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

Riccardo Spott[1,*], Mathias W. Pletz[1,2], Carolin Fleischmann-Struzek[1,2], Aurelia Kimmig[1], Christiane Hadlich[3], Mathias Hauert[3], Mara Lohde[1], Mateusz Jundzill[1], Mike Marquet[1], Petra Dickmann[4], Ruben Schüchner[5], Martin Hölzer[6], Denise Kühnert[7,8], Christian Brandt[1]

[1] Institute for Infectious Diseases and Infection Control, Jena University Hospital, Jena, Germany

[2] Center for Sepsis Control and Care, Jena University Hospital/Friedrich Schiller University Jena, Jena, Germany

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

[4] Department of Anaesthesiology and Intensive Care, Jena University Hospital, Jena, Germany

[5] Thuringian State Authority for Consumer Protection, Department Health Protection

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

[7] Centre for Artificial Intelligence in Public Health Research, Robert Koch Institute, Berlin, Germany

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

* Corresponding Author

Short Title

Mobility data for Surveillance

Keywords

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

Abstract

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

Introduction

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

Bioinformatics workflows developed in Germany, such as poreCov (for Oxford Nanopore data) and CoVpipe2 (for Illumina data), reconstruct SARS-CoV-2 consensus genomes from the sequencing data and prepare the results for upload and submission to the Robert Koch Institute (RKI).[7,8] As the German government's public health and biomedical research institute responsible for disease control and prevention, the RKI collected the genomes via the German Electronic Sequence Data Hub (DESH) and integrated them with additional epidemiological information to provide an up-to-date overview of the ongoing viral spread. For keeping track of the rapid SARS-CoV-2 evolution, PANGO (Phylogenetic Assignment of Named Global Outbreak) provides a standard naming convention based on unique mutation profiles and further criteria, resulting in the classification of over 3,660 lineages (as of August 2023).[9,10] Additionally, the WHO classified important viral lineages as "Variants of Concern" (VOC), "Variants of Interest" (VOI), or "Variants under Monitoring" (VUM), using Greek designations in the past (e.g., "Alpha" (Pango lineage main designation B.1.1.7) or "Omicron" (Pango lineage main designation B.1.1.529)). Further, the WHO also de-escalated former VOCs to reflect the current SARS-CoV-2 variant landscape better. The first defined VOC (now de-escalated), the Alpha lineage, rapidly replaced almost all previously circulating lineages globally by the end of 2020 until the VOC Delta (main lineage B.1.617.2) replaced it in mid-2021.[11–13]

To predict or monitor the rapid viral spread throughout regions, various data types, like travel data, passenger volumes, or passive wastewater monitoring were examined previously.[14–16] Furthermore, different studies explored mobility data with genomic data to retrace the origin and spatial expanse of Alpha or utilized geolocation data to model the spread in metropolitan areas to recreate case trajectories and the impact of mobility restrictions.[17,18] Mobility data was also used in Germany during the pandemic, revealing that lockdowns leave distant parts of the country less connected due to the sharp decline in long-distance travel.[19] These studies focused on analyzing residential movement and contact tracing to evaluate and inform health policies but were not applied to active molecular surveillance.

Here, we investigated whether mobile service data and fine-granular metadata (such as postal codes and genomic data) can help predict the spread of the Alpha lineage or guide the sampling for more targeted genomic surveillance with a focus on the German federal state of Thuringia.

Results

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

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

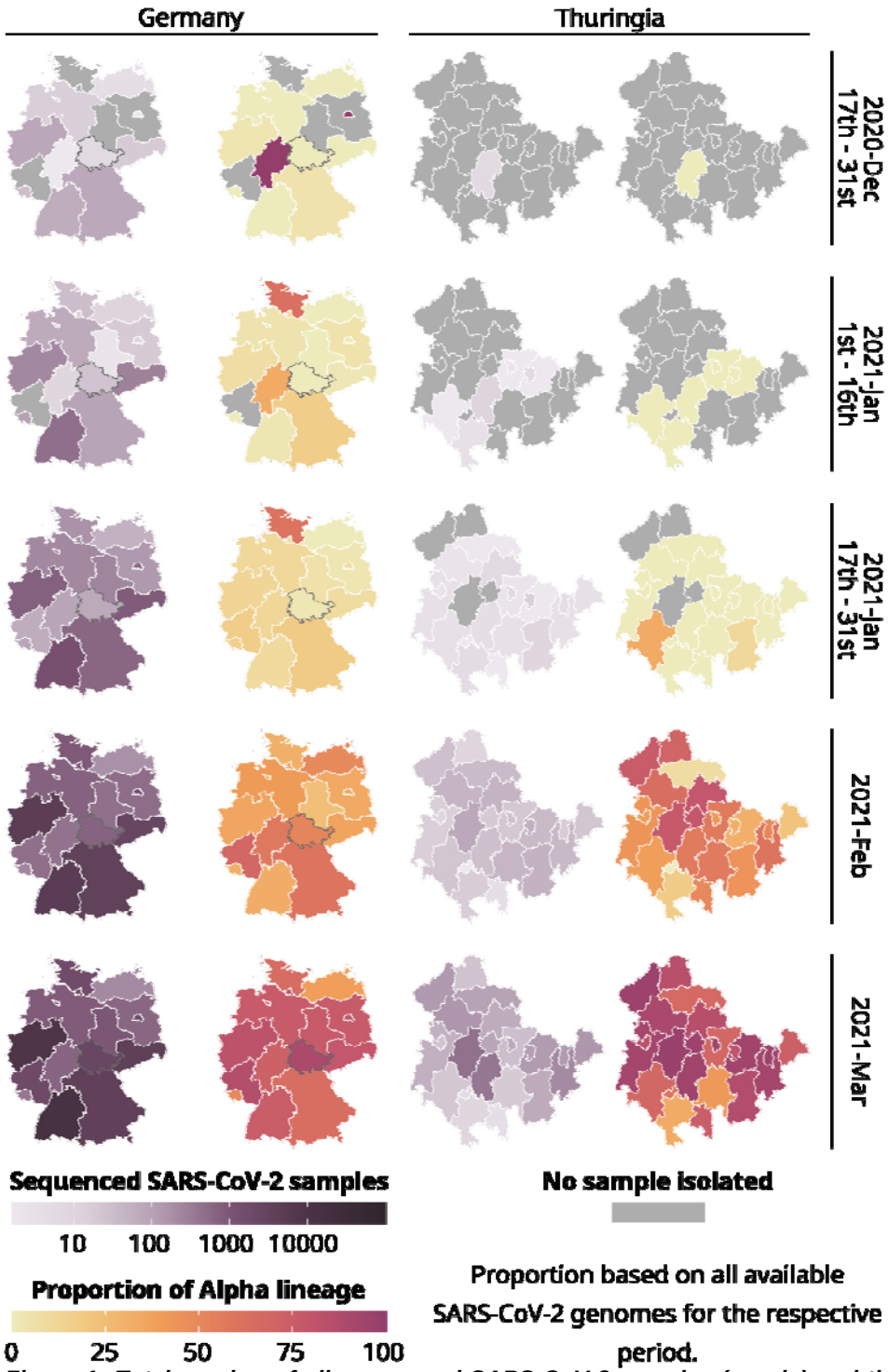


Figure 1: Total number of all sequenced SARS-CoV-2 samples (purple) and the proportion of the Alpha lineage for all sequenced samples (yellow-red) for each state of Germany and each district of Thuringia. *289,487 publicly available German SARS-CoV-2 genomes and their metadata were used for the general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

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

By March, Alpha had spread to nearly all states and districts (districts are similar to counties or provinces) in Germany (Median: 76·47 % Alpha samples among a federal states total sequenced samples compared to 36·03 % in February, excluding Thuringia) and Thuringia (Median: 85·29 %, up from 50·00 % in February). So, there was no noticeable difference in the Alpha proportions between Germany and Thuringia after February. During the summer of June and July 2021, sequencing declined in Germany (including Thuringia; Supplementary Figure S2) due to the decrease in overall daily cases, as reported by Meintrup *et al.* and Oh and Hölzer *et al.*.[20,21]

In summary, the spread of the Alpha lineage in Thuringia lagged roughly two weeks behind the general spread of other German federal states but showed similar proportions. This suggests that Thuringia experienced a delay in the initial arrival of Alpha. However, we did not observe any difference in the overall spread afterward. Thuringia was among the first states to adopt new containment measures, including contact limitations, closure of retail shops, and prohibition of tourist journeys (14th December 2020). Jena, a city in Thuringia, was also the first German city to implement mandatory public masking in March 2020.[22] Contacts were further restricted on January 9th, and people were urged to restrict their movement radius to 15 km, which might explain the delay besides the absence of major airports nearby.

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

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

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

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

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

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

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

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

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

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

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

Mobile service data indicates Alpha subcluster spread and sampling bias

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

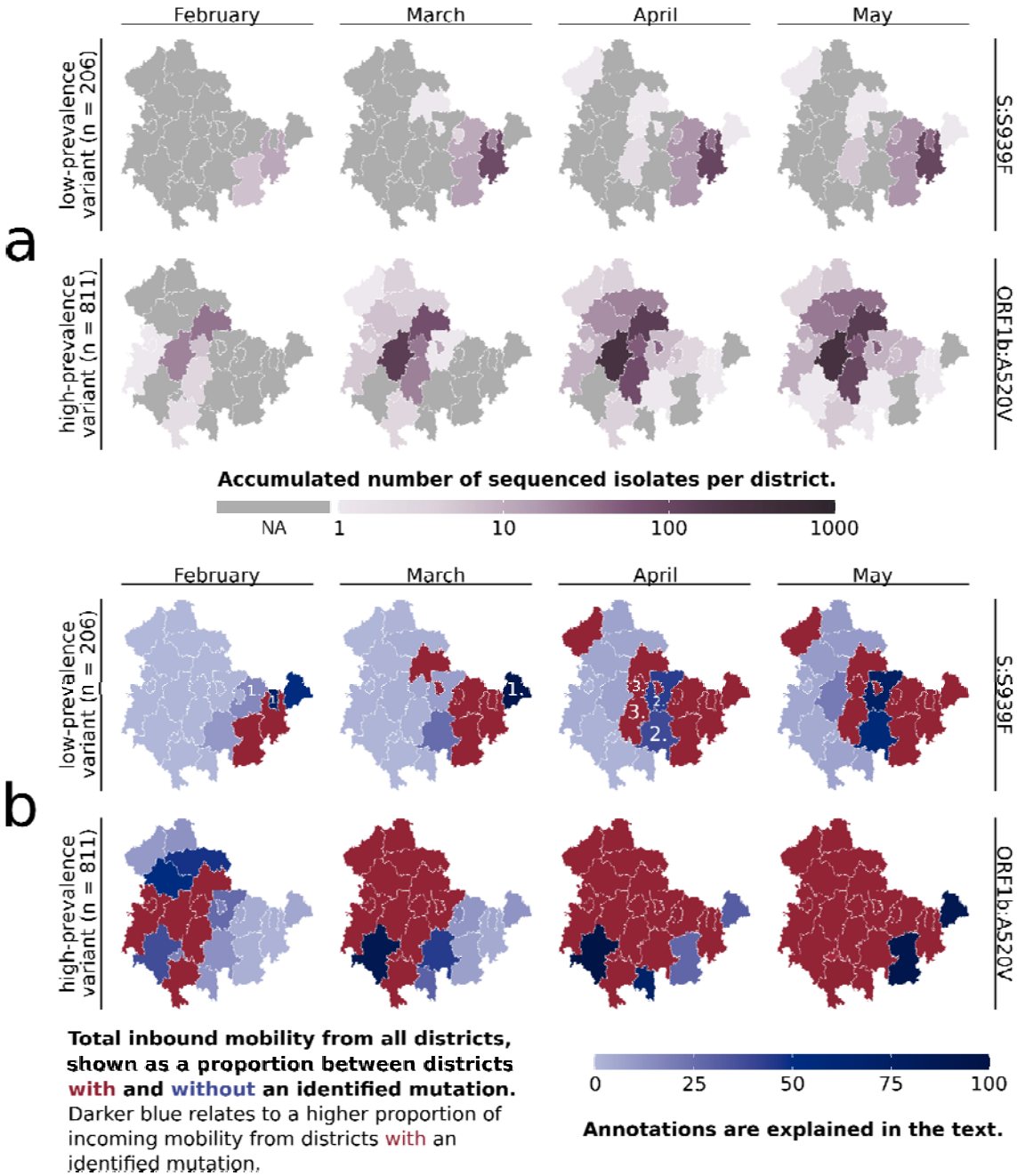


Figure 2: Overview of the subclusters "S:S939F" and "ORF1b:A520V" in Thuringian districts. *a) Accumulated number of sequenced samples for each subcluster per district and per month. b) Combined visualization of each district's "inbound mobility" from other districts (color intensity) and the occurrence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

The mobile service data-based prediction of a subcluster's spread aligned well with the subsequent regional coverage of fast-spreading, highly prevalent subclusters, such as subcluster 7, which covered 811 samples (see Figure 2). In contrast, the predicted spread for the low-prevalence subclusters did not correspond well with the actual occurrence. Yet, adding mobile service data resulted in three different "types" of districts (see Figure 2b, annotated districts). Type 1 included districts with high inbound mobility from areas with an identified variant, where the variant was eventually found afterward, while Type 2 included districts with high inbound mobility from areas with an identified variant, where the variant was never identified. Type 3 included districts not directly connected to a district with an identified variant, but a variant was eventually identified while they border Type 2 district(s).

Our previous analysis of the subclusters' spreading pattern across the districts, based solely on identified variants, indicated missed identifications in some districts due to the seemingly illogical spread to districts without a connection to others (Figure 2a, subcluster "S:S939F"). The inclusion of mobile service data revealed some of these districts to be Type 2 districts. This suggests that the specific variant should be identifiable within these districts due to the observed high incoming mobility from districts with identified variants. Type 2 districts were mainly observed for subclusters with low prevalence and, consequently, low numbers of covered samples that are usually more difficult to monitor. For example, we assumed missing identifications in some districts of subclusters 1, 2, and 3, which, through the mobile service data, are now partially identified as Type 2 districts. However, due to their low prevalence, it is also possible that these subclusters have not spread to the indicated districts.

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

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

Based on our previous findings, we implemented the "mobility-guided" sampling approach under real pandemic circumstances over one month in addition to our active surveillance.

As the subject of investigation, we searched for a newly emerging (based on global news reports) and ideally low prevalent SARS-CoV-2 lineage in Thuringia.

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

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

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

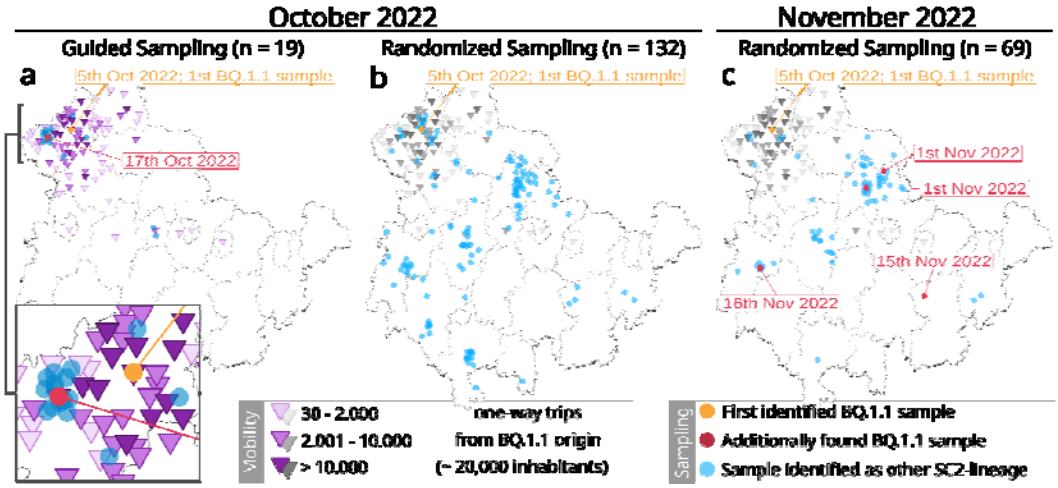


Figure 3: Overview of the mobility-guided sampling of the Omicron sublineage BQ.1.1 in Thuringia (a) compared to the default randomized sampling (surveillance) in October (b) and November 2022 (c). *The randomized surveillance results in November 2022 (c) have been added to highlight the spreading progress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

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

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

past two years and no current data were available. Nevertheless, the available datasets still reflect pandemic movement behavior since the pandemic has been ongoing for two years.

In summary, increasing the sampling depth in the suspected regions successfully identified the specified lineage using only a fraction of the samples from the randomized sampling. Conversely, randomized surveillance, the “gold standard” acting as our negative control, did not identify additional samples with similar sampling depths in regions with no or low incoming mobility or even in high mobility regions with less sampling depth. Implementing such an approach effectively under pandemic conditions poses difficult challenges due to the fluctuating sampling sizes. Although the finding of the sample may have been coincidental, our proof of concept demonstrated how we can leverage the potential of mobile service data for targeted surveillance sampling.

Conclusion

During the SARS-CoV-2 pandemic, diverse data sources like travel, wastewater, and mobility data have been employed in surveillance and transmission tracking.[14–19] In the present study, we analyzed over 296,800 German SARS-CoV-2 genomes to examine whether mobile service data can predict the spatial distribution of the Alpha lineage in the German state of Thuringia and how they potentially benefit pandemic surveillance.

A plausible explanation for the delayed spread of the Alpha lineage in Thuringia is the lack of major transport hubs, as Alpha first occurred in federal states with such hubs. Previous studies have already highlighted the impact of major transportation hubs in the spread of Sars-CoV-2[14,30]. However, its impact on the total distribution is limited, and the spread was ultimately comparable between Germany and Thuringia. While our findings on mobile service data may also apply to Germany, we could not verify this because the limited location data of publicly available German genomes prevented in-depth investigations outside of Thuringia. Thus, precise sampling location data are crucial to utilize mobile service data in genomic surveillance, but privacy regulations may restrict access to this data.

Shortly after Alpha’s emergence, mutation variants formed like the known sublineages Q.1 to Q.8 and the Thuringian subclusters identified by us. This reflects the ongoing evolution during active circulation and indicates an even greater sublineage diversity, which has not been surveyed as closely as in the subsequent Delta and Omicron waves. By monitoring the nine Thuringian subclusters, rather than focusing solely on the parental lineage B.1.1.7, we were again able to effectively track transmissions and gain a comprehensive understanding of the regional spread. So, it underscores the importance of sequencing in pandemic

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

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

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

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

Methods

Sampling

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

microbiology department and 23 randomly sourced by the Thuringian State Authority for Consumer Protection ("Thüringer Landesamt für Verbraucherschutz"; TLV).

Until June 2023, our institute sequenced 3,770 SARS-CoV-2 samples, and SYNLAB Holding Deutschland GmbH, Bioscientia Healthcare GmbH, and DIANOVIS GmbH provided additional 7,800 Thuringian SARS-CoV-2 genomes and their metadata.

Sample preparation and sequencing

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

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

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

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

Alpha spread datasets

From our local database, we extracted 8,397 samples with isolation dates before Oct 1st, 2021. After adding federal state and district information, 993 entries with non-Thuringian locations were excluded. Further, ten entries with unspecific isolation dates were excluded, yielding 7,394 samples (including 6,307 Alpha genomes (lineages B.1.1.7 and Q.1 to Q.8)).

The publicly available RKI SARS-CoV-2 dataset was downloaded, containing 1,091,655 genomes with the respective metadata (17th Oct 2022; Zenodo-version 2022-10-16).[33]

789,405 entries, isolated after Sep'21, and 59 entries without "sending laboratory" information were removed. For the resulting 302,191 entries, location information (location, federal state, district, longitude, latitude) were added based on the sending laboratory postal code. Five entries with a non-existing postal code and all 12,704 Thuringian samples were removed from the dataset, resulting in 289,487 samples (including 137,024 Alpha genomes). We investigated only Alpha-lineage samples collected from September 2020 onwards, after the first official reports of the Alpha-lineage.[11]

Analyzing both datasets, we calculated the monthly proportion of Alpha lineage samples in Thuringia and Germany per state/district, dividing Dec'20 and Jan'21 into first and second halves.

Subcluster identification

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

Mobile service data

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

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

Research in Context

Evidence before this study

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

Added value of this study

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

Implications of all the available evidence

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

Author contributions

Sample collection and preparation, R.Sp, R.Sc., M.L., and M.M.; sequencing, R.Sp., M.L., and M.M.; database setup and maintenance, M.J.; software, C.B.; bioinformatic analysis, R.Sp., M.L., M.M, and C.B.; mobile service data analysis, C.H., and M.Ha.; literature research, R.Sp.; writing first draft, R.Sp.; reviewing and editing manuscript, R.Sp., C.B., 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.; project administration, C.B.; funding acquisition, M.W.P.. All authors have read and agreed to the published version of the manuscript.

Funding

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

Conflicts of interest

The authors have declared that no competing interests exist.

Data sharing statement

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

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

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

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

References

1 Zhu N, Zhang D, Wang W, *et al.* A Novel Coronavirus from Patients with Pneumonia in China, 2019. *N Engl J Med* 2020; **382**: 727–33.

2 An integrated national scale SARS-CoV-2 genomic surveillance network. *Lancet Microbe* 2020; **1**: e99–100.

3 German 'Corona-Surveillanceverordnung' as issued on the 18th of January 2021 and updated on the 27th of June 2022. https://www.bundesgesundheitsministerium.de/fileadmin/Dateien/3_Downloads/C/Coronavirus/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.

4 Meredith LW, Hamilton WL, Warne B, *et al.* Rapid implementation of SARS-CoV-2 sequencing to investigate cases of health-care associated COVID-19: a prospective genomic surveillance study. *Lancet Infect Dis* 2020; **20**: 1263–71.

5 Page AJ, Mather AE, Le-Viet T, *et al.* Large-scale sequencing of SARS-CoV-2 genomes from one region allows detailed epidemiology and enables local outbreak management. *Microb Genomics* 2021; **7**. DOI:10.1099/mgen.0.000589.

6 Lambrou AS, Shirk P, Steele MK, *et al.* Genomic Surveillance for SARS-CoV-2 Variants: Predominance of the Delta (B.1.617.2) and Omicron (B.1.1.529) Variants — United States, June 2021–January 2022. *MMWR Morb Mortal Wkly Rep* 2022; **71**: 206–11.

7 Brandt C, Krautwurst S, Spott R, *et al.* poreCov-An Easy to Use, Fast, and Robust Workflow for SARS-CoV-2 Genome Reconstruction via Nanopore Sequencing. *Front Genet* 2021; **12**: 711437.

8 Lataretu M, Drechsel O, Kmiecinski R, Trappe K, Hölzer M, Fuchs S. Lessons learned: overcoming common challenges in reconstructing the SARS-CoV-2 genome from short-read sequencing data via CoVpipe2. *F1000Research* 2023; **12**: 1091.

9 Rambaut A, Holmes EC, O'Toole Á, *et al.* A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nat Microbiol* 2020; **5**: 1403–7.

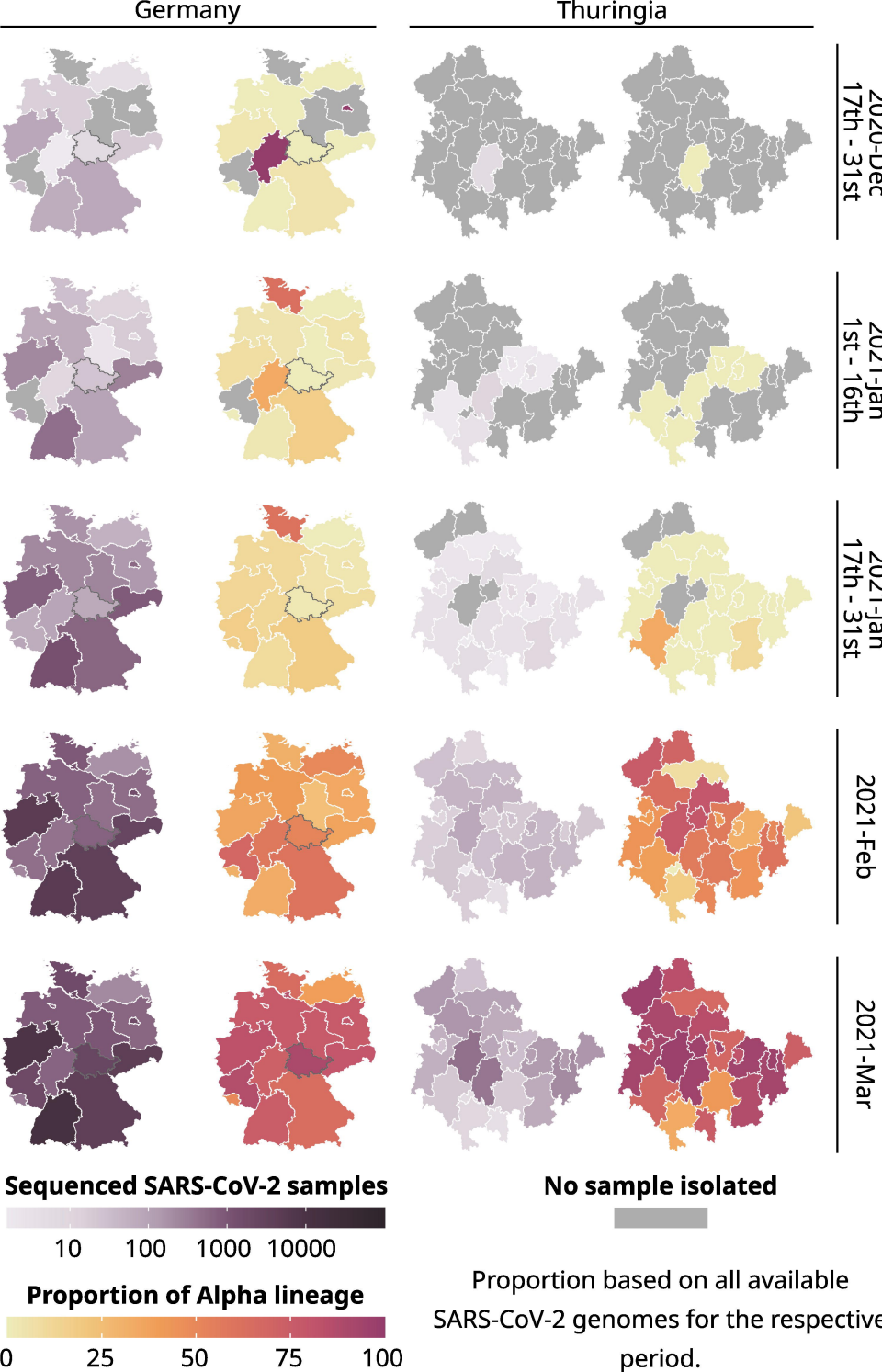
10 PANGO SARS-CoV-2 lineage overview; accessed 18th August 2023. https://github.com/cov-lineages/lineages-website/blob/master/_data/lineage_data.full.json.

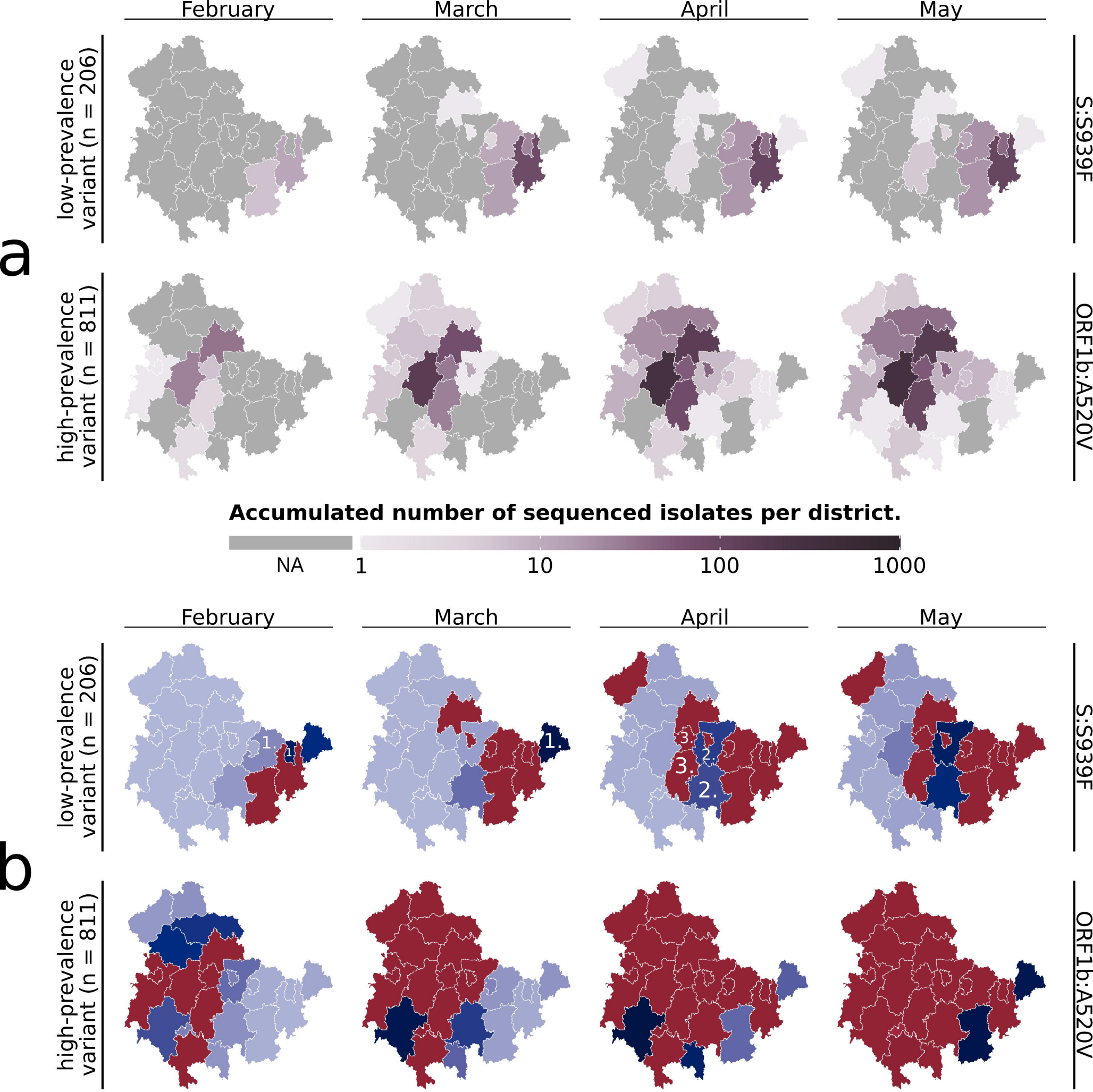
11 Washington NL, Gangavarapu K, Zeller M, *et al.* Emergence and rapid transmission of SARS-CoV-2 B.1.1.7 in the United States. *Cell* 2021; **184**: 2587-2594.e7.

12 Walker AS, Vihta K-D, Gethings O, *et al.* Tracking the Emergence of SARS-CoV-2 Alpha Variant in the United Kingdom. *N Engl J Med* 2021; **385**: 2582–5.

13 Michaelsen TY, Bennedbæk M, Christiansen LE, *et al.* Introduction and transmission of SARS-CoV-2 lineage B.1.1.7, Alpha variant, in Denmark. *Genome Med* 2022; **14**: 47.

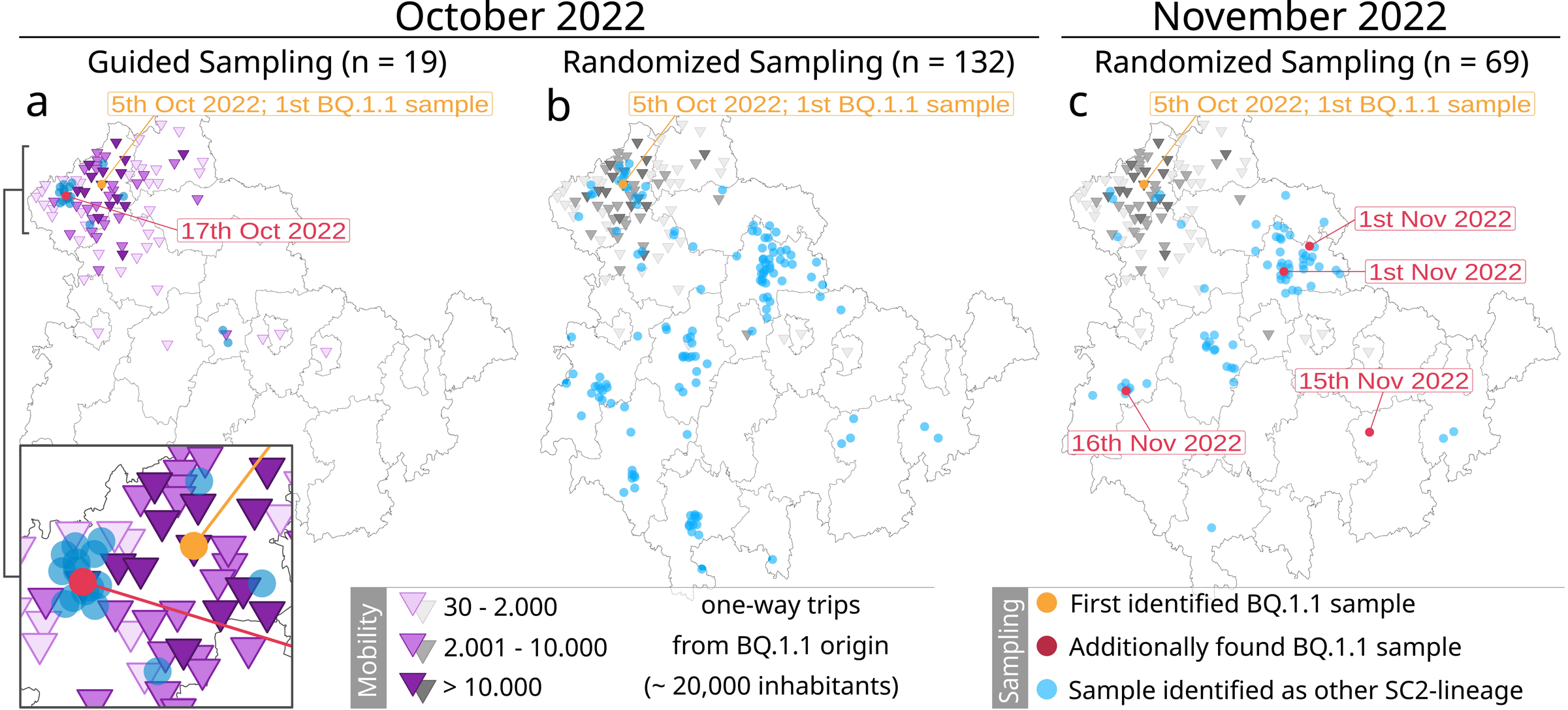
14 Alpert T, Brito AF, Lasek-Nesselquist E, *et al.* Early introductions and transmission of SARS-CoV-2 variant B.1.1.7 in the United States. *Cell* 2021; **184**: 2595-2604.e13.
15 O'Toole Á, Hill V, Pybus OG, *et al.* Tracking the international spread of SARS-CoV-2 lineages B.1.1.7 and B.1.351/501Y-V2 with grinch. *Wellcome Open Res* 2021; **6**: 121.
16 Li J, Ahmed W, Metcalfe S, *et al.* Monitoring of SARS-CoV-2 in sewersheds with low COVID-19 cases using a passive sampling technique. *Water Res* 2022; **218**: 118481.
17 Kraemer MUG, Hill V, Ruis C, *et al.* Spatiotemporal invasion dynamics of SARS-CoV-2 lineage B.1.1.7 emergence. *Science* 2021; **373**: 889–95.
18 Chang S, Pierson E, Koh PW, *et al.* Mobility network models of COVID-19 explain inequities and inform reopening. *Nature* 2021; **589**: 82–7.
19 Schlosser F, Maier BF, Jack O, Hinrichs D, Zachariae A, Brockmann D. COVID-19 lockdown induces disease-mitigating structural changes in mobility networks. *Proc Natl Acad Sci* 2020; **117**: 32883–90.
20 Meintrup D, Nowak-Machen M, Borgmann S. A Comparison of Germany and the United Kingdom Indicates That More SARS-CoV-2 Circulation and Less Restrictions in the Warm Season Might Reduce Overall COVID-19 Burden. *Life* 2022; **12**: 953.
21 Oh DY, Hölzer M, Paraskevopoulou S, *et al.* Advancing Precision Vaccinology by Molecular and Genomic Surveillance of Severe Acute Respiratory Syndrome Coronavirus 2 in Germany, 2021. *Clin Infect Dis* 2022; **75**: S110–20.
22 Pletz MW, Steiner A, Kesselmeier M, *et al.* Introduction of mandatory masking in health care and community: experience from Jena, Germany. *Infection* 2023; published online April 3. DOI:10.1007/s15010-023-02015-w.
23 Aksamentov I, Roemer C, Hodcroft E, Neher R. Nextclade: clade assignment, mutation calling and quality control for viral genomes. *J Open Source Softw* 2021; **6**: 3773.
24 Ozono S, Zhang Y, Ode H, *et al.* SARS-CoV-2 D614G spike mutation increases entry efficiency with enhanced ACE2-binding affinity. *Nat Commun* 2021; **12**: 848.
25 Du Y, Shi R, Zhang Y, *et al.* A broadly neutralizing humanized ACE2-targeting antibody against SARS-CoV-2 variants. *Nat Commun* 2021; **12**: 5000.
26 Liang R, Ye Z-W, Ong CP, *et al.* The spike receptor-binding motif G496S substitution determines the replication fitness of SARS-CoV-2 Omicron sublineage. *Emerg Microbes Infect* 2022; **11**: 2093–101.
27 Kimura I, Yamasoba D, Nasser H, *et al.* The SARS-CoV-2 spike S375F mutation characterizes the Omicron BA.1 variant. *iScience* 2022; **25**: 105720.
28 Asif A, Ilyas I, Abdullah M, Sarfraz S, Mustafa M, Mahmood A. The Comparison of Mutational Progression in SARS-CoV-2: A Short Updated Overview. *J Mol Pathol* 2022; **3**: 201–18.
29 Donzelli S, Spinella F, di Domenico EG, *et al.* Evidence of a SARS-CoV-2 double Spike mutation D614G/S939F potentially affecting immune response of infected subjects. *Comput Struct Biotechnol J* 2022; **20**: 733–44.
30 Tegally H, Wilkinson E, Tsui JL-H, *et al.* Dispersal patterns and influence of air travel during the global expansion of SARS-CoV-2 variants of concern. *Cell* 2023; **186**: 3277-3290.e16.
31 Freed N, Silander O. SARS-CoV2 genome sequencing protocol (1200bp amplicon "midnight" primer set, using Nanopore Rapid kit) v5. 2021 DOI:10.17504/protocols.io.btsrnnd6.
32 Jundzill M, Spott R, Lohde M, Hölzer M, Viehweger A, Brandt C. Managing and monitoring a pandemic: showcasing a practical approach for the genomic surveillance of SARS-CoV-2. *Database* 2023; **2023**: baad071.
33 Koch-Institut R. SARS-CoV-2 Sequenzdaten aus Deutschland. 2022; published online 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
0
25
50
75
100
No sample isolated
Proportion based on all available SARS-CoV-2 genomes for the respective period.



Total inbound mobility from all districts, shown as a proportion between districts with and without an identified mutation. Darker blue relates to a higher proportion of incoming mobility from districts with an identified mutation.

Annotations are explained in the text.


October 2022
November 2022
Guided Sampling (n = 19)
Randomized Sampling (n = 132)
Randomized Sampling (n = 69)
a
b
c
5th Oct 2022; 1st BQ.1.1 sample
5th Oct 2022; 1st BQ.1.1 sample
5th Oct 2022; 1st BQ.1.1 sample
17th Oct 2022
1st Nov 2022
1st Nov 2022
15th Nov 2022
16th Nov 2022
Mobility
30 - 2.000
2.001 - 10.000
> 10.000
one-way trips
from BQ.1.1 origin
(~ 20,000 inhabitants)
Sampling
First identified BQ.1.1 sample
Additionally found BQ.1.1 sample
Sample identified as other SC2-lineage