**Figure 1** and **Figure 3**). The intention to incorporate deletion information in our proposed
292 analyses also contributes to the enhanced sensitivity and accuracy (**Supplementary**
293 **Figure 2**). Earlier detection of VoCs from wastewater data then enables public health
294 authorities to implement timely and targeted interventions to mitigate the spread of the
295 virus⁴⁰.

296

297 ICA-Var does not require clinical data to identify a novel variant.

298 Wastewater monitoring and clinical sequencing data of SARS-CoV-2 can provide a
299 comprehensive understanding of the emergence, spread and prevalence of a virus^{33,49,50}.

300 A time-dependent and accurate reference barcode for circulating VoCs are often required
301 to identify emerging variants^{34,35}. Therefore, these methods (such as Freyja and COJAC)
302 are potentially restricted from identifying or forecasting potential novel variants in the
303 absence of clinical sequencing data and require a “correct” barcode of circulating VoCs
304 for accurate detection.

305

306 As a data-driven approach, our proposed pipeline (**Figure 1**) recognizes mutation
307 sites within the SARS-CoV-2 genome through the identification of co-varying and time-
308 evolving patterns in group sources. This crucial step allows us to identify contributing
309 mutation sites for various VoCs in wastewater at different time points from August 2021
310 to November 2023, without any prior knowledge of circulating VoCs (**Figure 4**). The
311 proposed pipeline additionally identifies co-varying mutation patterns that are more recent
312 and contribute to emerging group sources that have not been reported in clinical
313 sequencing data (**Figure 5**). Therefore, these mutation sites could potentially give rise to
314 novel SARS-CoV-2 variants.

315

316 VoCs spread from urban to rural areas.

317 Besides methodological developments, our study shows that each SARS-CoV-2 VoC is
318 in general first detected in wastewater samples from urban areas and later in wastewater
319 samples from rural areas (**Figure 3**). This observation is in concordance with a previous

320 report on COVID-19 epidemic dynamics in the United States⁵. In addition to highlighting
321 these dynamic patterns, our results underscore the feasibility of monitoring SARS-CoV-2
322 VoCs through wastewater samples obtained from rural areas. Given reports that residents
323 in rural locations are at a higher risk for disease and often lack healthcare resources⁷,
324 WBE provides a practical and effective way to monitor the disease emergence and
325 estimate the disease spread and prevalence.

326

327 Limitations.

328 As a data-driven method, ICA-Var requires a significant number of samples with high
329 genome coverage and depth to produce stable results. Therefore, the proposed pipeline
330 may not be suitable for scenarios with a limited number of wastewater samples or if
331 sequencing metrics indicate genome coverage below 50% and low sequencing depth (i.e.,
332 less than 10 reads per sequenced base). Moreover, one assumption of ICA-Var is that
333 sources are independent and linearly separable. In our application, the independence of
334 underlying signals comes from different dominant mutation sites for various VoCs and
335 different dominance of VoCs at different time points. Both conditions demand a relatively
336 large number of wastewater samples to generate meaningful results. The multivariate
337 nature of the proposed pipeline further restricts its application to detect the presence of
338 VoCs in a single wastewater sample. Despite the re-association of group sources with
339 individual samples during the dual-regression step, the regression nature constrains the
340 algorithm's stability for single-sample analyses. As a result, the proposed method can
341 only determine the existence of VoCs within a timeframe from multiple samples, in our
342 case, multiple samples from various wastewater sampling locations in southern Nevada.

343 Finally, the proposed pipeline cannot estimate the abundance of each VoC from the
344 wastewater sample. This limitation arises from ICA-Var's inability to distinguish between
345 signal and noise in mixed data, and its treatment of each source with equal weight. While
346 the former may not pose a significant concern in our application, given the bioinformatics
347 processing pipeline's retention of relevant SARS-CoV-2 signals, the latter impedes our
348 ability to discern the abundance or significance of each identified source.

349

350 **Methods**

351 Wastewater sample collection, processing, and sequencing. A total of 3,659 wastewater
352 samples were collected from urban and rural locations in Southern Nevada (detailed in
353 **Supplementary Table 1**) from August 2021 to November 2023. After collection,
354 samples were placed on ice in the field and stored under refrigeration until processing
355 (hold time < 36 h). Nucleic acids from wastewater samples were isolated using the
356 Promega Wizard Enviro Total Nucleic Acid Kit (Cat #A2991) following the
357 manufacturer's protocol. In addition, we modified the Promega protocol by lysing
358 wastewater with the protease solution and binding free nucleic acids using NucleoMag
359 Beads from Macherey-Nagel (Cat #744970). Total RNA (>10ng) was processed for first-
360 strand cDNA synthesis using the LunaScript RT SuperMix Kit (New England BioLabs).
361 Amplicon-based sequencing libraries were constructed using the CleanPlex SARS-CoV-
362 2 FLEX Panel from Paragon Genomics. Libraries were sequenced on an Illumina
363 NextSeq 500 or NextSeq 1000 platform with 300 cycle flow cells.

364

365 Wastewater sequence data processing. Processing of sequencing data followed a

366 modification of our previously published pipeline¹⁸. Briefly, upon sequencing, Illumina
367 adapter sequences were trimmed from read pairs using cutadapt version 4.2⁵¹.
368 Sequencing reads were then mapped to the SARS-CoV-2 reference genome
369 (NC_045512.2) using bwa mem, version 0.7.17-r1188⁵². Paragon Genomics CleanPlex
370 SARS-CoV-2 FLEX tiled-amplicon primers were trimmed from the aligned reads using
371 fgbio TrimPrimers version 2.1.0 in hard-clip mode. Variants were called by iVar variants
372 v1.4.1³⁹ using mutation sites with alternative allele frequencies with respect to the
373 reference Wuhan SARS-CoV-2 genome⁵³, Genome coverage and read depth were
374 calculated using samtools v1.16.1⁵⁴. Strict quality control (QC) was enforced as only
375 wastewater samples with 50x depth covering more than 80% of SARS-CoV-2 genome
376 were retained in the following analyses. Collectively, a total of 1,385 samples, from August
377 2021 to November 2023, covering 59,422 mutation sites of SARS-Cov-2 variants were
378 used for the following analyses (**Supplementary Figure 1C-D**).

379
380 Public health sample analyses. Public health samples were processed and sequenced
381 at the Southern Nevada Public Health Laboratory (SNPHL) as part of the Southern
382 Nevada Health District's surveillance of the COVID-19 pandemic. Visual inspection of
383 sequencing reads using the Integrative Genomics Viewer (IGV) was performed to
384 assess whether mutations had sufficient sequencing support. A TheiaCoV_Illumina_PE
385 workflow using Nextclade version 2.14.0 was used to assign lineages.

386
387 Retrospective independent component analysis of Variants (ICA-Var). Mathematically, let
388 $\mathbf{Y} \in \mathbb{R}^{1,385 \times 59,422}$ denote the 59,422 mutation frequencies (i.e., the proportion of reads at

389 a site that contains the mutation) from 1,385 wastewater samples. Since wastewater
390 samples are aggregations of genomes from multiple infected individuals with various virus
391 lineages, Y could be considered as a multivariate mixed signal of SARS-CoV-2 variants
392 spanning the local community. The data-driven ICA approach separates this multivariate
393 signal into additive subcomponents⁴⁴: $Y = AS$, where $A \in \mathbb{R}^{1,385 \times n_{ica}}$ denotes the mixing
394 matrix and $S \in \mathbb{R}^{n_{ica} \times 59,422}$ represents the source matrix (**Figure 1A**, shaded grey box). In
395 ICA-Var, the number of ICA components (n_{ica}) was determined from the minimum
396 description length criterion, and fastICA algorithm was utilized to perform ICA. In our
397 analysis, ICA was repeated 50 times with different initial values and components from
398 each run were clustered and visualized⁵⁵. Only reliable estimates corresponding to tight
399 clusters were retained as final sources (S , **Supplementary Figure 1A**).

400

401 The original ICA method assumes that all sources (i.e., subcomponents S) are non-
402 Gaussian and that the sources are statistically independent from one another⁴⁴. In
403 analyzing our wastewater samples using ICA-Var, the independence comes mostly from
404 different mutation patterns and various circulating windows of time for each VoCs^{41–43}.
405 The sparsity in Y contributes to the non-Gaussianness. In this case, the source matrix
406 could represent a co-varying mutation pattern in different time windows, and could
407 therefore serve as data-driven reference barcodes for mutation frequencies co-existing in
408 wastewater samples. From this perspective, the ICA-Var pipeline can be considered as
409 running a multivariate regression and determining the design matrix (S , i.e., reference
410 barcodes) from the data under the constraint of independence of the sources. In our study,
411 we conducted a retrospective analysis using ICA-Var on 1,385 wastewater samples

412 spanning from August 2021 to November 2023. We predicted that the ICA sources could
413 capture the evolving dominant SARS-CoV-2 VoCs over time, each characterized by
414 unique determinant mutation patterns.

415

416 Next, we identified a set of contributing mutations (i.e., significant mutations) for each
417 source in \mathcal{S} . We selected mutation sites with values exceeding the mean ± 2 standard
418 deviations (i.e., 4.55% of all mutation sites) in each row of \mathbf{S} as contributing mutations⁴⁸.

419 A binary matrix $\widehat{\mathbf{S}}$ was then computed to retain only these contributing mutations
420 (**Supplementary Figure 1B**). The pipeline developed in this manuscript and the data
421 used to generate the results are available at <https://github.com/zhuangx15/ICAvAr>.

422

423 Dual-regression to back-project source matrix onto weekly wastewater samples. To
424 further determine the dominant mutations and VoCs for each week, we performed a dual-
425 regression analysis⁴⁵ to project the ICA source matrix (\mathbf{S}) back onto weekly wastewater
426 samples (**Figure 1A**, shaded grey boxes). The term "dual-regression" stems from the
427 utilization of two regression procedures employed to estimate source and de-mixing
428 dynamics for each week against the original data. More specifically, let $\mathbf{y}_i \in$
429 $\mathbb{R}^{N_{sample_i} \times 59,422}$ denote mutation frequencies for N_{sample_i} samples in the i^{th} week, we
430 then, 1) used the all-sample source matrix (\mathbf{S}) as a set of source regressors in a general
431 linear model (GLM), to find week-specific de-mixing dynamics ($\mathbf{a}_i = \mathbf{y}_i \mathbf{S}^{-1}$, $\mathbf{a}_i \in$
432 $\mathbb{R}^{N_{sample_i} \times n_{ica}}$) associated with all-sample source matrix (\mathbf{S}); and 2) used week-specific
433 de-mixing dynamics (\mathbf{a}_i) as a set of regressors in a second GLM, to find the week-specific
434 source matrix ($\mathbf{s}_i = \mathbf{a}_i^{-1} \mathbf{y}_i$, $\mathbf{s}_i \in \mathbb{R}^{n_{ica} \times 59,422}$) that were still associated with the all-sample

435 source matrix (\mathbf{S}). This process yields pairs of estimates forming a dual space, jointly
436 providing the best approximation for the original all-sample ICA source matrix in each
437 weekly sample. In summary, we obtained dual-regressed week-specific source matrix s_i
438 for 113 weeks from August 2021 to November 2023.

439

440 ICA source matrix annotation to SARs-CoV-2 VoCs. To delineate the VoCs each week,
441 we annotated our dual-regressed ICA source matrix s_i by comparing them against the
442 known mutations in VoCs from clinical SARS-CoV-2 sequencing data (**Figure 1A**, bottom
443 row). Next, we focused on 18 VoCs that have either been or were circulating in Southern
444 Nevada between 2021 and 2023. Due to the potential shared dominant mutations among
445 VoCs during evolution, a hierarchical structure was formed from these 18 VoCs based on
446 the phylogenetic tree from www.covspectrum.org (**Figure 1B**, top panel). Dominant
447 mutation sites for each VoC were determined as follows: 1) mutations with more than 90%
448 prevalence among clinical sequences reported at www.covspectrum.org were retained;
449 2) for lineages in level 2, 3, and 4 of the hierarchical tree, mutations that existed in their
450 higher level VoCs were excluded to maintain a unique determinant mutation set (**Figure**
451 **1A**, bottom row); and 3) mutations with both substitutions and deletions were included in
452 step 1) and 2). The number of dominant mutations of each VoC were listed in parentheses
453 in **Figure 1B**, top panel.

454

455 We next binarized each row of week-specific source matrix (s_i) by keeping only mutations
456 with values greater than mean ± 2 standard deviations, as these mutations were
457 contributing significantly towards the source (\hat{s}_i). We annotated the binarized week-

458 specific source matrix (\hat{s}_i) using the dominant mutations from the VoCs by computing the
459 following six matrices: 1) Spearman's rank correlation coefficient (ρ); 2) sensitivity; 3)
460 specificity; 4) area under the receiver operating characteristic curve (AUC); 5) F1 score;
461 and 6) Jaccard Index (JI). As shown in the dendrogram of these six matrices
462 (**Supplementary Figure 4C**), JI and F1score, AUC and sensitivity were highly similar to
463 each other, respectively. The measure of specificity is dependent on the count of non-
464 dominant mutations within each VoC. This count is arbitrarily determined and influences
465 the number of dominant mutations observed in other VoCs. Therefore, we established
466 our annotation criteria using the F1 score, sensitivity, and the Spearman's correlation
467 values (**Supplementary Figure 4B**), and further based on hierarchical levels and number
468 of determinant mutations of each VoC (detailed in **Figure 1B**).

469
470 We compared VoC annotations of the proposed pipeline against results from the
471 state-of-the-art tool Freyja³⁴ (version 1.4.5, **Figure 1A**, dashed boxes). For this
472 comparison, Freyja was retrospectively and independently applied to each of the 1,385
473 samples, utilizing a barcode comprising 18 VoCs, generated in October 2023. We
474 organized samples into individual weeks, ranging from August 2021 to November 2023.
475 In each week, if the results from Freyja indicated that any wastewater sample contained
476 a VoC with an abundance exceeding 15%, we considered this VoC as detected by Freyja
477 in that specific week (**Supplement Figure 4A**).

478
479 Potential novel mutations. Given that the identification of dominant mutation sites did not
480 necessitate prior knowledge of reference barcodes for VoCs, ICA-Var provides a

481 distinctive approach to discern emerging mutations. This capability extends to
482 contributions that might give rise to novel lineages across local communities, even in the
483 absence of clinical sequencing data. From this perspective, we focused on capturing the
484 time-evolving contributions of significant mutations in each week-specific source matrix
485 (**Figure 1A**, top row), using: $|s_i^{(j)}| = \mathbf{T}\boldsymbol{\beta} + \epsilon$, where $i = 1, \dots, 113$, $s_i^{(j)}$ denoted the source
486 values for j^{th} contributing mutations in i^{th} week, \mathbf{T} denoted a time vector for 113 weeks
487 from August 2021 to November 2023, and $\boldsymbol{\beta}$ represented the time-evolving effect of each
488 contributing mutation. Since flipping signs of the de-mixing matrix in ICA would result in a
489 flipping sign of the source matrix⁴⁴, we focused on the amplitude (absolute value) of each
490 week-specific source (s_i).

491
492 A significant β indicated a critical time-changing contribution for this mutation from
493 2021 to 2023. As a proof of concept, we cross-checked mutations with significant β
494 against known mutations in Delta (B.1.617.2), Omicron (BA.1), and more recent XBB.1
495 variants, and examined their time-evolving contributions. Following clinical reports,
496 mutations in Delta variant (B.1.617.2) should demonstrate significant contributions in
497 2021, mutations in Omicron variant (BA.1) should demonstrate significant contributions
498 from late 2021 to 2022, and mutations in XBB.1 variants should demonstrate significant
499 contributions after late 2022.

500
501 Subsequently, we refined our pool of potential novel mutations by cross-referencing
502 known SARS-CoV-2 variants with mutations exhibiting significant time-evolving
503 contributions. Among the mutations retained, those demonstrating emerging contributions

504 in more recent weeks were identified as candidates with the potential to give rise to novel
505 lineages. To delve deeper, our focus extended to their time-evolving contributions over
506 recent three months, from August 2023 to October 2023. We further performed a
507 hierarchical clustering on these mutations with contributions in the three months as
508 features, to identify co-varying patterns among these mutations. To validate the identified
509 co-varying patterns, we examined co-varying patterns of mutations within each cluster,
510 and cross-referenced them with dominant mutations from the emerging VoCs EG.5, HV.1,
511 and BA.2.

512

513 **Figure Legends**

514 **Figure 1** ICA-Var pipeline and comparisons with Freyja. **(A)** Proposed independent
515 component analyses (ICA) pipeline. Two matrices are reported: SARS-CoV-2 lineages
516 detection each week (bottom row), and potential novel mutations (top row). **(B)**
517 Hierarchical structure of 18 variants of concerns (VoCs). Lineage-defining mutations for
518 each VoC were obtained from clinical data summarized at covspectrum.org, and the
519 number of defining mutations were listed in brackets. Criteria for calling a detection in the
520 proposed pipeline were listed in shaded boxes. Abbreviations: ρ : the Spearman's
521 Correlation coefficient; FDR: false discovery rate. **(C)**. Detection of the emerging VoCs in
522 wastewater from Southern Nevada from August 2021 to November 2023 in the proposed
523 method (first reporting matrix in **(A)**) and the state-of-art tool Freyja. An asterisk (*)
524 indicates at least one clinical sample was reported within that week. Earlier detections of
525 the proposed method were observed for emerging variants EG. 5, HV.1, and BA.2.86 (red

526 triangle boxes). The yellow triangle box indicates the week without wastewater sampling
527 due to technical issues.

528

529 **Figure 2.** Detection of three variants using ICA-Var and Freyja. **(A)** Earlier detection of
530 EG.5, **(B)** HV.1, and **(C)** BA.2.86 made by the proposed method, as compared to Freyja.
531 In each panel, top and bottom rows plot the alternative allele frequencies at the dominant
532 mutation sites for samples at the first detection date made by the proposed method and
533 Freyja, respectively. X-axis represents each determinant mutation for the variant, y-axis
534 represents each individual sample at the first detection date, and the color represents the
535 alternative allele frequencies. In the y-axis, for each sample, if the Freyja pipeline outputs
536 an abundance, the abundance value will be listed after the sample name. An asterisk (*)
537 before the sample name indicates that the variant could be detected by this sample
538 following the proposed criteria.

539

540 **Figure 3.** Variant detection in urban and rural samples. **(A-B)** Detection of the emerging
541 variants of concerns (VoCs) in urban and rural samples using **(A)** Freyja and the **(B)** ICA-
542 Var pipeline. **(C)** Earliest date of variant detection in clinical cases from Southern Nevada;
543 first detection date with all wastewater samples (lightest grey), urban wastewater samples
544 (lighter grey) and rural wastewater samples (darker grey) using Freyja and the proposed
545 pipeline. A red asterisk (*) indicates the earliest detection dates among clinical reports,
546 Freyja, and the proposed method. If the variant was captured by the Freyja and ICA-Var
547 on the same earliest date, no * would be indicated. Earlier detection dates in rural samples
548 than urban samples are highlighted in bold.

549

550 **Figure 4.** Mutations with significant time-evolving contributions in the proposed method.

551 **(A)** Following the proposed method, 16 out of 25 determinant mutations in B.1.617.2 were
552 identified to maintain a significant time-evolving contributions to the group, with major
553 contributions in 2021 and early 2022. **(B)** The 25 dominant mutations in BA.1 were
554 identified to have significant time-evolving contributing to the group, with major
555 contributions in early 2022. **(C)** We identified 22 out of 25 determinant mutations in XBB.1
556 to have significant time-evolving contributions to the group, with major contributions after
557 2022/09. Similar contributing patterns were observed for several determinant mutations
558 (orange boxes).

559

560 **Figure 5.** Potential novel mutation patterns. **(A)** Hierarchical clustering leads to six
561 clusters at 113 potential novel mutation sites. Clusters 2, 3, 4, and 5 have overlapping
562 mutation sites with emerging variants EG.5, HV.1, and BA.2.86. Cluster 1 and 6 show no
563 overlapping mutation sites with known variants, and therefore, are more likely to give rise
564 to novel lineages. **(B)** Co-varying patterns of mutation sites in cluster 3 show major
565 fluctuating contributions after June 2023, consistent with EG.5 and HV.1 variants. **(C)** Co-
566 varying patterns of mutation sites in cluster 1, with major fluctuating contributions after
567 August 2023.

568

569 **Supplementary Figure 1.** Summary statistics of samples. **(A)** Independent component
570 analysis (ICA) source matrix (\mathcal{S}). **(B)** Most contributing mutations (top 5%) in each ICA
571 source ($\hat{\mathcal{S}}$). These mutation sites were focused on evaluating time-evolving contributions.

572 **(C).** Wastewater sample coverage at 50x depth from August 2021 to November 2023.
573 Each line represents a sequencing run and color indicates SARS-CoV-2 genome
574 coverage. **(D)** Number of weekly wastewater samples passing quality control (50x depth
575 covers >80% of SARS-CoV-2 genome) from August 2021 till November 2023. Urban and
576 rural wastewater samples are plotted separately.

577
578 **Supplementary Figure 2.** Early (or simultaneous) detection of VoCs **(A)** BA.2, BQ.1, **(B)**
579 BA.4, **(C)** BA.5, **(D)** BF.7, **(E)** XBB.1, XBB.1.5, **(F)** XBB.1.9, XBB.2.3, and **(G)** FL.1.5.1 in
580 the proposed independent component analysis pipeline (ICA). The x-axis represents each
581 determinant mutation for the variant and the y-axis represents each individual sample at
582 the first detection date. The colors represent the alternative allele frequencies. Looking at
583 the y-axis (for each sample), if Freyja outputs an abundance, the abundance value will
584 be listed after the sample name. An asterisk (*) before the sample name indicates that
585 the variant could be detected in the sample following the proposed criteria. For each VoC,
586 dominant mutation sites represented by deletions are highlighted in orange boxes.

587
588 **Supplementary Figure 3.** Manual cross-refencing of mutation sites in cluster 1 for Figure
589 5 with clinical sequencing reports in GISAID. According to clinical reports, all eight
590 mutations are currently not dominant in any variant.

591
592 **Supplementary Figure 4.** A comparison of detection matrices between Freyja and ICA-
593 Var. **(A)** Maximum Freyja abundance for samples in each week. Final detection criteria in
594 Figure 1C were set as maximum Freyja abundance greater than 15% (0.15). **(B)**

595 Detection matrices by the proposed ICA method based on three different criteria:
596 Spearman's correlation value (top panel), sensitivity (middle panel) and F1score (last
597 panel). **(C)** Hierarchical clustering results for six detection criteria computed from the
598 proposed methods. Jaccard index and F1 score are highly similar, sensitivity and area
599 under the ROC curve are highly correlated, and Spearman's correlation is a unique
600 criterion. Therefore, we utilized Spearman's correlation, sensitivity, and F1score to
601 establish our detection criteria in the ICA-Var pipeline (Figure 1B).

602

603 **Conflict of Interest Disclosures:** No disclosures to report.

604

605 **Acknowledgments**

606 VV, ECO are supported by NIH grants: GM103440 and MH109706 and a CARES Act
607 grant from the Nevada Governor's Office of Economic Development. VV, CL, DG, HK,
608 and ECO are supported by Grant Number NH75OT000057-01-00 from the Centers for
609 Disease Control and Prevention. The project contents are solely the responsibility of the
610 authors and do not necessarily represent the official views of the Centers for Disease
611 Control and Prevention. DPM was supported by the Nevada Water Resources Research
612 Institute/USGS under Grant/Cooperative Agreement No. G21AP10578 through the
613 Division of Hydrologic Sciences at Desert Research Institute. We would like to
614 acknowledge personnel at DRI and the collaborating wastewater agencies for their
615 assistance with sample logistics and data access. Special thanks to Daniel Fischer,
616 Latoya Blanche, LeAnna Risso and Alycia S. Ybarra; to Sean Twomey, James Eason, Bill
617 Coates, Brian Magna, Jose Rodriguez, George Veliz, Deborah Woodland; and to Chad

618 Marchand, Teresa Gomez, and Jeremy Singleton for contributing samples from
619 Wastewater Treatment Plants. We additionally thank Dr. Mira Han for her time reviewing
620 this manuscript.

621

622 References

- 623 1. Espinoza, B. *et al.* Coupled models of genomic surveillance and evolving pandemics with
624 applications for timely public health interventions. *Proc Natl Acad Sci U S A* **120**, (2023).
- 625 2. Li, J., Lai, S., Gao, G. F. & Shi, W. The emergence, genomic diversity and global spread
626 of SARS-CoV-2. *Nature* **600**, 408–418 (2021).
- 627 3. Ling-Hu, T., Rios-Guzman, E., Lorenzo-Redondo, R., Ozer, E. A. & Hultquist, J. F.
628 Challenges and Opportunities for Global Genomic Surveillance Strategies in the COVID-
629 19 Era. *Viruses* **14**, (2022).
- 630 4. Robishaw, J. D. *et al.* Genomic surveillance to combat COVID-19: challenges and
631 opportunities. *Lancet Microbe* **2**, e481–e484 (2021).
- 632 5. Cuadros, D. F., Branscum, A. J., Mukandavire, Z., Miller, F. D. W. & MacKinnon, N.
633 Dynamics of the COVID-19 epidemic in urban and rural areas in the United States. *Ann*
634 *Epidemiol* **59**, 16–20 (2021).
- 635 6. Paul, R., Arif, A. A., Adeyemi, O., Ghosh, S. & Han, D. Progression of COVID-19 From
636 Urban to Rural Areas in the United States: A Spatiotemporal Analysis of Prevalence
637 Rates. *J Rural Health* **36**, 591–601 (2020).
- 638 7. Souch, J. M. & Cossman, J. S. A Commentary on Rural-Urban Disparities in COVID-19
639 Testing Rates per 100,000 and Risk Factors. *The Journal of Rural Health* **37**, 188 (2021).
- 640 8. Fontenele, R. S. *et al.* High-throughput sequencing of SARS-CoV-2 in wastewater
641 provides insights into circulating variants. *Water Res* **205**, 117710 (2021).
- 642 9. Gerrity, D., Papp, K., Stoker, M., Sims, A. & Frehner, W. Early-pandemic wastewater
643 surveillance of SARS-CoV-2 in Southern Nevada: Methodology, occurrence, and
644 incidence/prevalence considerations. *Water Res X* **10**, (2021).
- 645 10. Medema, G., Been, F., Heijnen, L. & Petterson, S. Implementation of environmental
646 surveillance for SARS-CoV-2 virus to support public health decisions: Opportunities and
647 challenges. *Curr Opin Environ Sci Health* **17**, 49–71 (2020).
- 648 11. Ahmed, W. *et al.* First confirmed detection of SARS-CoV-2 in untreated wastewater in
649 Australia: A proof of concept for the wastewater surveillance of COVID-19 in the
650 community. *Sci Total Environ* **728**, (2020).

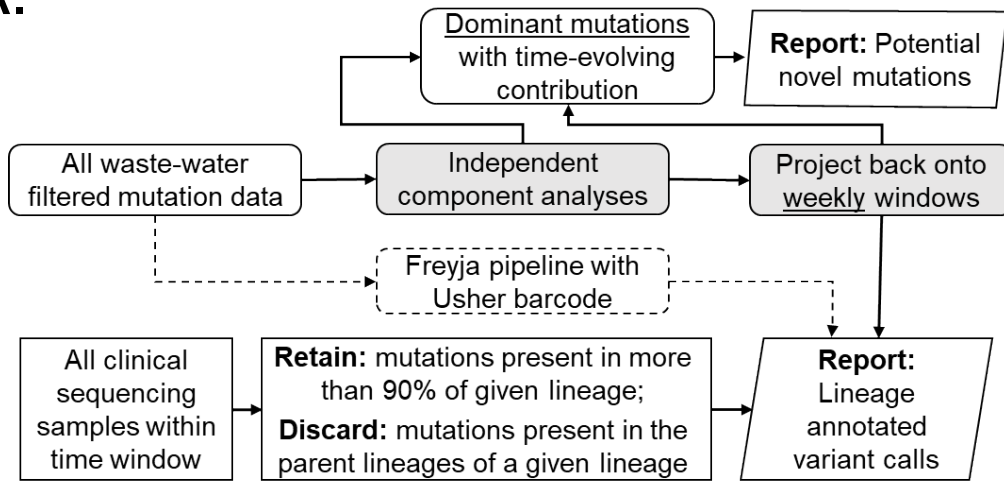
- 651 12. Hart, O. E. & Halden, R. U. Computational analysis of SARS-CoV-2/COVID-19
652 surveillance by wastewater-based epidemiology locally and globally: Feasibility, economy,
653 opportunities and challenges. *Sci Total Environ* **730**, (2020).
- 654 13. Crank, K., Chen, W., Bivins, A., Lowry, S. & Bibby, K. Contribution of SARS-CoV-2 RNA
655 shedding routes to RNA loads in wastewater. *Sci Total Environ* **806**, (2022).
- 656 14. Wölfel, R. *et al.* Virological assessment of hospitalized patients with COVID-2019. *Nature*
657 **581**, 465–469 (2020).
- 658 15. Jakariya, M. *et al.* Wastewater-based epidemiological surveillance to monitor the
659 prevalence of SARS-CoV-2 in developing countries with onsite sanitation facilities.
660 *Environ Pollut* **311**, (2022).
- 661 16. Stockdale, S. R. *et al.* RNA-Seq of untreated wastewater to assess COVID-19 and
662 emerging and endemic viruses for public health surveillance. *The Lancet regional health.*
663 *Southeast Asia* **14**, (2023).
- 664 17. Betancourt, W. Q. *et al.* COVID-19 containment on a college campus via wastewater-
665 based epidemiology, targeted clinical testing and an intervention. *Sci Total Environ* **779**,
666 (2021).
- 667 18. Vo, V. *et al.* Use of wastewater surveillance for early detection of Alpha and Epsilon
668 SARS-CoV-2 variants of concern and estimation of overall COVID-19 infection burden.
669 *Science of the Total Environment* **835**, (2022).
- 670 19. Harrington, A. *et al.* Environmental Surveillance of Flood Control Infrastructure Impacted
671 by Unsheltered Individuals Leads to the Detection of SARS-CoV-2 and Novel Mutations
672 in the Spike Gene. *Environ Sci Technol Lett* (2024) doi:10.1021/ACS.ESTLETT.3C00938.
- 673 20. Kayikcioglu, T. *et al.* Performance of methods for SARS-CoV-2 variant detection and
674 abundance estimation within mixed population samples. *PeerJ* **11**, (2023).
- 675 21. Amman, F. *et al.* Viral variant-resolved wastewater surveillance of SARS-CoV-2 at
676 national scale. *Nat Biotechnol* **40**, 1814–1822 (2022).
- 677 22. Brunner, F. S. *et al.* City-wide wastewater genomic surveillance through the successive
678 emergence of SARS-CoV-2 Alpha and Delta variants. *Water Res* **226**, 119306 (2022).
- 679 23. Vo, V. *et al.* Detection of the Omicron BA.1 Variant of SARS-CoV-2 in Wastewater From a
680 Las Vegas Tourist Area. *JAMA Netw Open* **6**, E230550 (2023).

- 681 24. Harrington, A. *et al.* Urban monitoring of antimicrobial resistance during a COVID-19
682 surge through wastewater surveillance. *Sci Total Environ* **853**, (2022).
- 683 25. Vo, V. *et al.* Identification of a rare SARS-CoV-2 XL hybrid variant in wastewater and the
684 subsequent discovery of two infected individuals in Nevada. *Sci Total Environ* **858**,
685 (2023).
- 686 26. Crits-Christoph, A. *et al.* Genome sequencing of sewage detects regionally prevalent
687 SARS-CoV-2 variants. *mBio* **12**, 1–9 (2021).
- 688 27. Wurtz, N. *et al.* Monitoring the Circulation of SARS-CoV-2 Variants by Genomic Analysis
689 of Wastewater in Marseille, South-East France. *Pathogens* **10**, (2021).
- 690 28. Izquierdo-Lara, R. *et al.* Monitoring SARS-CoV-2 Circulation and Diversity through
691 Community Wastewater Sequencing, the Netherlands and Belgium. *Emerg Infect Dis* **27**,
692 1405–1415 (2021).
- 693 29. Bar-Or, I. *et al.* Detection of SARS-CoV-2 variants by genomic analysis of wastewater
694 samples in Israel. *Sci Total Environ* **789**, (2021).
- 695 30. Rainey, A. L. *et al.* Wastewater surveillance for SARS-CoV-2 in a small coastal
696 community: Effects of tourism on viral presence and variant identification among low
697 prevalence populations. *Environ Res* **208**, (2022).
- 698 31. Pérez-Cataluña, A. *et al.* Spatial and temporal distribution of SARS-CoV-2 diversity
699 circulating in wastewater. *Water Res* **211**, (2022).
- 700 32. Gregory, D. A., Wieberg, C. G., Wenzel, J., Lin, C. H. & Johnson, M. C. Monitoring sars-
701 cov-2 populations in wastewater by amplicon sequencing and using the novel program
702 sam refiner. *Viruses* **13**, (2021).
- 703 33. Jahn, K. *et al.* Early detection and surveillance of SARS-CoV-2 genomic variants in
704 wastewater using COJAC. *Nature Microbiology* **2022 7:8 7**, 1151–1160 (2022).
- 705 34. Karthikeyan, S. *et al.* Wastewater sequencing reveals early cryptic SARS-CoV-2 variant
706 transmission. *Nature* **609**, 101–108 (2022).
- 707 35. Sapoval, N. *et al.* Enabling accurate and early detection of recently emerged SARS-CoV-
708 2 variants of concern in wastewater. *Nature Communications* **2023 14:1 14**, 1–7 (2023).
- 709 36. Valieris, R. *et al.* A mixture model for determining SARS-Cov-2 variant composition in
710 pooled samples. *Bioinformatics* **38**, 1809–1815 (2022).

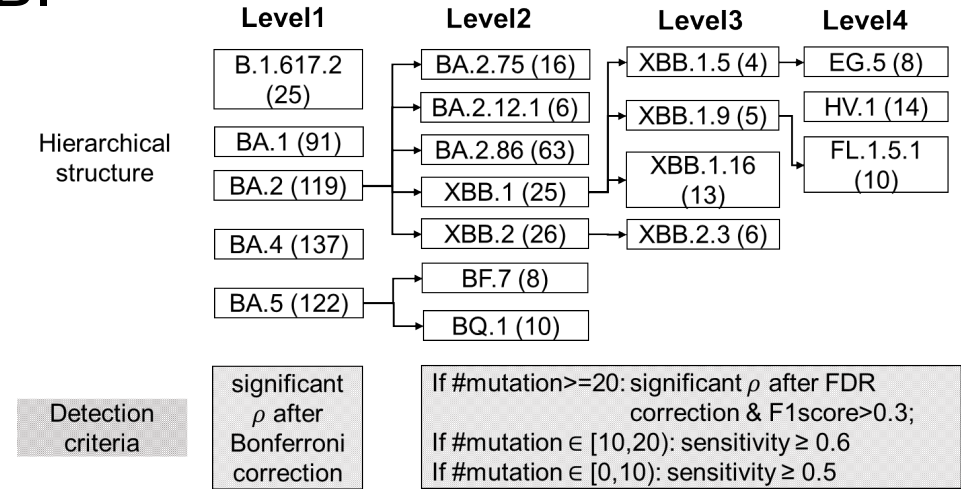
- 711 37. Posada-Cespedes, S. *et al.* V-pipe: a computational pipeline for assessing viral genetic
712 diversity from high-throughput data. *Bioinformatics* **37**, 1673–1680 (2021).
- 713 38. Wilm, A. *et al.* LoFreq: a sequence-quality aware, ultra-sensitive variant caller for
714 uncovering cell-population heterogeneity from high-throughput sequencing datasets.
715 *Nucleic Acids Res* **40**, 11189–11201 (2012).
- 716 39. Grubaugh, N. D. *et al.* An amplicon-based sequencing framework for accurately
717 measuring intrahost virus diversity using PrimalSeq and iVar. *Genome Biol* **20**, 1–19
718 (2019).
- 719 40. Singer, A. C. *et al.* A world of wastewater-based epidemiology. *Nature Water* **1**, 408–415
720 (2023).
- 721 41. Yamasoba, D. *et al.* Virological characteristics of the SARS-CoV-2 omicron XBB.1.16
722 variant. *Lancet Infect Dis* **23**, 655–656 (2023).
- 723 42. Araf, Y. *et al.* Omicron variant of SARS-CoV-2: Genomics, transmissibility, and responses
724 to current COVID-19 vaccines. *J Med Virol* **94**, 1825–1832 (2022).
- 725 43. Mlcochova, P. *et al.* SARS-CoV-2 B.1.617.2 Delta variant replication and immune
726 evasion. *Nature* **599**, 114–119 (2021).
- 727 44. Hyvärinen, A., Karhunen, J. & Oja, E. Independent Component Analysis. *Appl Comput*
728 *Harmon Anal* **21**, 135–144 (2001).
- 729 45. Beckmann, C., Mackay, C., Filippini, N. & Smith, S. Group comparison of resting-state
730 FMRI data using multi-subject ICA and dual regression. *Neuroimage* **47**, S148 (2009).
- 731 46. Groves, A. R., Beckmann, C. F., Smith, S. M. & Woolrich, M. W. Linked independent
732 component analysis for multimodal data fusion. *Neuroimage* **54**, 2198–2217 (2011).
- 733 47. Beckmann, C. F., DeLuca, M., Devlin, J. T. & Smith, S. M. Investigations into resting-state
734 connectivity using independent component analysis. *Philosophical Transactions of the*
735 *Royal Society B: Biological Sciences* **360**, 1001–1013 (2005).
- 736 48. Calhoun, V. D., Adali, T., Pearlson, G. D. & Pekar, J. J. A Method for Making Group
737 Inferences from Functional MRI Data Using Independent Component Analysis. (2001)
738 doi:10.1002/hbm.

- 739 49. Hasing, M. E. *et al.* Wastewater surveillance monitoring of SARS-CoV-2 variants of
740 concern and dynamics of transmission and community burden of COVID-19. *Emerg*
741 *Microbes Infect* **12**, (2023).
- 742 50. Reynolds, L. J. *et al.* SARS-CoV-2 variant trends in Ireland: Wastewater-based
743 epidemiology and clinical surveillance. *Science of The Total Environment* **838**, 155828
744 (2022).
- 745 51. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing
746 reads. *EMBnet J* **17**, 10–12 (2011).
- 747 52. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler
748 transform. *Bioinformatics* **25**, 1754–1760 (2009).
- 749 53. Wu, F. *et al.* A new coronavirus associated with human respiratory disease in China.
750 *Nature* **2020 579:7798 579**, 265–269 (2020).
- 751 54. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–
752 2079 (2009).
- 753 55. Himberg, J. & Hyvärinen, A. ICASSO: Software for investigating the reliability of ICA
754 estimates by clustering and visualization. *Neural Networks for Signal Processing -*
755 *Proceedings of the IEEE Workshop 2003-Janua*, 259–268 (2003).
- 756
- 757

A.



B.



C.

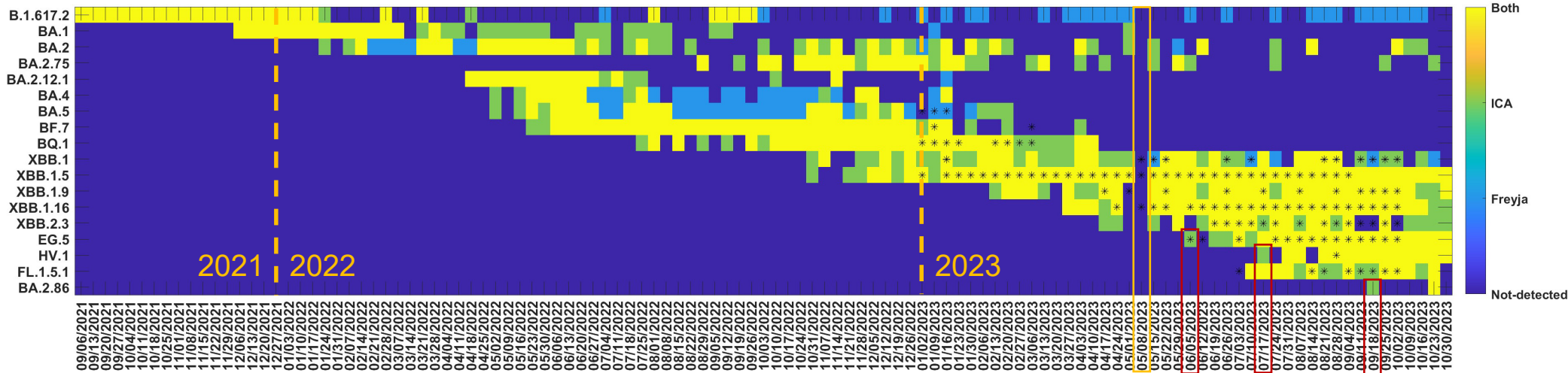
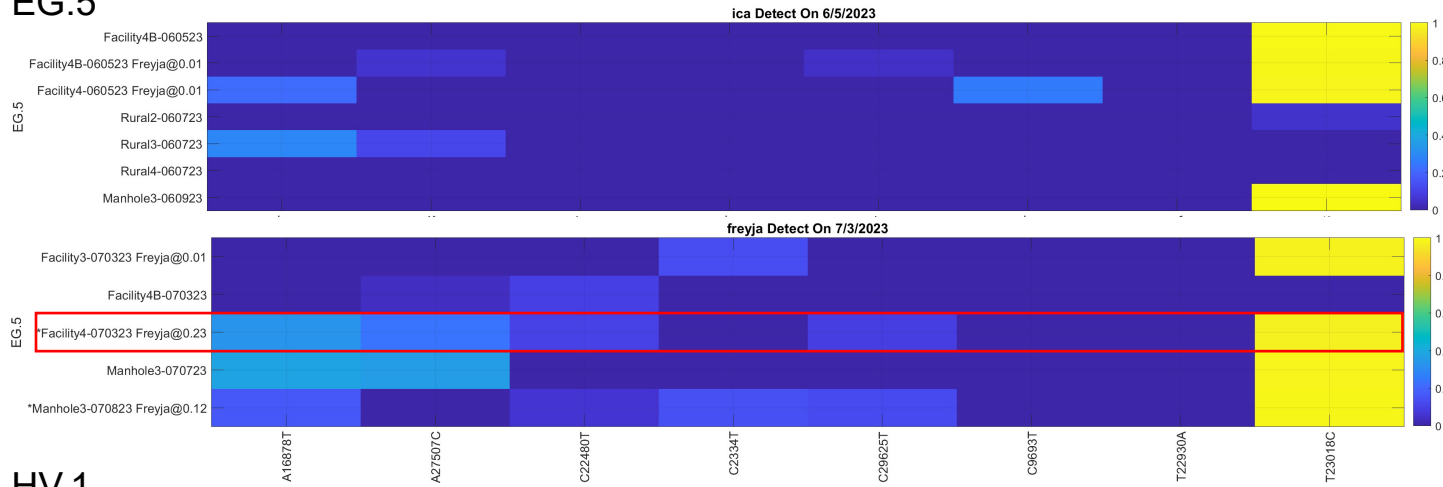
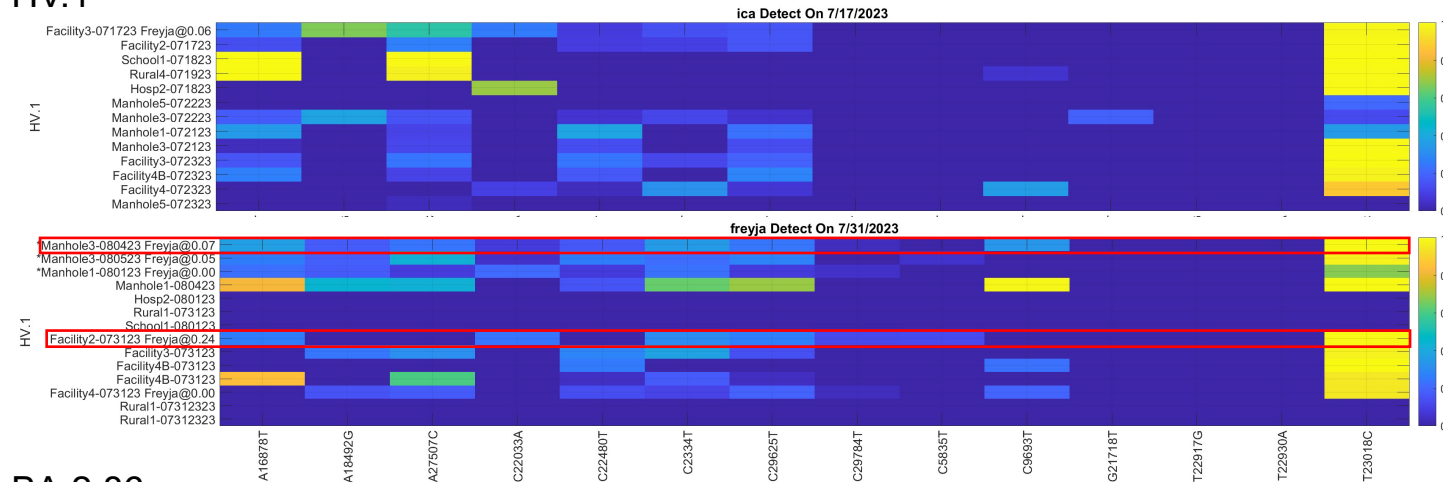


Figure 1

A. EG.5



B. HV.1



C. BA.2.86

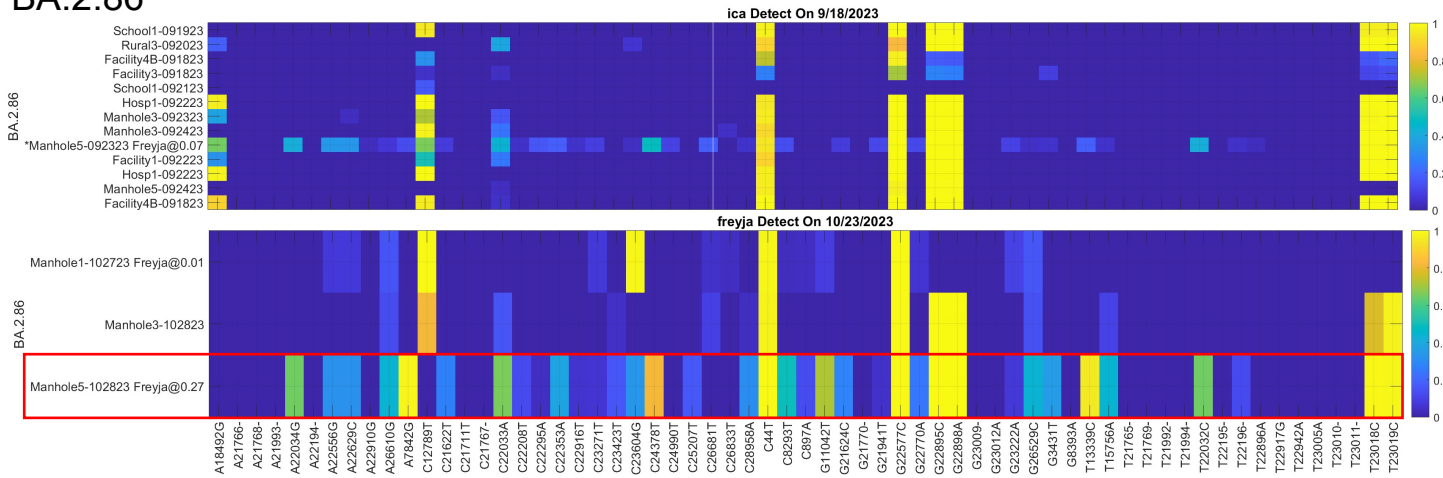
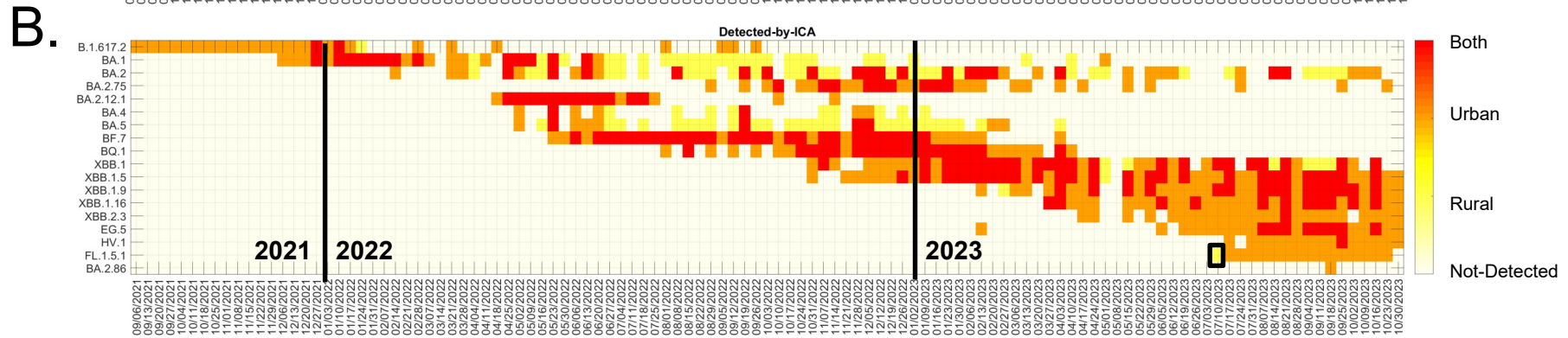
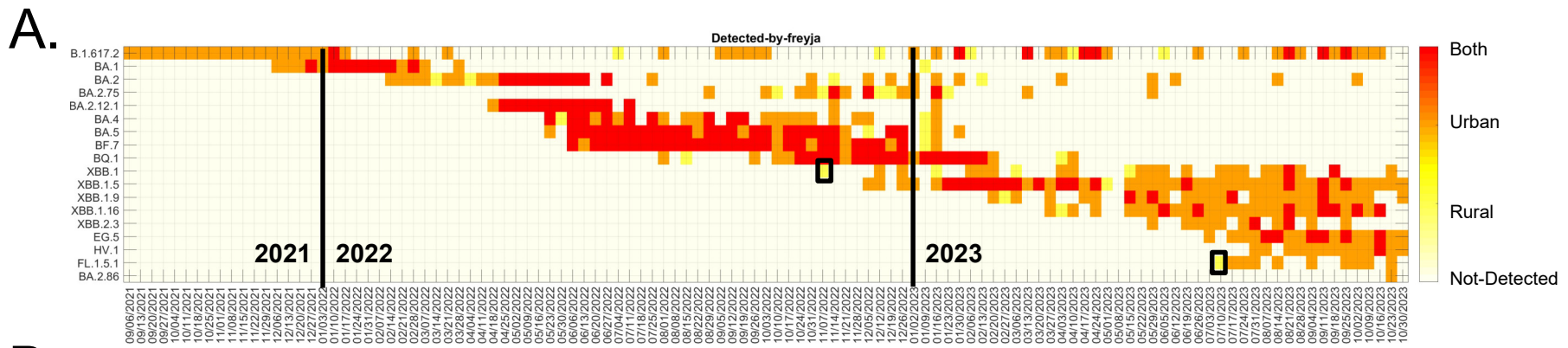


Figure 2

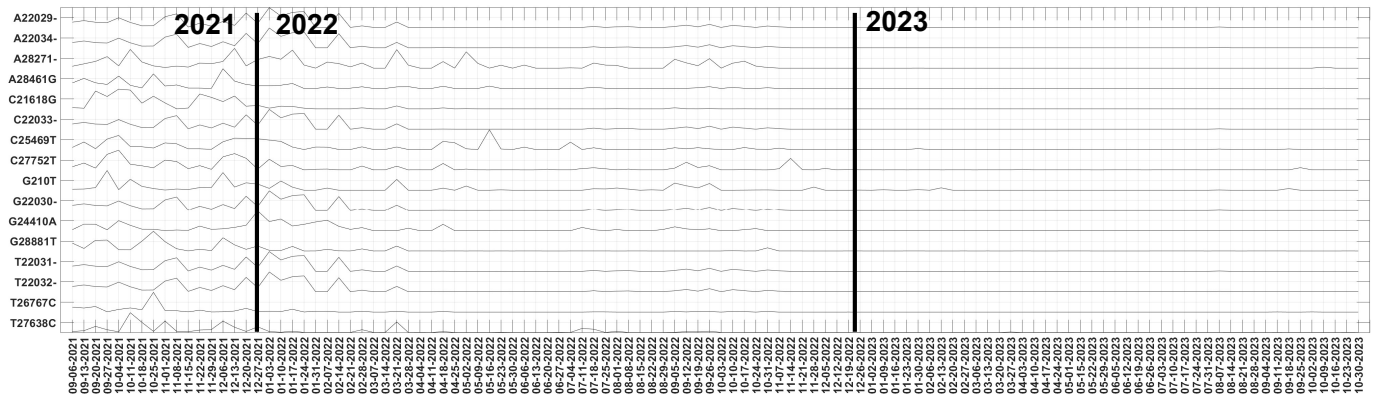


C.

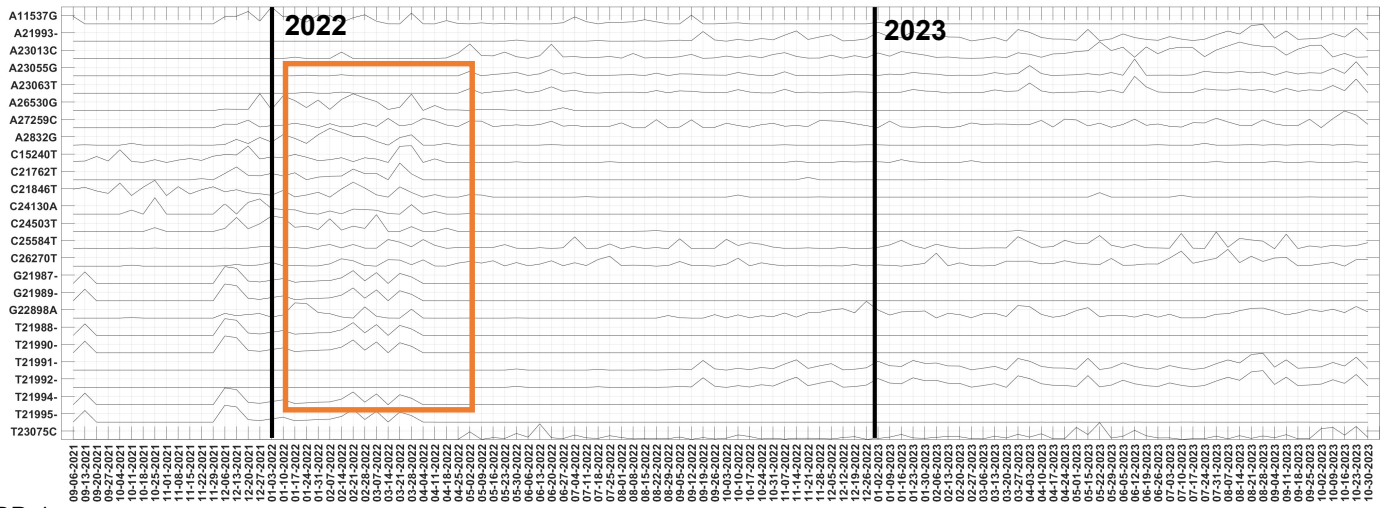
Variants of Interest	First clinical report	First Detection date by Freyja			First Detection date by ICA		
		In All samples	In Urban samples	In Rural Samples	In All samples	In Urban samples	In Rural samples
B.1.617.2		09/06/2021	09/06/2021	1/10/2022	09/06/2021	09/06/2021	12/27/2021
BA.1		12/6/2021	12/6/2021	12/27/2021	12/6/2021	12/6/2021	12/27/2021
BA.2		2/14/2022	2/14/2022	3/14/2022	1/24/2022*	2/14/2022	4/4/2022
BA.2.75		8/29/2022	8/29/2022	10/10/2022	8/29/2022	8/29/2022	10/10/2022
BA.2.12.1		4/18/2022	4/18/2022	4/25/2022	4/18/2022	4/18/2022	4/25/2022
BA.4		5/23/2022	5/23/2022	5/23/2022	5/2/2022*	5/2/2022	5/23/2022
BA.5	1/2/2023	5/23/2022	5/23/2022	6/6/2022	5/2/2022*	5/2/2022	5/16/2022
BF.7	1/9/2023	6/6/2022	6/6/2022	6/6/2022	5/23/2022*	5/23/2022	6/20/2022
BQ.1	1/2/2023	8/1/2022	8/1/2022	8/15/2022	7/25/2022*	8/1/2022	8/15/2022
XBB.1	1/16/2023	11/7/2022	12/12/2022	11/7/2022	10/31/2022*	10/31/2022	11/7/2022
XBB.1.5	1/2/2023	12/5/2022	12/5/2022	1/23/2023	10/31/2022*	10/31/2022	12/26/2022
XBB.1.9	4/17/2023	2/20/2023	2/20/2023	2/27/2023	2/13/2023*	2/13/2023	2/27/2023
XBB.1.16	4/24/2023	3/27/2023	3/27/2023	4/3/2023	3/27/2023	3/27/2023	3/27/2023
XBB.2.3	6/19/2023	5/29/2023	5/29/2023	7/31/2023	4/17/2023*	4/17/2023	8/21/2023
EG.5	6/5/2023*	7/3/2023	7/3/2023	8/7/2023	6/5/2023*	2/13/2023	8/7/2023
HV.1	8/28/2023	7/31/2023	7/31/2023	10/16/2023	7/17/2023*	7/17/2023	9/25/2023
FL.1.5.1	7/3/2023*	7/10/2023	7/17/2023	7/10/2023	7/10/2023	7/17/2023	7/10/2023
BA.2.86		10/23/2023	10/23/2023		9/18/2023*	9/18/2023	

Figure 3

A. B.1.617.2



B. BA.1



C. XBB.1

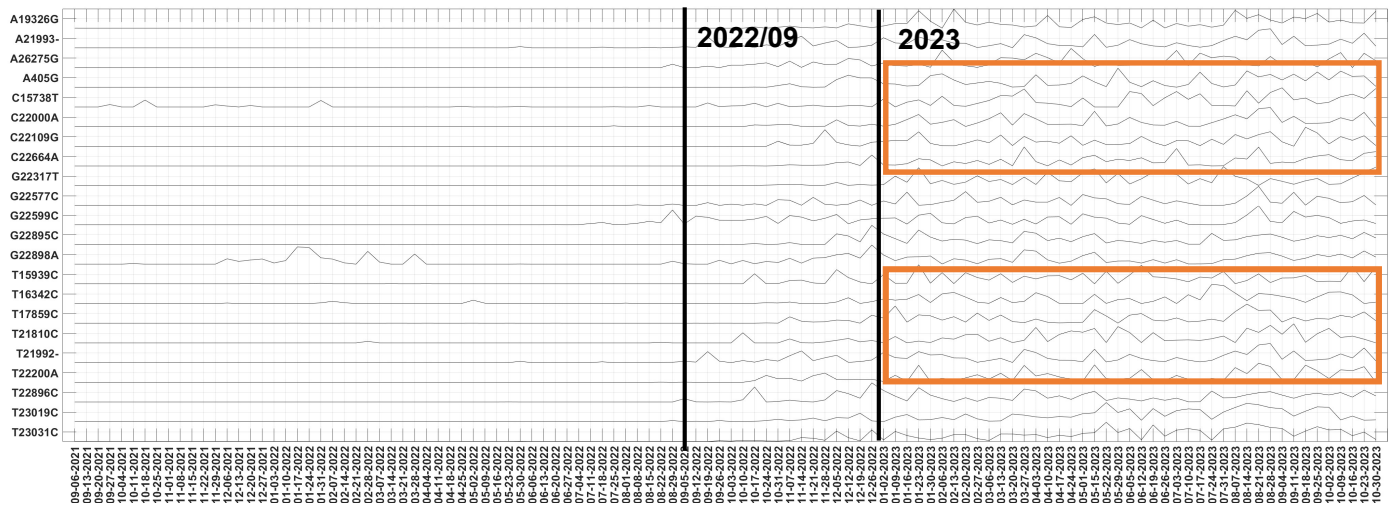
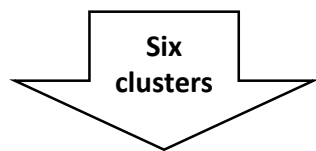
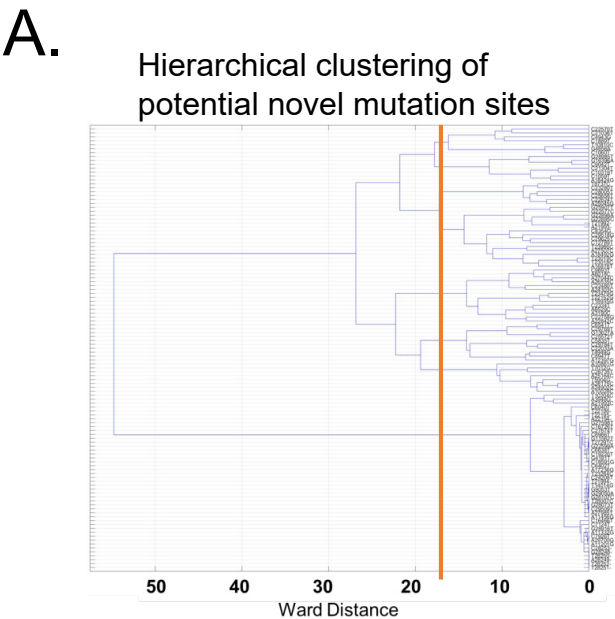
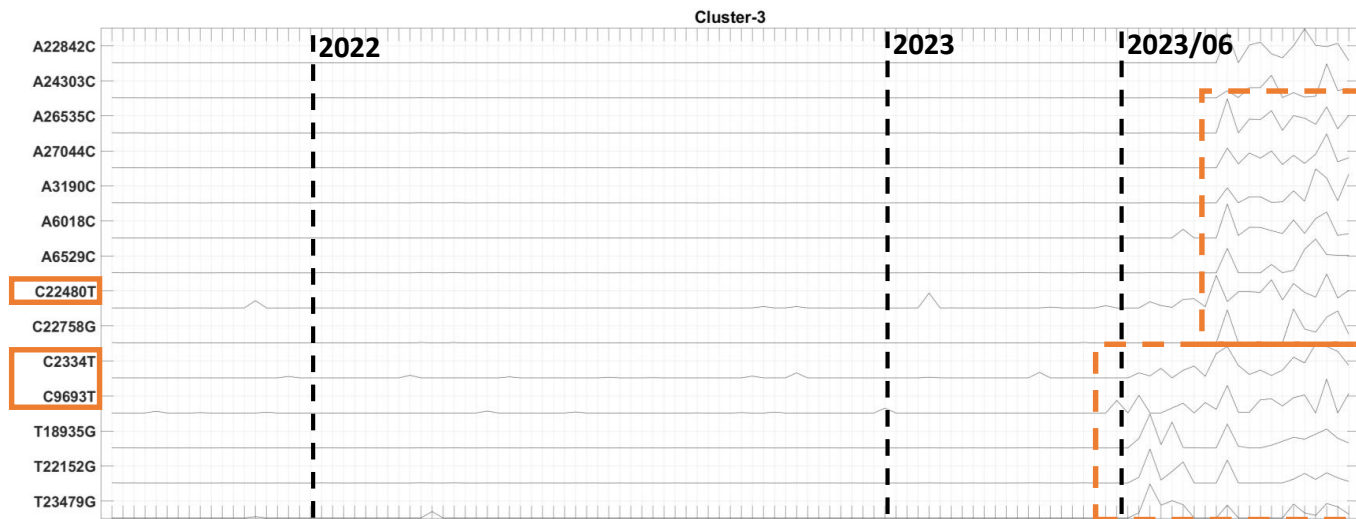


Figure 4



	cluster -1	cluster -2	cluster -3	cluster -4	cluster -5	cluster -6
EG.5	0	4	3	0	0	0
HV.1	0	5	3	0	3	0
BA.2.86	0	9	0	4	1	0

B. Cluster-3: Overlapping with 3 dominant mutations in EG.5 and HV.1



C. Cluster-1: No overlap with known emerging lineages

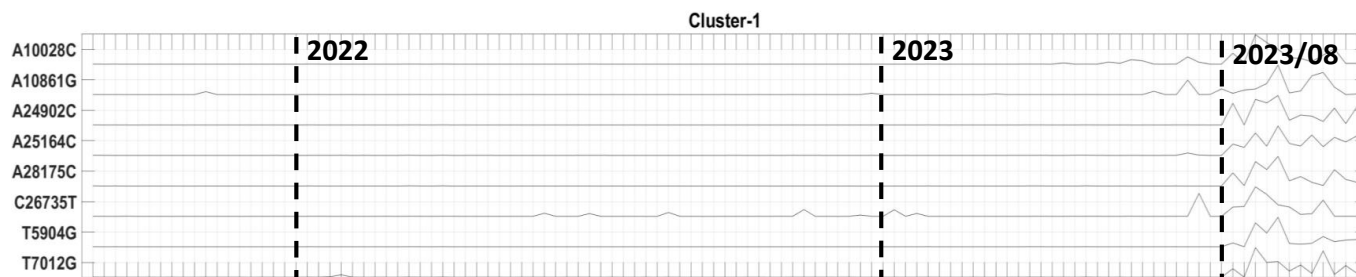


Figure 5