

1 Exploring the Spatial Distribution of Persistent 2 SARS-CoV-2 Mutations - Leveraging mobility 3 data for targeted sampling

4 Riccardo Spott^{1,*}, Mathias W. Pletz^{1,2}, Carolin Fleischmann-Struzek^{1,2}, Aurelia Kimmig¹,
5 Christiane Hadlich³, Mathias Hauert³, Mara Lohde¹, Mateusz Jundzill¹, Mike Marquet¹, Petra
6 Dickmann⁴, Ruben Schüchner⁵, Martin Hölzer⁶, Denise Kühnert^{7,8}, Christian Brandt¹

7
8 ¹ Institute for Infectious Diseases and Infection Control, Jena University Hospital, Jena,
9 Germany

10 ² Center for Sepsis Control and Care, Jena University Hospital/Friedrich Schiller University
11 Jena, Jena, Germany

12 ³ SMA Development GmbH - epicinsights Agentur für Künstliche Intelligenz und Big Data
13 Analytics, Jena, Germany

14 ⁴ Department of Anaesthesiology and Intensive Care, Jena University Hospital, Jena,
15 Germany

16 ⁵ Thuringian State Authority for Consumer Protection, Department Health Protection

17 ⁶ Methodology and Research Infrastructure, Genome Competence Center (MF1), Robert
18 Koch Institute, Berlin, Germany

19 ⁷ Centre for Artificial Intelligence in Public Health Research, Robert Koch Institute, Berlin,
20 Germany

21 ⁸ Transmission, Infection, Diversification and Evolution Group, Max Planck Institute for
22 Geoanthropology, 07745 Jena, Germany.

23 * Corresponding Author

24 Short Title

25 Mobility data for Surveillance

26 Keywords

27 Nanopore sequencing; SARS-CoV-2; WGS; mobility data; cluster tracking

28 Abstract

29 Given the rapid cross-country spread of SARS-CoV-2 and the resulting difficulty in tracking
30 lineage spread, we investigated the potential of combining mobile service data and fine-
31 granular metadata (such as postal codes and genomic data) to advance integrated genomic
32 surveillance of the pandemic in the federal state of Thuringia, Germany. We sequenced over
33 6,500 SARS-CoV-2 Alpha genomes (B.1.1.7) across seven months within Thuringia while
34 collecting patients' isolation dates and postal codes. Our dataset is complemented by over
35 66,000 publicly available German Alpha genomes and mobile service data for Thuringia. We
36 identified the existence and spread of nine persistent mutation variants within the Alpha
37 lineage, seven of which formed separate phylogenetic clusters with different spreading
38 patterns in Thuringia. The remaining two are sub-clusters. Mobile service data can indicate
39 these clusters' spread and highlight a potential sampling bias, especially of low-prevalence
40 variants. Thereby, mobile service data can be used either retrospectively to assess
41 surveillance coverage and efficiency from already collected data or to actively guide part of a
42 surveillance sampling process to districts where these variants are expected to emerge. The
43 latter concept was successfully implemented as a proof-of-concept for a mobility-guided
44 sampling strategy in response to the surveillance of Omicron sublineage BQ.1.1. The
45 combination of mobile service data and SARS-CoV-2 surveillance by genome sequencing is
46 a valuable tool for more targeted and responsive surveillance.

47 Introduction

48 On March 11, 2020, the World Health Organization (WHO) classified the SARS-CoV-2 virus
49 (Severe Acute Respiratory Syndrome - Corona Virus - 2) as a global pandemic due to its
50 rapid spread and high infection rate.¹ The airborne virus has since caused significant
51 morbidity and mortality worldwide (<https://covid19.who.int/>). In an attempt to control its
52 spread, many countries initiated comprehensive surveillance efforts with molecular
53 techniques such as Polymerase Chain Reaction (PCR) and Whole Genome Sequencing
54 (WGS).^{2,3} Consequently, nearly 15.8 million SARS-CoV-2 sequences have been deposited
55 into the "Global Initiative on Sharing All Influenza Data" database (as of July 21, 2023,
56 GISAID). Many research groups have undertaken studies examining the viral spread by
57 integrating sequencing and epidemiological data to monitor the pandemic and investigate
58 local outbreaks.^{4,5} Most of these local projects are part of national surveillance programs
59 such as the UK's Genomics Consortium (COG-UK) or "national genomic surveillance" in the
60 USA.^{2,6} In Germany, the "Coronavirus-Surveillanceverordnung" (CorSurV) enacted by the
61 State Ministry of Health on January 19, 2021, mandated that laboratories with sequencing
62 capabilities process SARS-CoV-2-positive samples, offering financial compensation until
63 April 2023.³

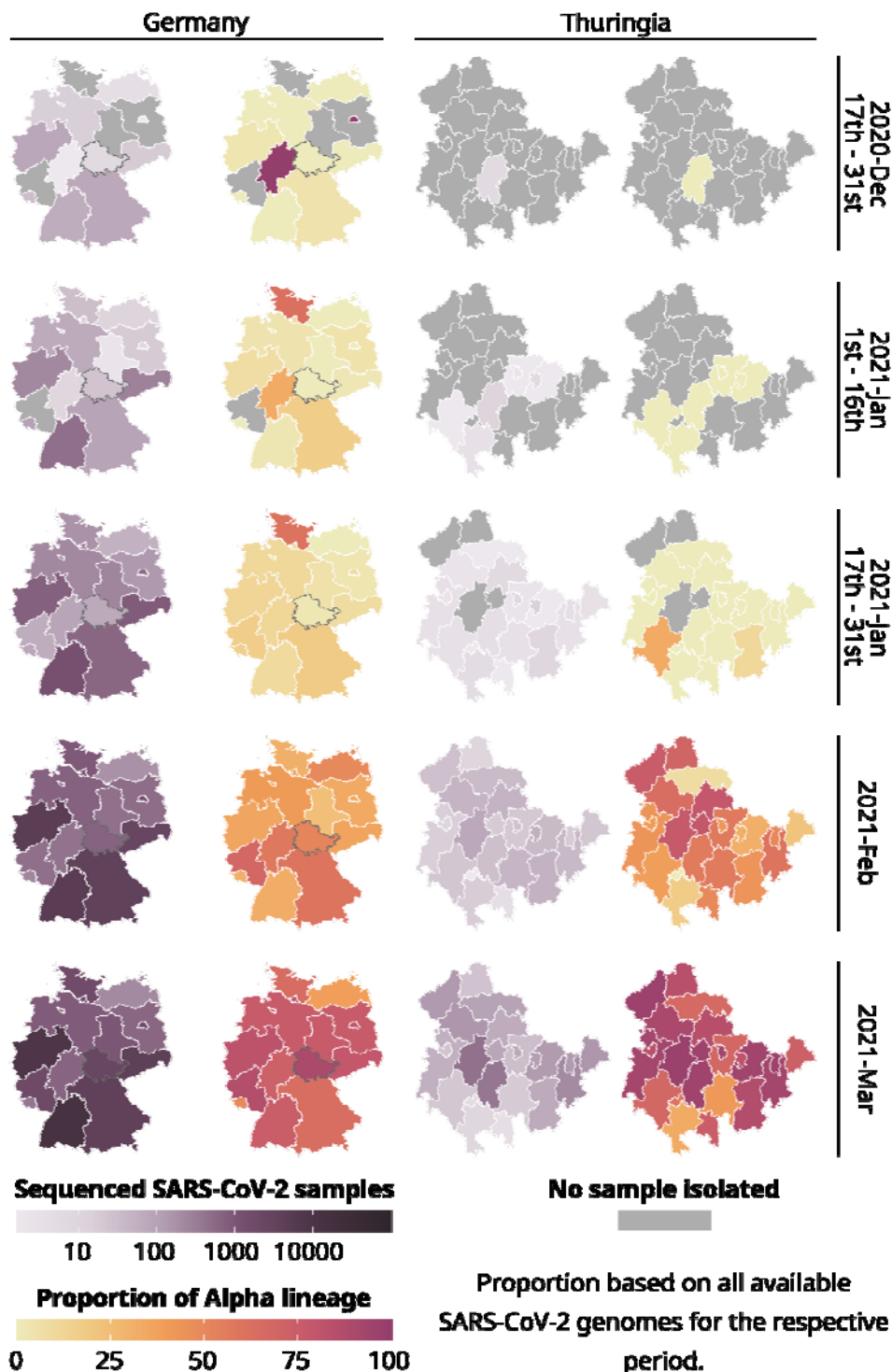
64 Bioinformatics workflows developed in Germany, such as poreCov (for Oxford Nanopore
65 data) and CoVpipe2 (for Illumina data), reconstruct SARS-CoV-2 consensus genomes from
66 the sequencing data and prepare the results for upload and submission to the Robert Koch
67 Institute (RKI).^{7,8} As the German government's public health and biomedical research
68 institute responsible for disease control and prevention, the RKI collected the genomes via
69 the German Electronic Sequence Data Hub (DESH) and integrated them with additional
70 epidemiological information to provide an up-to-date overview of the ongoing viral spread.
71 For keeping track of the rapid SARS-CoV-2 evolution, PANGO (Phylogenetic Assignment of
72 Named Global Outbreak) provides a standard naming convention based on unique mutation
73 profiles and further criteria, resulting in the classification of over 3,660 lineages (as of August
74 2023).^{9,10} Additionally, the WHO classified important viral lineages as "Variants of Concern"
75 (VOC), "Variants of Interest" (VOI), or "Variants under Monitoring" (VUM), using Greek
76 designations in the past (e.g., "Alpha" (Pango lineage main designation B.1.1.7) or
77 "Omicron" (Pango lineage main designation B.1.1.529)). Further, the WHO also de-
78 escalated former VOCs to reflect the current SARS-CoV-2 variant landscape better. The first
79 defined VOC (now de-escalated), the Alpha lineage, rapidly replaced almost all previously
80 circulating lineages globally by the end of 2020 until the VOC Delta (main lineage B.1.617.2)
81 replaced it in mid-2021.¹¹⁻¹³

82 To predict or monitor the rapid viral spread throughout regions, various data types, like travel
83 data, passenger volumes, or passive wastewater monitoring were examined previously.¹⁴⁻¹⁶
84 Furthermore, different studies explored mobility data with genomic data to retrace the origin
85 and spatial expanse of Alpha or utilized geolocation data to model the spread in metropolitan
86 areas to recreate case trajectories and the impact of mobility restrictions.^{17,18} Mobility data
87 was also used in Germany during the pandemic, revealing that lockdowns leave distant parts
88 of the country less connected due to the sharp decline in long-distance travel.¹⁹ These
89 studies focused on analyzing residential movement and contact tracing to evaluate and
90 inform health policies but were not applied to active molecular surveillance.
91 Here, we investigated whether mobile service data and fine-granular metadata (such as
92 postal codes and genomic data) can help predict the spread of the Alpha lineage or guide
93 the sampling for more targeted genomic surveillance with a focus on the German federal
94 state of Thuringia.

95 Results

96 The Alpha lineage spread rapidly through Thuringia, showing a pattern
97 similar to its nationwide spread

98 Thuringia is a rural federal state in central Germany with a population of 2.1 million and no
99 major airports (overview of Thuringia's population density in Supplementary Figure S1). We
100 investigated if the spread of the Alpha lineage of SARS-CoV-2 behaved differently compared
101 to the whole of Germany. To understand its spread, we used 289,487 public SARS-CoV-2
102 genomes from Germany (excluding Thuringia; including 137,024 Alpha genomes) and 7,394
103 genomes from our own sequencing data for Thuringia (including 6,307 Alpha genomes) to
104 track Alpha's occurrence from December 2020 to August 2021 (see Figure 1,
105 Supplementary Figure S2, and Supplementary Tables S1 and S2; for details, see Methods
106 section "Alpha spread datasets"). For Thuringia, district-level data (full postal code) per
107 genome were available, whereas, for Germany, only postal code data of the sending
108 laboratories (referred to as "primary diagnostic laboratory" by the RKI where the SARS-CoV-
109 2 positive sample was detected) and sequencing laboratories were publicly available.



110 **Figure 1: Total number of all sequenced SARS-CoV-2 samples (purple) and the proportion of**
 111 **the Alpha lineage for all sequenced samples (yellow-red) for each state of Germany and each**
 112 **district of Thuringia. 289,487 publicly available German SARS-CoV-2 genomes and their metadata**
 113 **were used for the general German maps, excluding data from Thuringia. For Thuringia, we always**
 114 **used 7,394 genomes and their metadata from our database for the German and Thuringian maps.**
 115 **Please note that for all states except Thuringia, we used the postal code of the sending laboratory as**
 116 **a proxy for the geographic location of a sample. A gray border on the maps of Germany highlights**
 117 **Thuringia.**
 118

119 In late December 2020, six federal states in Germany (from here on called states) reported
120 the first cases of the Alpha variant. Although sequencing was initially low, it gradually
121 increased in the following month. However, the Corona-Surveillance regulation was passed
122 at the end of January 2021, leading to a rapid increase in sampling and sequencing by
123 February since sequencing costs could be reimbursed. Even though Thuringia sequenced a
124 similar amount of SARS-CoV-2 samples compared to other German states (as shown in
125 Figure 1), the proportion of the Alpha variant to other lineages was relatively low. However,
126 the proportion of Alpha increased heavily in February.

127 By March, Alpha had spread to nearly all states and districts (districts are similar to counties
128 or provinces) in Germany (Median: 76.47 % Alpha samples among a federal states total
129 sequenced samples compared to 36.03 % in February, excluding Thuringia) and Thuringia
130 (Median: 85.29 %, up from 50.00 % in February). So, there was no noticeable difference in
131 the Alpha proportions between Germany and Thuringia after February. During the summer
132 of June and July 2021, sequencing declined in Germany (including Thuringia;
133 Supplementary Figure S2) due to the decrease in overall daily cases, as reported by
134 Meinrup *et al.* and Oh and Hölzer *et al.*^{20,21}

135 In summary, the spread of the Alpha lineage in Thuringia lagged roughly two weeks behind
136 the general spread of other German federal states but showed similar proportions. This
137 suggests that Thuringia experienced a delay in the initial arrival of Alpha. However, we did
138 not observe any difference in the overall spread afterward. Thuringia was among the first
139 states to adopt new containment measures, including contact limitations, closure of retail
140 shops, and prohibition of tourist journeys (14th December 2020). Jena, a city in Thuringia,
141 was also the first German city to implement mandatory public masking in March 2020.²²
142 Contacts were further restricted on January 9th, and people were urged to restrict their
143 movement radius to 15 km, which might explain the delay besides the absence of major
144 airports nearby.

145 All Thuringian genomes were evenly distributed between other German samples in the
146 phylogenetic time tree (see Supplementary Figure S3). However, due to its rapid spread
147 from February onwards, it is difficult to accurately track how the Alpha lineage specifically
148 expanded (point of entries, exact origins, etc.). Consequently, we investigated whether
149 "sublineages" might be identifiable and trackable to address this.

150 Monitoring of Alpha subclusters in Thuringia reveals temporally and 151 regionally restricted distribution patterns

152 To identify possible clusters among the Alpha lineage spreading in Thuringia, we called each
153 Alpha genome's mutations via Nextclade by analyzing them using poreCov.^{7,23} We identified

154 nine clusters out of 70,429 Alpha genomes, based on their mutation profile, time period, and
 155 phylogenetic distance (from here on called Alpha subclusters; for details, see Methods
 156 “Subcluster identification”). All subclusters, their time period, and sample size in Thuringia
 157 are summarized in Table 1. An overview of each subcluster (phylogenetic time tree, location,
 158 and period) is also provided on microreact.org as interactive views (see Methods “Subcluster
 159 identification”). Note that our subcluster definition is similar to the definition of a sublineage.
 160 However, PANGO sublineages were rarely defined during the Alpha wave (PANGO
 161 designation: Q.1 to Q.8; compared to the Delta and Omicron waves).

162 **Table 1: Overview of nine Alpha subclusters in Thuringia, their sample count, their time period,**
 163 **and their specific mutations that are shared across all members of the subcluster (excluding**
 164 **characteristic Alpha mutations that are shared across all subclusters).** The mutation used to
 165 define the subcluster is highlighted in bold.

Designation	Mutations	Number of samples	Time period	Remarks
1	S:H49Y , ORF1a:I841V	44	Feb-May 2021	S:H49Y eases cell entry in S-pseudotyped lentiviral system. ²⁴
2	S:N354K	63	Feb-May 2021	S:N354K slightly impaired mAb h11B11. ²⁵
3	S:G496S , ORF1a:E1013K	12	Mar-May 2021	S:G496S: compromises BA.1 replication fitness and changed mAb sensitivities, reduces ACE2 binding affinity, and increases immune evasion. ²⁶⁻²⁸
4	S:N703D , ORF1a:D1228G, ORF1a:A2123V	51	Mar-May 2021	-
5	S:T716V , N:G204P, ORF1a:D1600N	22	Apr-May 2021	-
6	S:S939F	206	Feb-May 2021	S:S939F: modulates T-cell propensity. ²⁹
6.1 [#]	S:V90F , S:S939F	55	Feb-May 2021	-
7	ORF1b:A520V	811	Feb-Jun 2021*	-
7.1 [§]	S:N185D , ORF1b:A520V, ORF1b:L1504F	40	Feb-May 2021	-

166 * only one sample for June # branch from subcluster 6 § branch from subcluster 7

167 Eight of these subclusters are based around a specific spike protein mutation, while the
 168 other contains a mutation within the ORF1b region. The subcluster 7.1 “S:N185D” branched
 169 out from the subcluster 7 “ORF1b:A520V” and subcluster 6.1 “S:V90F” is a branch of
 170 subcluster 6 “S:S939F” (see Table 1). These two branched subclusters still carry the specific
 171 mutation of their originating subcluster. The subclusters 3, 4, and 5 were observable

172 between two and three months, and the other subclusters over at least four months. To
173 investigate these subclusters' regional spread, each sample was mapped to its Thuringian
174 district based on the resident's postal code from which it was isolated. We then sorted the
175 samples according to their subclusters and visualized them throughout the subcluster's
176 observed period. The spread of two representative subclusters is exemplary visualized in
177 Figure 2a, and all the subclusters are available via Supplementary Figure S4 and their data
178 in Supplementary Table S3. Additionally, all subclusters and their metadata are also
179 available via Microreact (see Methods "Subcluster identification").

180

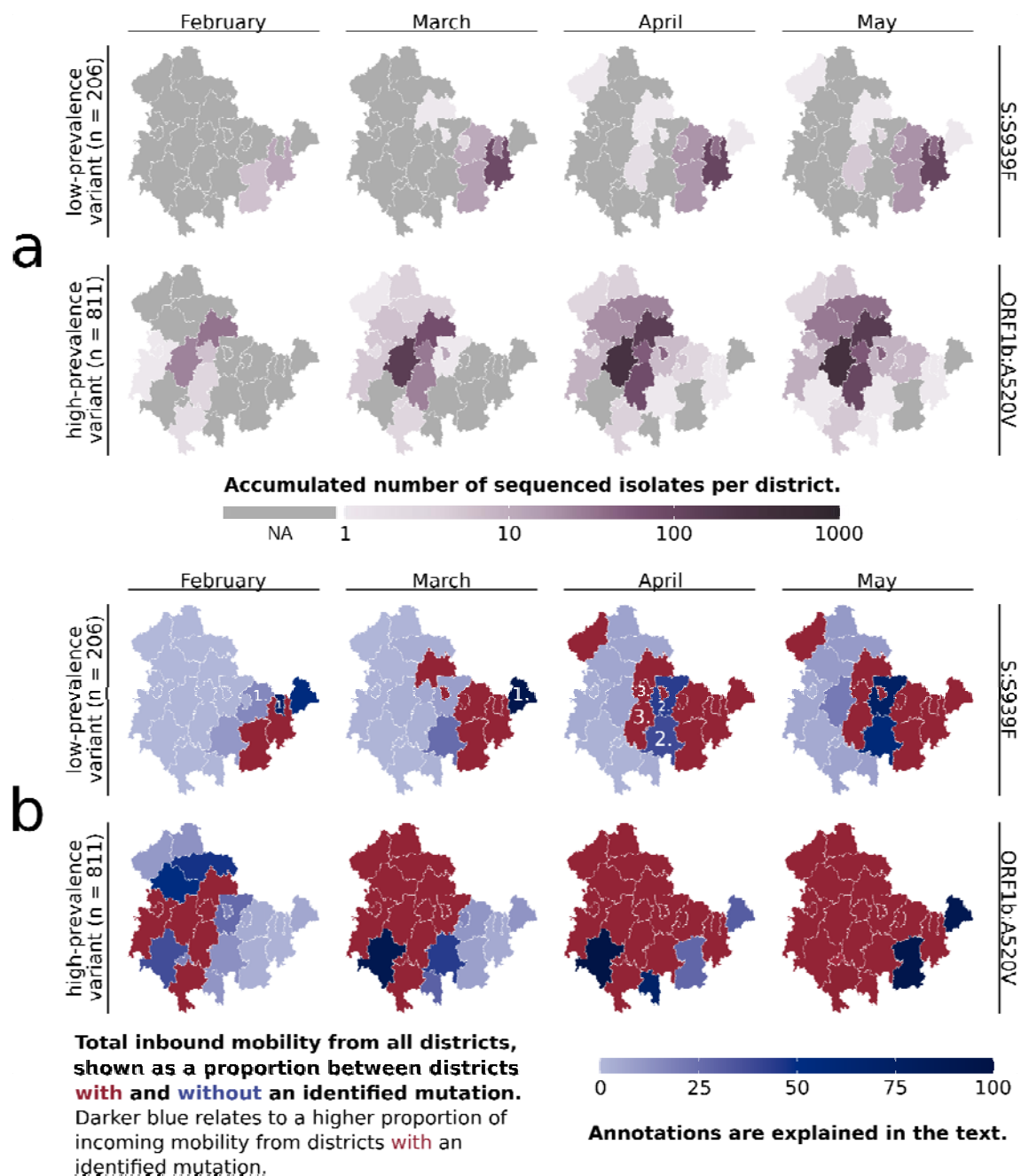
181 Each of the seven main mutation variant clusters originated from a different Thuringian
182 district. At the same time, two subclusters, 6.1 and 7.1, branched out from the same districts
183 as their original clusters (6 and 7) 12 or 13 days after their first emergence. The subclusters
184 mainly spread regionally confined and not across all of Thuringia (see Figure 2a and
185 Supplementary Figure S4) but were also identified in other states of Germany (see
186 "microreact.org"-project). For example, the "S:S939F" subcluster spread across 15 states,
187 with the first samples being isolated outside of Thuringia. The eight Spike-mutation
188 subclusters had expanded between four to twelve of the 23 Thuringian districts within the
189 observable time period of each subcluster. They expanded by one to six districts per month,
190 with a greater expansion accompanied mainly by a larger increase in the subcluster sample
191 number. In contrast, the ORF1b-variant even comprised 21 districts and expanded between
192 two to seven districts per month (see Figure 2a). Most of each subcluster's samples were
193 identified in their region of first occurrence, and no additional samples were found after the
194 given periods.

195 Several limitations need to be considered. The identified subclusters may have multiple
196 origins or may not originate from Thuringia. Due to the lack of precise zip codes (publicly
197 available German genomes are limited to postal codes of sending and sequencing
198 laboratories), monitoring the subclusters in other states on a district level was impossible.
199 Nevertheless, we could follow how the subclusters developed in Thuringia, even if multiple
200 origins may have affected the overall speed or length of each subcluster's occurrence.

201 Our surveillance sampling heavily relies on various institutions and partners, and only a
202 portion of the provided samples can be sequenced (see "Sampling" in Methods). For
203 example, the spread of subcluster "S:S939F" revealed two districts in April where no
204 respective samples were found (Figure 2a) despite being surrounded by districts with
205 "S:S939F"-samples present. This could be due to the lack of samples sent to sequencing
206 from those regions or the low prevalence. We, therefore, investigated if mobile service data
207 of residents, in addition to molecular surveillance, might be utilized to counteract this issue.

208 Mobile service data indicates Alpha subcluster spread and sampling bias

209 With the aim to predict the subcluster spread and, thereby, reduce surveillance-based
210 sampling bias, we utilized anonymized mobile service data from T-Systems International
211 GmbH. Around 200 million trips were used to determine the number of daily trips between
212 the Thuringian districts. We then combined this information with our fine-granular genomic
213 data to specify each district's monthly proportion of inbound mobility from subcluster-
214 affiliated districts (see Methods "Mobile service data"). The results are visualized in Figure
215 2b (complete overview in Supplementary Figure S5; data provided in Supplementary Table
216 S4).



217

218 **Figure 2: Overview of the subclusters "S:S939F" and "ORF1b:A520V" in Thuringian districts.**
 219 a) Accumulated number of sequenced samples for each subcluster per district and per month. b)
 220 Combined visualization of each district's "inbound mobility" from other districts (color intensity) and the
 221 occurrence of a subcluster sample (red = sample found, blue = no sample found). The inbound
 222 mobility of each district (blue color intensity) is shown as a proportion of incoming mobility from other
 223 districts with or without an identified sample. The darker the blue color of a district, the higher the
 224 proportion of inbound mobility from other districts with an identified subcluster sample (red districts).
 225 The light blue color describes that most of the inbound mobility of a district comes from other districts
 226 without an identified subcluster sample (blue districts). Numbers refer to district types 1, 2, and 3,
 227 as further defined in the main text. The last month of subcluster "ORF1b:A520V" is not visualized, as
 228 affected districts were unchanged.

229 The mobile service data-based prediction of a subcluster's spread aligned well with the
230 subsequent regional coverage of fast-spreading, highly prevalent subclusters, such as
231 subcluster 7, which covered 811 samples (see Figure 2). In contrast, the predicted spread
232 for the low-prevalence subclusters did not correspond well with the actual occurrence. Yet,
233 adding mobile service data resulted in three different “types” of districts (see Figure 2b,
234 annotated districts). Type 1 included districts with high inbound mobility from areas with an
235 identified variant, where the variant was eventually found afterward, while Type 2 included
236 districts with high inbound mobility from areas with an identified variant, where the variant
237 was never identified. Type 3 included districts not directly connected to a district with an
238 identified variant, but a variant was eventually identified while they border Type 2 district(s).
239 Our previous analysis of the subclusters' spreading pattern across the districts, based solely
240 on identified variants, indicated missed identifications in some districts due to the seemingly
241 illogical spread to districts without a connection to others (Figure 2a, subcluster “S:S939F”).
242 The inclusion of mobile service data revealed some of these districts to be Type 2 districts.
243 This suggests that the specific variant should be identifiable within these districts due to the
244 observed high incoming mobility from districts with identified variants. Type 2 districts were
245 mainly observed for subclusters with low prevalence and, consequently, low numbers of
246 covered samples that are usually more difficult to monitor. For example, we assumed
247 missing identifications in some districts of subclusters 1, 2, and 3, which, through the mobile
248 service data, are now partially identified as Type 2 districts. However, due to their low
249 prevalence, it is also possible that these subclusters have not spread to the indicated
250 districts.

251 Despite analyzing the mobile service data of districts from other federal states than
252 Thuringia, we could not apply them, as the lack of precise location data for samples outside
253 of Thuringia prevented the correct calculation of the incoming mobility. Based on the nine
254 observable clusters, we concluded that mobile service data is a good prediction marker for
255 the spread of high-prevalence variants but, more importantly, a good indication of districts
256 that should have an identified low-prevalence variant. Next, we investigated if mobile service
257 data can improve active surveillance via guiding sample collection for genomic sequencing.

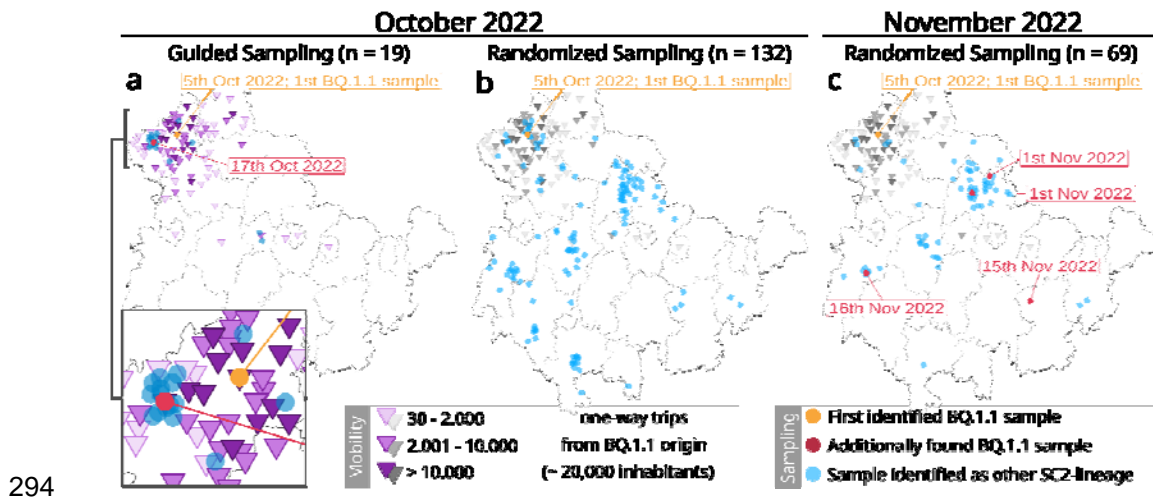
258 Proof of principle: Mobile service data-guided sampling for genomic 259 surveillance for Omicron BQ.1.1

260 Based on our previous findings, we implemented the “mobility-guided” sampling approach
261 under real pandemic circumstances over one month in addition to our active surveillance.
262 As the subject of investigation, we searched for a newly emerging (based on global news
263 reports) and ideally low prevalent SARS-CoV-2 lineage in Thuringia.

264 Among the various emerging Omicron sublineages during that time, sublineage BQ.1.1
265 fulfilled the defined criteria. First isolated in a northwestern Thuringian community with
266 around 20,000 inhabitants on October 5, 2022, we identified this particular sublineage on
267 October 14, 2022, among a routine batch of 42 samples. BQ.1.1 was a low-prevalence
268 sublineage that was identified worldwide ([https://outbreak.info/situation-](https://outbreak.info/situation-reports?pango=BQ.1.1)
269 [reports?pango=BQ.1.1](https://outbreak.info/situation-reports?pango=BQ.1.1)).

270 Following its first Thuringian identification, we utilized the latest available dataset of the past
271 two years of mobile service data (October 2020 and June 2021) to investigate the residential
272 movements for the community of first detection. Considering the highest incoming mobility
273 from both datasets, we identified 18 communities with high (> 10,000), 34 with medium
274 (2,001-10,000), and 82 with low (30-2,000) number of incoming one-way trips from the
275 originating community (purple triangles in Figure 3a). As a result, we specifically requested
276 all the available samples from the eight communities with the highest incoming mobility. Still,
277 we were restricted to the submission of third parties over whom we had no influence. This
278 led to the inclusion of the following eight communities with the most residential movement
279 from the originating community: four in central and three in NW of Thuringia, one in NW-
280 neighboring state Saxony-Anhalt. The samples requested from central Thuringia were also
281 due to their geographic arrangement as a “belt” in central Thuringia, linking three major cities
282 (see Supplementary Figure S1). Subsequently, we collected 19 additional samples (isolated
283 between the 17th and 25th of October 2022; see “Guided Sampling” for October 2022,
284 Figure 3a) besides the randomized sampling strategy. Thus, the sampling depth was
285 increased in communities with high incoming mobility from the first origin.

286 As part of the general Thuringian surveillance, we collected 132 samples for October
287 (covering dates between the 5th and 31st) and 69 samples in November (covering dates
288 between the 1st and 25th; see Figure 3b and 3c). Randomized sampling was not influenced
289 or adjusted based on the mobility-guided sample collection. Thus, it also contains samples
290 from communities with a mobility link towards the first occurrence of BQ.1.1, as they were
291 part of the regular random collection (see gray triangles in Figure 3b). A complete overview
292 of all samples is provided in Supplementary Table S5. The mobility datasets from October
293 2020 and June 2021 for all sampled communities are provided in Supplementary Table S6.



295 **Figure 3: Overview of the mobility-guided sampling of the Omicron sublineage BQ.1.1 in**
296 **Thuringia (a) compared to the default randomized sampling (surveillance) in October (b) and**
297 **November 2022 (c).** The randomized surveillance results in November 2022 (c) have been added to
298 highlight the spreading progress of BQ.1.1. Dots reflect the location of each sample (based on
299 residents' zip codes). Orange dot: First identified BQ.1.1 sample; Red dot: Additionally identified
300 BQ.1.1 sample; Blue dot: Another SARS-CoV-2 lineage. Purple triangles represent the number of
301 one-way trips a community received from the community of the first BQ.1.1 occurrence (orange dot)
302 based on mobility data from October 2020 and June 2021. The same mobility data from mobility-
303 guided sampling (a) were added in grayscales to the randomized surveillance (b and c) as a visual
304 reference only. Sampling dots are slightly scattered to improve visibility.

305 Among the 19 samples specifically collected based on mobile service data, we identified one
306 additional sample of the specific Omicron sublineage BQ.1.1 in a community with high
307 incoming mobility ($n = 14$, number of trips = 37,499) with a distance of approximately 16 km
308 between both towns. Our randomly sampled routine surveillance strategy did not detect
309 another sample during the same period. This was despite a seven times higher overall
310 sample rate, which included 31 samples from communities with an identified incoming
311 mobility from the community of the first occurrence (October 2022, Figure 3b). Only in the
312 one-month follow-up were four other samples identified across Thuringia through routine
313 surveillance (November 2022, Figure 3c).

314 During our attempt to implement the mobility-guided sampling approach in real-time during
315 the pandemic, we encountered three distinct limitations, some of which are commonly
316 observed in surveillance practices. The guided sampling depended on the individual sample
317 submitting institutions, affecting the availability of suitable samples, especially for the
318 communities of interest. By choosing a newly emerging Omicron sublineage for our
319 experiment, spread and, therefore, suitability were uncertain. In our case, BQ.1.1's
320 prevalence in Thuringia was even lower than expected, and it also remained rare in
321 subsequent months, with only 42 samples found until June 2023, eight months after the first
322 occurrence in Thuringia. Due to the short preparation time, only mobile service data from the

323 past two years and no current data were available. Nevertheless, the available datasets still
324 reflect pandemic movement behavior since the pandemic has been ongoing for two years.
325 In summary, increasing the sampling depth in the suspected regions successfully identified
326 the specified lineage using only a fraction of the samples from the randomized sampling.
327 Conversely, randomized surveillance, the “gold standard” acting as our negative control, did
328 not identify additional samples with similar sampling depths in regions with no or low
329 incoming mobility or even in high mobility regions with less sampling depth. Implementing
330 such an approach effectively under pandemic conditions poses difficult challenges due to the
331 fluctuating sampling sizes. Although the finding of the sample may have been coincidental,
332 our proof of concept demonstrated how we can leverage the potential of mobile service data
333 for targeted surveillance sampling.

334 Conclusion

335 During the SARS-CoV-2 pandemic, diverse data sources like travel, wastewater, and
336 mobility data have been employed in surveillance and transmission tracking.^{14–19} In the
337 present study, we analyzed over 296,800 German SARS-CoV-2 genomes to examine
338 whether mobile service data can predict the spatial distribution of the Alpha lineage in the
339 German state of Thuringia and how they potentially benefit pandemic surveillance.
340 A plausible explanation for the delayed spread of the Alpha lineage in Thuringia is the lack of
341 major transport hubs, as Alpha first occurred in federal states with such hubs. Previous
342 studies have already highlighted the impact of major transportation hubs in the spread of
343 Sars-CoV-2^{14,30}. However, its impact on the total distribution is limited, and the spread was
344 ultimately comparable between Germany and Thuringia. While our findings on mobile
345 service data may also apply to Germany, we could not verify this because the limited
346 location data of publicly available German genomes prevented in-depth investigations
347 outside of Thuringia. Thus, precise sampling location data are crucial to utilize mobile
348 service data in genomic surveillance, but privacy regulations may restrict access to this data.
349 Shortly after Alpha’s emergence, mutation variants formed like the known sublineages Q.1 to
350 Q.8 and the Thuringian subclusters identified by us. This reflects the ongoing evolution
351 during active circulation and indicates an even greater sublineage diversity, which has not
352 been surveyed as closely as in the subsequent Delta and Omicron waves. By monitoring the
353 nine Thuringian subclusters, rather than focusing solely on the parental lineage B.1.1.7, we
354 were again able to effectively track transmissions and gain a comprehensive understanding
355 of the regional spread. So, it underscores the importance of sequencing in pandemic

356 surveillance to explore such genomic changes and, thereby, keep track of the transmission
357 chains and potential outbreaks.

358 Mobile service data can support such surveillance in different ways. Previous studies
359 examined the capabilities of mobility data in the context of, e.g., case trajectories, but
360 retrospectively applied to already collected data, it can be used to examine surveillance
361 sampling coverage and possible sampling bias. We exemplified this approach with the Alpha
362 lineage, where mobile service data indicated a putative sampling bias and partially predicted
363 the spread of our Thuringian subclusters.

364 Another approach is actively guiding the sampling process through mobile service data,
365 which we demonstrated with our proof of principle focusing on the Omicron-lineage BQ.1.1
366 as a real-life example. This approach could allow for a flexible allocation of surveillance
367 resources, enabling adaptation to specific circumstances and increasing sampling depth in
368 regions where a variant is anticipated. By incorporating guided sampling, much fewer
369 resources may be needed for unguided or random sampling, thereby reducing overall
370 surveillance costs.

371 Additionally, while this approach is particularly useful for identifying low-prevalence variants,
372 it is not limited to such variants. Still, it can provide a guided, more cost-efficient, low-
373 sampling alternative to general randomized surveillance that can also be applied to other
374 viruses or lineages. For this purpose, Pre-generated mobility networks automatically tailored
375 to each state's unique infrastructure and population dynamics could provide better-targeted
376 sampling guidance rather than simple geographical proximity. However, the feasibility
377 depends on the availability and cost of such mobile service data. Alternatively, financial
378 resources could also be invested directly in increasing sampling capacity and coverage,
379 which ultimately depends on individual factors of the respective surveillance. Mobile service
380 data can also be used with other surveillance approaches and elements. For example,
381 wastewater surveillance can give further indications to supplement guided sampling. At the
382 same time, passenger data offers additional insights into traffic hubs as sources of regional
383 movement.

384 Methods

385 Sampling

386 Starting mid-2020, we initially sequenced hospital-intern samples, transitioning by January
387 2021 to approximately 43 PCR-positive samples per week: 20 from the hospital's

388 microbiology department and 23 randomly sourced by the Thuringian State Authority for
389 Consumer Protection (“Thüringer Landesamt für Verbraucherschutz”; TLV).
390 Until June 2023, our institute sequenced 3,770 SARS-CoV-2 samples, and SYNLAB Holding
391 Deutschland GmbH, Bioscientia Healthcare GmbH, and DIANOVIS GmbH provided
392 additional 7,800 Thuringian SARS-CoV-2 genomes and their metadata.

393 Sample preparation and sequencing

394 RNA isolation used the ZymoResearch “Quick-RNA Viral Kit” (Zymo Research Europe
395 GmbH, Germany, Product-ID: R1035), according to the manufacturer's instructions with
396 100 µl patient sample input and a centrifuge speed of 16,000 g.

397 The viral RNA underwent a Reverse Transcriptase (RT)-PCR followed by a multiplex-PCR
398 using the ARTIC V1200 primer set, according to Freed and Silander's SARS-CoV-2
399 sequencing protocol (version 4, updating to version 5 by March 2021).³¹ Subsequent DNA
400 quantification utilized the Qubit dsDNA HS assay (Invitrogen, USA).

401 From the amplified DNA, a sequencing library was prepared using the Nanopore SQK-
402 LSK109 and SQK-RBK004 kits (Oxford Nanopore Technologies, Oxford, UK), sequenced for
403 a maximum of 72 h utilizing an Oxford Nanopore MinION Mk1b sequencer with R.9-flowcells
404 and the MinKNOW software (versions MKE_1013_v1_revBC_11Apr2016 to
405 MKE_1013_v1_revBR_11Apr2016 in the respective period), and analyzed with the software
406 pipeline poreCov (versions 0.3.5 to 0.11.7; including basecalling, demultiplexing, adapter
407 removal, quality filtering, and genome alignment) to reconstruct consensus genomes.⁷

408 Sequencing data and the respective metadata (e.g., isolation date, sending laboratory
409 details) were submitted to the RKI through DESH. We also collected the postal code of the
410 isolation location or at least of the sending local health authority, storing all data additionally
411 in a local database.³² Due to data protection, such data is limited on the RKI's public GitHub
412 repository ([https://github.com/robert-koch-institut/SARS-CoV-2-](https://github.com/robert-koch-institut/SARS-CoV-2-Sequenzdaten_aus_Deutschland)
413 [Sequenzdaten aus Deutschland](https://github.com/robert-koch-institut/SARS-CoV-2-Sequenzdaten_aus_Deutschland)), providing instead postal codes of the sequencing and
414 sending laboratories.

415 Alpha spread datasets

416 From our local database, we extracted 8,397 samples with isolation dates before Oct 1st,
417 2021. After adding federal state and district information, 993 entries with non-Thuringian
418 locations were excluded. Further, ten entries with unspecific isolation dates were excluded,
419 yielding 7,394 samples (including 6,307 Alpha genomes (lineages B.1.1.7 and Q.1 to Q.8)).
420 The publicly available RKI SARS-CoV-2 dataset was downloaded, containing 1,091,655
421 genomes with the respective metadata (17th Oct 2022; Zenodo-version 2022-10-16).³³

422 789,405 entries, isolated after Sep'21, and 59 entries without “sending laboratory”
423 information were removed. For the resulting 302,191 entries, location information (location,
424 federal state, district, longitude, latitude) were added based on the sending laboratory postal
425 code. Five entries with a non-existing postal code and all 12,704 Thuringian samples were
426 removed from the dataset, resulting in 289,487 samples (including 137,024 Alpha genomes).
427 We investigated only Alpha-lineage samples collected from September 2020 onwards, after
428 the first official reports of the Alpha-lineage.¹¹
429 Analyzing both datasets, we calculated the monthly proportion of Alpha lineage samples in
430 Thuringia and Germany per state/district, dividing Dec'20 and Jan'21 into first and second
431 halves.

432 Subcluster identification

433 Using a total of 70,429 German and Thuringian Alpha genomes, a phylogenetic time tree
434 was created (see Supplementary Method “Phylogenetic time tree construction” and
435 Supplementary Figure S3). We determined the frequency of all non-Alpha-specific mutations
436 among the 6,522 Thuringian Alpha genomes. We then manually screened for mutations
437 present in at least 20 genomes with a small phylogenetic distance and a time occurrence of
438 at least two months. This led to nine mutations, each of them creating a defined cluster
439 covering between 12 and 811 closely related genomes. We only kept mutation information of
440 these nine subclusters in the respective metadata, which, together with the tree file of the
441 phylogenetic time tree, was uploaded to a “microreact.org”-project, provided as
442 Supplementary File 1 and found under the following link:
443 <https://microreact.org/project/ftR2GfjF6iXtSwbmN4ARTx-thuringianalpha-linclusters#76ir->
444 [complete-overview](https://microreact.org/project/ftR2GfjF6iXtSwbmN4ARTx-thuringianalpha-linclusters#76ir-complete-overview).

445 Mobile service data

446 T-Systems International GmbH collected and aggregated mobile service data via the Cell ID
447 method, dividing a geographical area into so-called traffic cells. Each cell is assigned to
448 exactly one transmitter mast, with a spatial resolution from 500 m x 500 m up to 8 km x 8 km
449 (depending on the transmitter mast network density). Cell phones always register to the
450 closest traffic cell, which is recorded and stored in an Origin-Destination Matrix (ODM). For
451 population representation, the data was extrapolated with Deutsche Telekom's market
452 share. Due to data privacy, the registration data is combined into movement streams
453 between traffic cells, the status resolution is reduced to one hour (greater time intervals =
454 less resolution), and individual traffic cells are grouped into districts. The degree of
455 anonymization (k-value = 30, data privacy regulation) removed movement streams with less

456 than 30 participants, resulting in approximately 200 Mio trips in the ODM. SMA Development
457 GmbH analyzed all movements between the single Thuringian districts, adding each Alpha
458 sample's isolation time and location data (per subcluster). The movements were further
459 divided by months and originating district (subcluster-affiliated vs. -unaffiliated), determining
460 each district's monthly inbound mobility proportion from cluster-affiliated districts.

461 Research in Context

462 Evidence before this study

463 We searched Pubmed for studies about the use of mobile service data for surveillance
464 written in English. For the broadest possible search, we included any publication covering
465 mobile data and surveillance aspects, using the following search string: ("cellular data" OR
466 "cell phone data" OR "mobility data" OR "movement data" OR "migration data" OR "phone
467 data") AND ("Surveillance" OR "Monitoring" OR "Survey" OR "Pandemic" OR "Disease" OR
468 "Epidemic" OR "Outbreak"). Our search yielded 1,285 publications published between 1966
469 and 2023. We manually screened all these publications but found no study that applied
470 mobile service data for active, targeted surveillance. Across all studies, the general focus
471 was on tracking contacts or analyzing movements to assess, for instance, the efficiency of
472 non-pharmaceutical interventions or generate prediction models. Some studies suggested
473 targeted surveillance based on their results, but it was not yet applied. Additionally, we used
474 "suite.ai" and "chatGPT" (with BING-search access) to let them search for "studies that
475 utilize mobile service data to guide the sampling process for infectious disease surveillance".
476 While "suite.ai" found two studies and "chatGPT" found another ten studies and reviews,
477 none covered the direct application of the mobility data in active surveillance.

478 Added value of this study

479 This study highlights the value of combining mobile service data with fine-granular metadata
480 for integrated genomic surveillance during the SARS-CoV-2 pandemic in a German federal
481 state. We illustrated this strategy with the Omicron sublineage BQ.1.1 and how to guide the
482 sampling processes toward areas where the new variant was expected to emerge.
483 Additionally, we used mobile service data during the pandemic to assess our sampling
484 coverage. Our study is the first to actively guide part of the genomic surveillance process
485 during a pandemic.

486 Implications of all the available evidence

487 Efficient molecular surveillance setups are crucial in managing outbreaks from the local to
488 the global scale. Different data sources are investigated to increase this efficiency,
489 addressing factors like the more efficient usage of scarce surveillance resources and the
490 prediction of spread. Extending molecular surveillance with such data should improve the
491 future management of pandemics and outbreaks.

492 Author contributions

493 Sample collection and preparation, R.Sp, R.Sc., M.L., and M.M.; sequencing, R.Sp., M.L.,
494 and M.M.; database setup and maintenance, M.J.; software, C.B.; bioinformatic analysis,
495 R.Sp., M.L., M.M, and C.B.; mobile service data analysis, C.H., and M.Ha.; literature
496 research, R.Sp.; writing first draft, R.Sp.; reviewing and editing manuscript, R.Sp., C.B.,
497 M.M., C.F.-S., A.K., C.H., M.Ha., M.Hö., D.K., R.Sc., P.D., and M.W.P.; supervision, C.B.;
498 project administration, C.B.; funding acquisition, M.W.P.. All authors have read and agreed
499 to the published version of the manuscript.

500 Funding

501 This work was supported by grants from the Federal Ministry of Education and Research
502 (project “SARS-CoV-2Dx“), [grant number 13N15745] and the Thüringer Aufbaubank
503 (project “Pandemie Analyse mittels Advanced Analytics Methoden“), [grant number 2021
504 VF 0035]. We acknowledge support by the German Research Foundation Projekt-Nr.
505 512648189 and the Open Access Publication Fund of the Thueringer Universitaets- und
506 Landesbibliothek Jena.

507 Conflicts of interest

508 The authors have declared that no competing interests exist.

509 Data sharing statement

510 All genomic data (genomes and respective metadata) are available in the provided microreact
511 project (project file available under <https://osf.io/n5qj6/>;
512 [https://microreact.org/project/ftR2GfjF6iXtSwbmN4ARTx-thuringianalpha-linclusters#76ir-](https://microreact.org/project/ftR2GfjF6iXtSwbmN4ARTx-thuringianalpha-linclusters#76ir-complete-overview)
513 complete-overview). Supplementary tables and the R-scripts used to generate all figures are

514 also provided in the repository under <https://osf.io/n5qj6/>. These include the mobile service
515 data used in this study, which is available in processed and anonymized form. The original
516 mobile service data can not be made public due to legal reasons/ownership.

517 Declaration of generative AI and AI-assisted technologies in the 518 writing process

519 During the preparation of this work, the author(s) used Grammarly in order to correct general
520 English and improve readability. After using this tool/service, the author(s) reviewed and
521 edited the content as needed and take(s) full responsibility for the content of the publication.

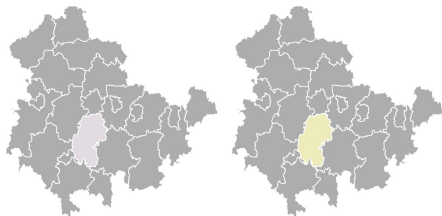
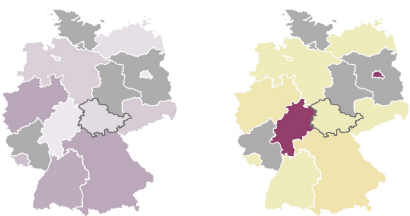
522 References

- 523 1 Zhu N, Zhang D, Wang W, *et al.* A Novel Coronavirus from Patients with Pneumonia in
524 China, 2019. *N Engl J Med* 2020; **382**: 727–33.
- 525 2 An integrated national scale SARS-CoV-2 genomic surveillance network. *Lancet Microbe*
526 2020; **1**: e99–100.
- 527 3 German ‘Corona-Surveillanceverordnung’ as issued on the 18th of January 2021 and
528 updated on the 27th of June 2022.
529 [https://www.bundesgesundheitsministerium.de/fileadmin/Dateien/3_Downloads/C/Corona](https://www.bundesgesundheitsministerium.de/fileadmin/Dateien/3_Downloads/C/Corona_virus/Verordnungen/CorSurV_BAnz_AT_19.01.2021_V2.pdf)
530 [virus/Verordnungen/CorSurV_BAnz_AT_19.01.2021_V2.pdf](https://www.bundesanzeiger.de/pub/publication/hEG1y5vUCsSa43vJZSP/content/hEG1y5vUCsSa43vJZSP/BAnz%20AT%2028.06.2022%20V1.pdf?inline);
531 [https://www.bundesanzeiger.de/pub/publication/hEG1y5vUCsSa43vJZSP/content/hEG1y5vUCsSa43vJZSP/](https://www.bundesanzeiger.de/pub/publication/hEG1y5vUCsSa43vJZSP/content/hEG1y5vUCsSa43vJZSP/BAnz%20AT%2028.06.2022%20V1.pdf?inline)
532 [BAnz%20AT%2028.06.2022%20V1.pdf?inline](https://www.bundesanzeiger.de/pub/publication/hEG1y5vUCsSa43vJZSP/content/hEG1y5vUCsSa43vJZSP/BAnz%20AT%2028.06.2022%20V1.pdf?inline).
- 533 4 Meredith LW, Hamilton WL, Warne B, *et al.* Rapid implementation of SARS-CoV-2
534 sequencing to investigate cases of health-care associated COVID-19: a prospective
535 genomic surveillance study. *Lancet Infect Dis* 2020; **20**: 1263–71.
- 536 5 Page AJ, Mather AE, Le-Viet T, *et al.* Large-scale sequencing of SARS-CoV-2 genomes
537 from one region allows detailed epidemiology and enables local outbreak management.
538 *Microb Genomics* 2021; **7**. DOI:10.1099/mgen.0.000589.
- 539 6 Lambrou AS, Shirk P, Steele MK, *et al.* Genomic Surveillance for SARS-CoV-2 Variants:
540 Predominance of the Delta (B.1.617.2) and Omicron (B.1.1.529) Variants — United
541 States, June 2021–January 2022. *MMWR Morb Mortal Wkly Rep* 2022; **71**: 206–11.
- 542 7 Brandt C, Krautwurst S, Spott R, *et al.* poreCov-An Easy to Use, Fast, and Robust
543 Workflow for SARS-CoV-2 Genome Reconstruction via Nanopore Sequencing. *Front*
544 *Genet* 2021; **12**: 711437.
- 545 8 Lataretu M, Drechsel O, Kmiecinski R, Trappe K, Hölzer M, Fuchs S. Lessons learned:
546 overcoming common challenges in reconstructing the SARS-CoV-2 genome from short-
547 read sequencing data via CoVpipe2. *F1000Research* 2023; **12**: 1091.
- 548 9 Rambaut A, Holmes EC, O’Toole Á, *et al.* A dynamic nomenclature proposal for SARS-
549 CoV-2 lineages to assist genomic epidemiology. *Nat Microbiol* 2020; **5**: 1403–7.
- 550 10 PANGO SARS-CoV-2 lineage overview; accessed 18th August 2023.
551 https://github.com/cov-lineages/lineages-website/blob/master/_data/lineage_data.full.json.
- 552 11 Washington NL, Gangavarapu K, Zeller M, *et al.* Emergence and rapid transmission of
553 SARS-CoV-2 B.1.1.7 in the United States. *Cell* 2021; **184**: 2587-2594.e7.
- 554 12 Walker AS, Vihta K-D, Gethings O, *et al.* Tracking the Emergence of SARS-CoV-2 Alpha
555 Variant in the United Kingdom. *N Engl J Med* 2021; **385**: 2582–5.
- 556 13 Michaelsen TY, Bennedbæk M, Christiansen LE, *et al.* Introduction and transmission of
557 SARS-CoV-2 lineage B.1.1.7, Alpha variant, in Denmark. *Genome Med* 2022; **14**: 47.

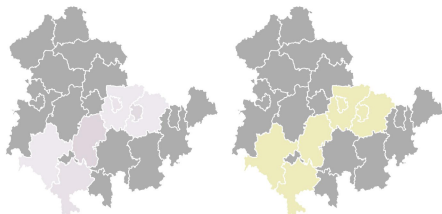
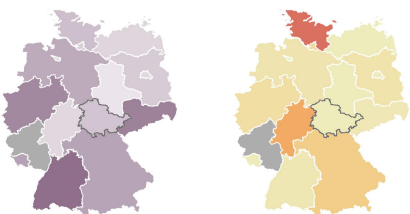
- 558 14 Alpert T, Brito AF, Lasek-Nesselquist E, *et al.* Early introductions and transmission of
559 SARS-CoV-2 variant B.1.1.7 in the United States. *Cell* 2021; **184**: 2595-2604.e13.
- 560 15 O'Toole Á, Hill V, Pybus OG, *et al.* Tracking the international spread of SARS-CoV-2
561 lineages B.1.1.7 and B.1.351/501Y-V2 with grinch. *Wellcome Open Res* 2021; **6**: 121.
- 562 16 Li J, Ahmed W, Metcalfe S, *et al.* Monitoring of SARS-CoV-2 in sewersheds with low
563 COVID-19 cases using a passive sampling technique. *Water Res* 2022; **218**: 118481.
- 564 17 Kraemer MUG, Hill V, Ruis C, *et al.* Spatiotemporal invasion dynamics of SARS-CoV-2
565 lineage B.1.1.7 emergence. *Science* 2021; **373**: 889–95.
- 566 18 Chang S, Pierson E, Koh PW, *et al.* Mobility network models of COVID-19 explain
567 inequities and inform reopening. *Nature* 2021; **589**: 82–7.
- 568 19 Schlosser F, Maier BF, Jack O, Hinrichs D, Zachariae A, Brockmann D. COVID-19
569 lockdown induces disease-mitigating structural changes in mobility networks. *Proc Natl*
570 *Acad Sci* 2020; **117**: 32883–90.
- 571 20 Meintrup D, Nowak-Machen M, Borgmann S. A Comparison of Germany and the United
572 Kingdom Indicates That More SARS-CoV-2 Circulation and Less Restrictions in the Warm
573 Season Might Reduce Overall COVID-19 Burden. *Life* 2022; **12**: 953.
- 574 21 Oh DY, Hölzer M, Paraskevopoulou S, *et al.* Advancing Precision Vaccinology by
575 Molecular and Genomic Surveillance of Severe Acute Respiratory Syndrome Coronavirus
576 2 in Germany, 2021. *Clin Infect Dis* 2022; **75**: S110–20.
- 577 22 Pletz MW, Steiner A, Kesselmeier M, *et al.* Introduction of mandatory masking in health
578 care and community: experience from Jena, Germany. *Infection* 2023; published online
579 April 3. DOI:10.1007/s15010-023-02015-w.
- 580 23 Aksamentov I, Roemer C, Hodcroft E, Neher R. Nextclade: clade assignment, mutation
581 calling and quality control for viral genomes. *J Open Source Softw* 2021; **6**: 3773.
- 582 24 Ozono S, Zhang Y, Ode H, *et al.* SARS-CoV-2 D614G spike mutation increases entry
583 efficiency with enhanced ACE2-binding affinity. *Nat Commun* 2021; **12**: 848.
- 584 25 Du Y, Shi R, Zhang Y, *et al.* A broadly neutralizing humanized ACE2-targeting antibody
585 against SARS-CoV-2 variants. *Nat Commun* 2021; **12**: 5000.
- 586 26 Liang R, Ye Z-W, Ong CP, *et al.* The spike receptor-binding motif G496S substitution
587 determines the replication fitness of SARS-CoV-2 Omicron sublineage. *Emerg Microbes*
588 *Infect* 2022; **11**: 2093–101.
- 589 27 Kimura I, Yamasoba D, Nasser H, *et al.* The SARS-CoV-2 spike S375F mutation
590 characterizes the Omicron BA.1 variant. *iScience* 2022; **25**: 105720.
- 591 28 Asif A, Ilyas I, Abdullah M, Sarfraz S, Mustafa M, Mahmood A. The Comparison of
592 Mutational Progression in SARS-CoV-2: A Short Updated Overview. *J Mol Pathol* 2022; **3**:
593 201–18.
- 594 29 Donzelli S, Spinella F, di Domenico EG, *et al.* Evidence of a SARS-CoV-2 double Spike
595 mutation D614G/S939F potentially affecting immune response of infected subjects.
596 *Comput Struct Biotechnol J* 2022; **20**: 733–44.
- 597 30 Tegally H, Wilkinson E, Tsui JL-H, *et al.* Dispersal patterns and influence of air travel
598 during the global expansion of SARS-CoV-2 variants of concern. *Cell* 2023; **186**: 3277-
599 3290.e16.
- 600 31 Freed N, Silander O. SARS-CoV2 genome sequencing protocol (1200bp amplicon
601 "midnight" primer set, using Nanopore Rapid kit) v5. 2021
602 DOI:10.17504/protocols.io.btsrmd6.
- 603 32 Jundzill M, Spott R, Lohde M, Hölzer M, Viehweger A, Brandt C. Managing and
604 monitoring a pandemic: showcasing a practical approach for the genomic surveillance of
605 SARS-CoV-2. *Database* 2023; **2023**: baad071.
- 606 33 Koch-Institut R. SARS-CoV-2 Sequenzdaten aus Deutschland. 2022; published online
607 Oct 16. DOI:10.5281/ZENODO.7212725.

Germany

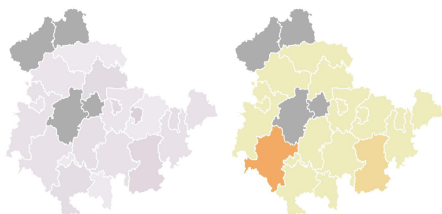
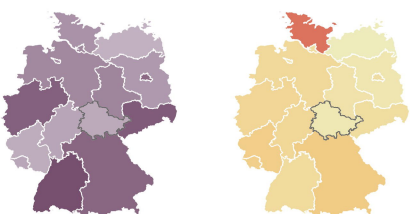
Thuringia



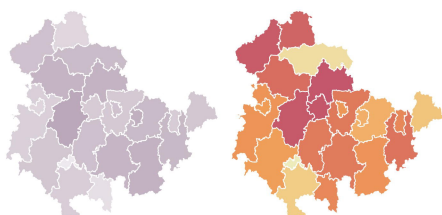
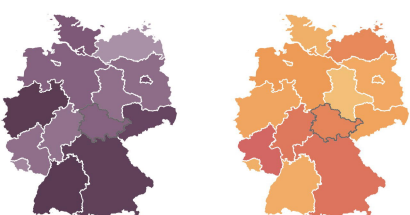
2020-Dec
17th - 31st



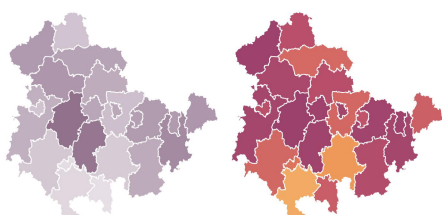
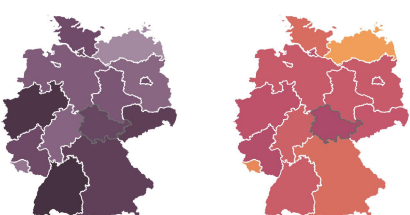
2021-Jan
1st - 16th



2021-Jan
17th - 31st



2021-Feb



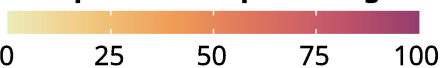
2021-Mar

Sequenced SARS-CoV-2 samples



10 100 1000 10000

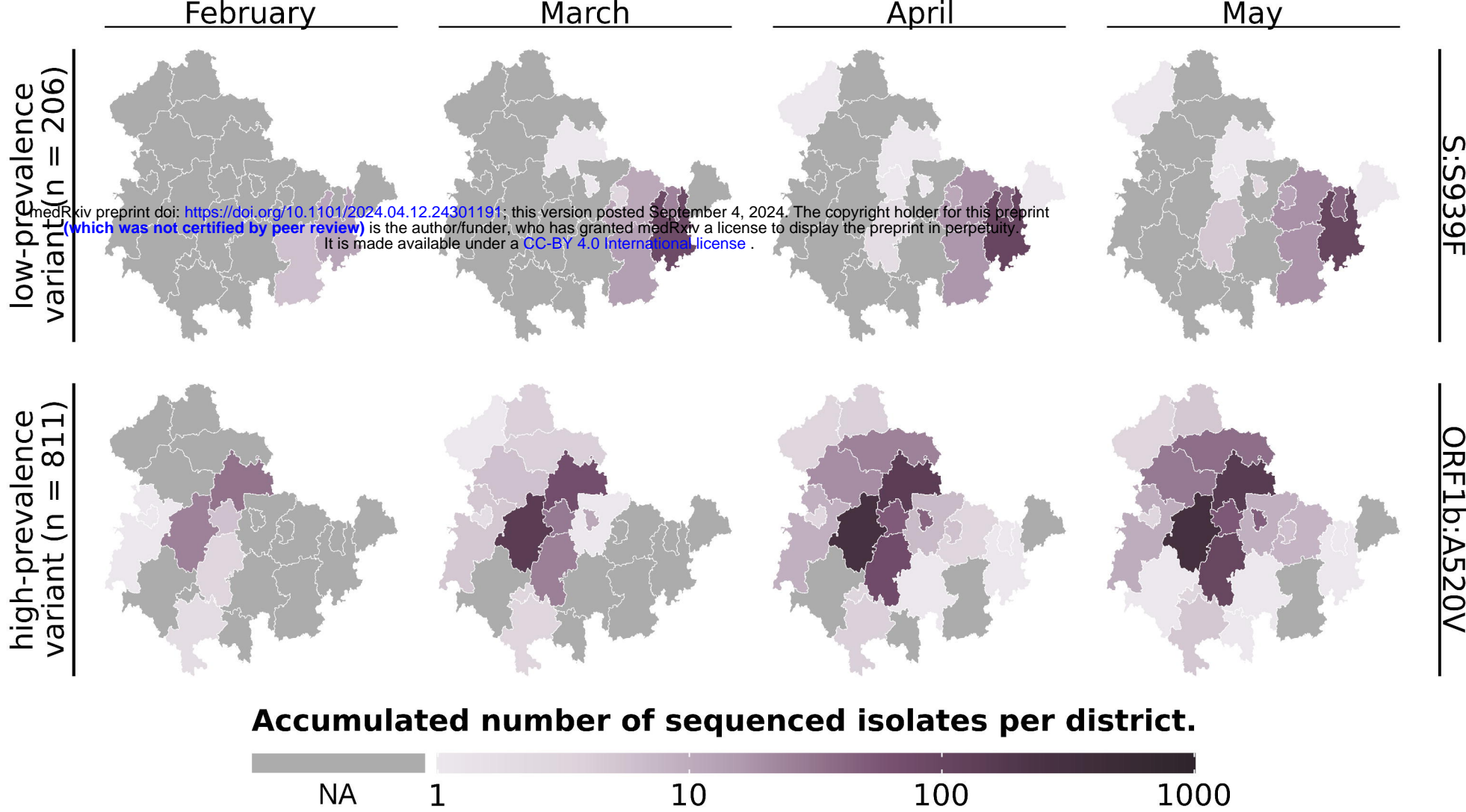
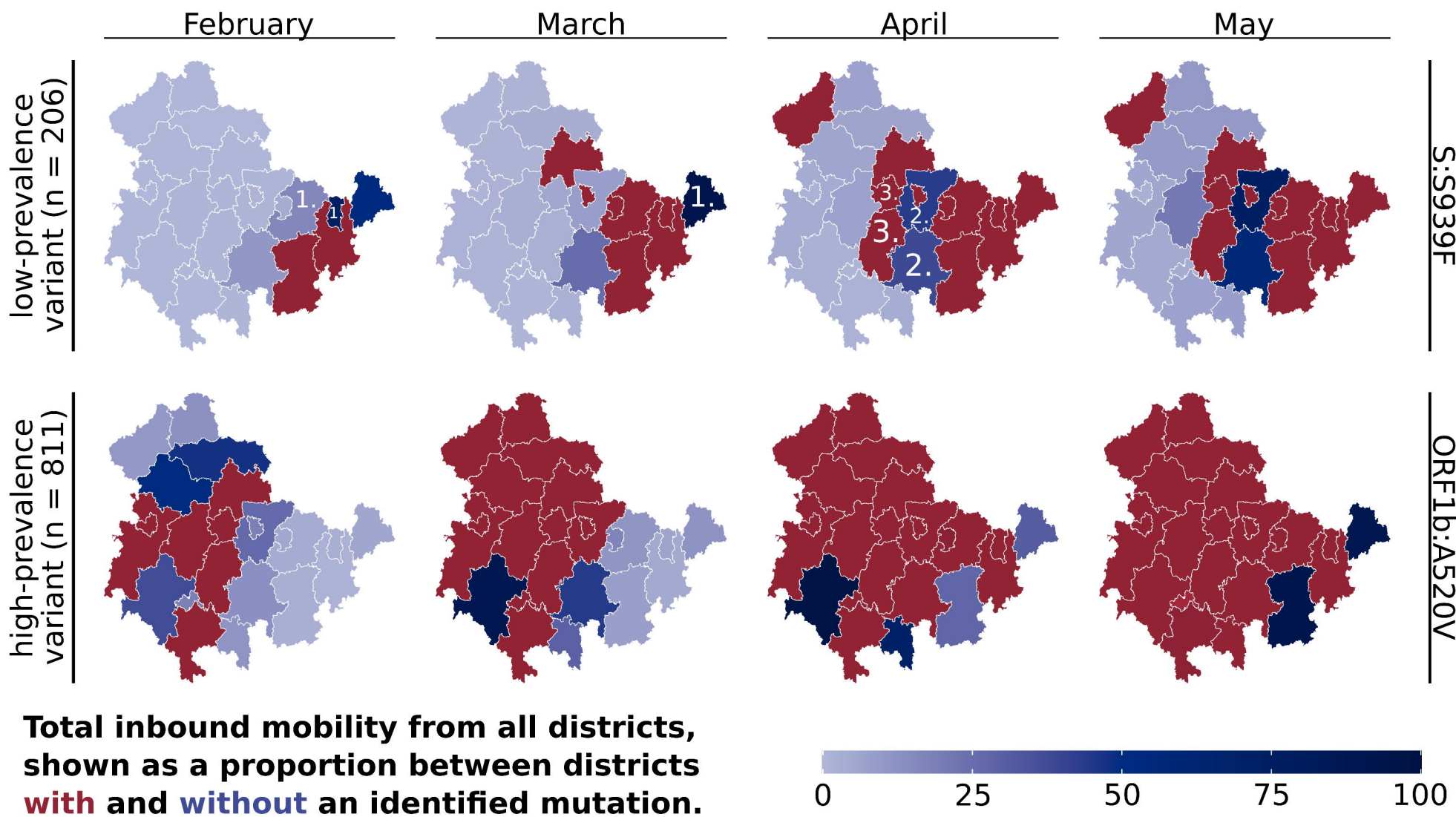
Proportion of Alpha lineage



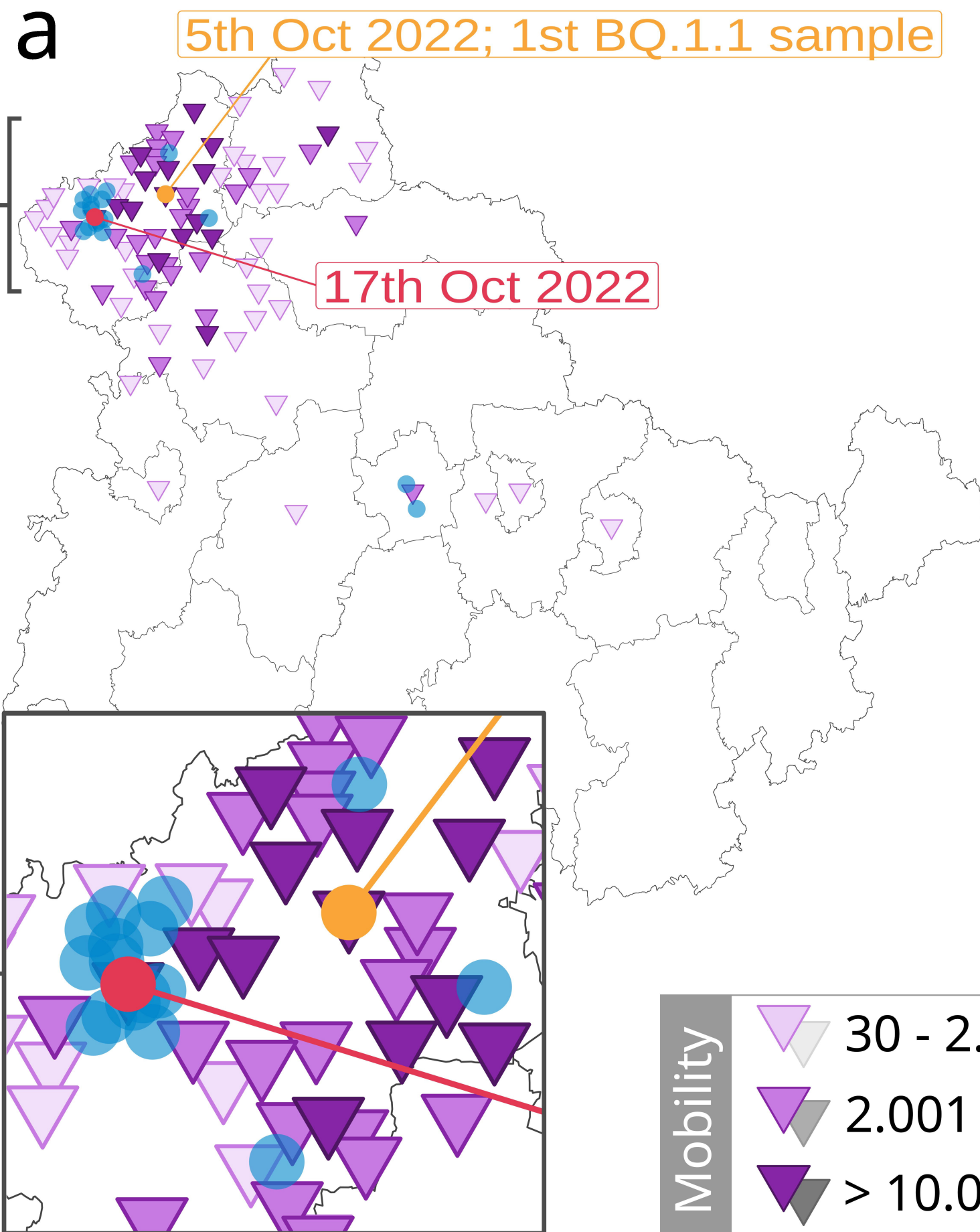
No sample isolated



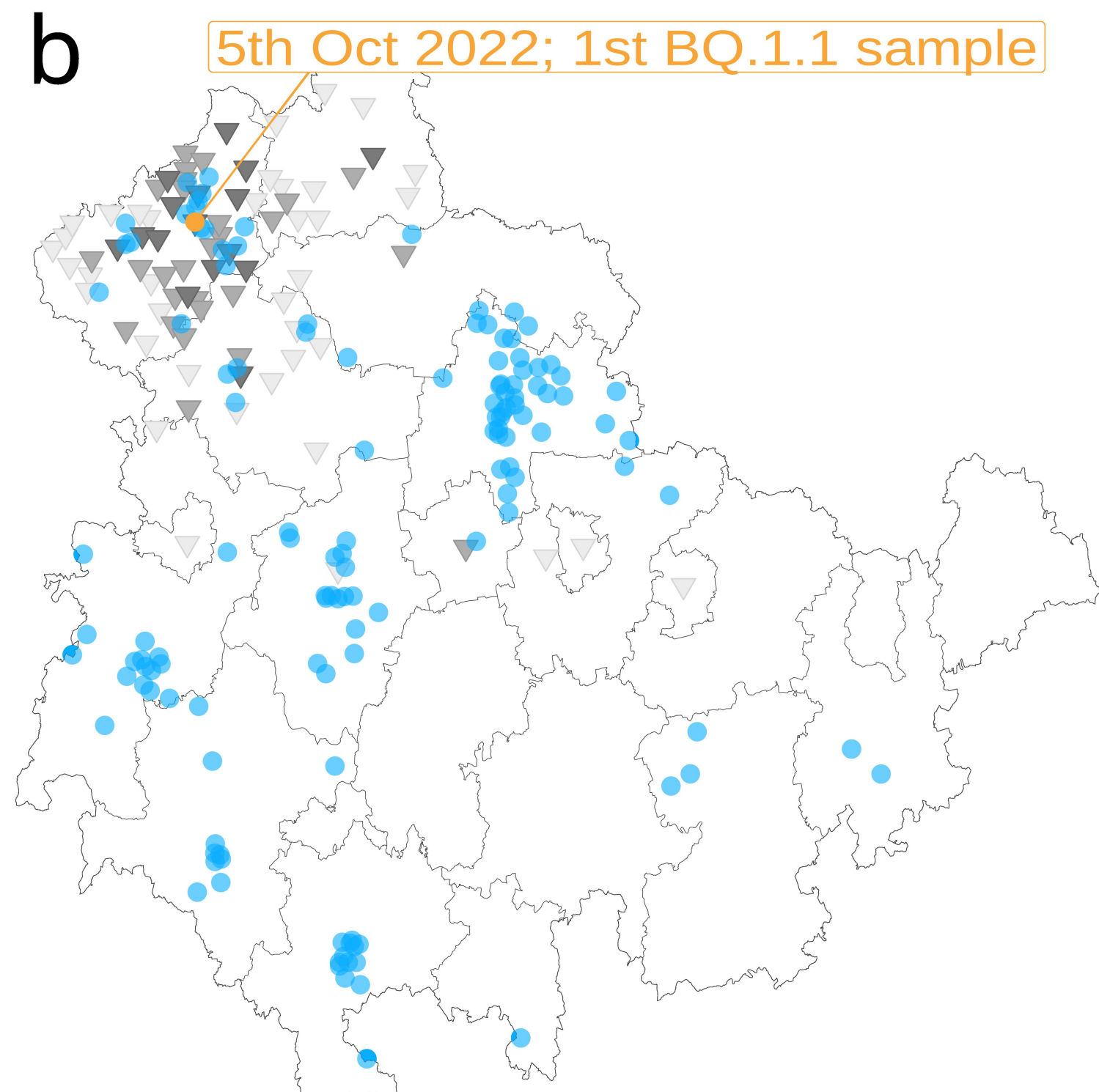
Proportion based on all available SARS-CoV-2 genomes for the respective period.

a**b****Annotations are explained in the text.**

Guided Sampling (n = 19)



Randomized Sampling (n = 132)



Randomized Sampling (n = 69)

