COVID-19 infection dynamics revealed by SARS-CoV-2 wastewater sequencing analysis and deconvolution

Vic-Fabienne Schumann[1,*], Rafael Ricardo de Castro Cuadrat[1,*], Emanuel Wyler[2,*], Ricardo Wurmus[1], Aylina Deter[3], Claudia Quedenau[3], Jan Dohmen[1], Miriam Faxel[1], Tatiana Borodina[3], Janine Altmüller[3], Frederik Zietzschmann[4], Regina Gnirss[4], Uta Böckelmann[4], Bora Uyar[1], Alexander Blume[1], Vedran Franke[1], Nikolaus Rajewsky[5,#], Markus Landthaler[2,#] and Altuna Akalin[1,#]

1 - Bioinformatics & Omics Data Science Platform, Berlin Institute for Medical Systems Biology, Max-Delbrück-Center for Molecular Medicine Berlin, Berlin, Germany
2 - RNA Biology and Posttranscriptional Regulation, Berlin Institute for Medical Systems Biology, Max-Delbrück-Center for Molecular Medicine Berlin, Berlin, Germany
3 - Genomics Platform, Berlin Institute for Medical Systems Biology, Max-Delbrück-Center for Molecular Medicine Berlin, Berlin, Germany
4 - Berliner Wasserbetriebe, Berlin, Germany
5 - Systems Biology of Gene Regulatory Elements, Berlin Institute for Medical Systems Biology, Max-Delbrück-Center for Molecular Medicine Berlin, Berlin, Germany
* These authors contributed equally to this work.
# Correspondence should be addressed to: altuna.akalin@mdc-berlin.de, markus.landthaler@mdc-berlin.de , rajewsky@mdc-berlin.de

Abstract

The use of RNA sequencing from wastewater samples is proven to be a valuable way for estimating infection dynamics and circulating lineages of SARS-CoV-2. This approach has the advantage of being independent from patient population testing and symptomatic disease courses. However, it is equally important to develop easily accessible and scalable tools which can highlight critical changes in infection rates and dynamics over time across different locations given sequencing data from wastewater. Here we provide an analysis of variant dynamics in Germany using wastewater sequencing and present PiGx SARS-CoV-2, a highly reproducible end-to-end pipeline with comprehensive reports. This complete pipeline includes all steps from raw-data to shareable reports, additional taxonomic analysis, deconvolution and geospatial time series analysis. Using our pipeline on a dataset of wastewater samples from different locations across Berlin from February 2021 to June 2021, we could reconstruct the dynamic of the Variant of Concern (VoC) B.1.1.7 (alpha). Additionally, we detected the unique signature mutation M:T26767C for the VoC delta B.1.617.2 (delta) and its rise in late May. This is around 1 week

earlier than the increase of the proportion of detected delta cases with 6% in the first week of June and 18% in the second week . We also show that SARS-CoV-2 mutation load measured from wastewater sequencing is correlated with actual case numbers and it has potential to be used in a predictive manner. All in all, our study provides additional evidence that systematic wastewater analysis using sequencing and computational methods can be used for modeling the infection dynamics of SARS-CoV-2. In addition, the results show that our tool can be used to identify new mutations and to detect any emerging new lineages of concern. Our approach can support efforts for establishing continuous monitoring and early-warning projects for detecting SARS-CoV-2 or any other detectable pathogen.

Introduction

The ongoing COVID-19 pandemic highlighted the need for better monitoring systems for emerging pathogens and pathogenic variants in order to quickly respond to changing epidemic dynamics. Acknowledging the importance and potential impact of wastewater-borne epidemiological analysis, the European Commission has recently recommended to implement continuous monitoring on SARS-CoV-2 through wastewater in all member states [1]. SARS-CoV-2 is a positive strand RNA virus from the family Coronaviridae, genus *Betacoronavirus* [2, 3] and several studies showed that it can be shed in feces, urine, and saliva [4–6], raising concerns about possible environment-based transmission [7] but also opening up the possibility of Wastewater Based Epidemiology (WBE) vigilance. An alternative to individual patient tests that are expensive and have privacy consent issues, WBE has been used, on a small scale, for different enteric microorganisms such as vaccine and wildtype polioviruses [8], rotaviruses, hepatitis A, astroviruses, adenoviruses, and noroviruses [9]. In the COVID-19 pandemic, wastewater monitoring has been shown to be an effective tool for monitoring incidence rates. Multiple studies showed that it is possible to detect viral RNA even before widespread clinical reports [10–13], suggesting a potential as a monitoring and early alert system.

Several WBE initiatives for SARS-CoV-2 monitoring were established worldwide, and currently, the COVIDpoops19 initiative [14] lists 88 dashboards from 263 universities monitoring 2302 sites. However, most studies are based on RT-qPCR analyses, limited to quantifying the viral titer and/or tracking a few known variants, correlating the results with the reported number of cases in the area. A few studies have been using amplicon sequencing or metagenomics covering the whole viral genome, allowing to track the change of proportions on signature mutations [15–17]. However, quantifying Variants of Concern (VoC) by NGS reads remains challenging, because phasing is difficult with the fragmented sequences generated. Moreover, sequencing and quantifying variants are just the first steps in understanding the dynamics of the outbreaks. The sequencing results should be easily analyzed and combined with geospatial time series analysis. Tracking of VoCs over time and space, can inform policy-making decisions in order to control new outbreaks.

Overall, we aimed to build computational and methodological capacity to monitor emerging SARS-CoV-2 lineages and mutations via wastewater sequencing. For monitoring purposes, we sampled wastewater treatment plants from February 19th to June 10th, 2021. We used the ARTIC protocol [18] for Illumina amplicon sequencing covering the whole SARS-CoV-2 genome to sequence the samples. Finally, we developed a reproducible, open-source pipeline for analyzing continuous sampling of wastewater treatment plants to track signature mutations and Variants of Concern.

Results

A reproducible computational pipeline for tracking SARS-CoV-2 in wastewater

We developed a new pipeline - PiGx SARS-CoV-2 - in the framework of our previously published set of pipelines called PiGx [19]. They are designed with a special focus on usability and reproducibility. The new pipeline was added to the PiGx collection of pipelines and it is distributed together, using GNU Guix (See Figure 1 for a diagram of the workflow). The pipeline comes with all the needed tools and their dependencies and can thus be reproduced on different systems independent of any other installed software.

General description of the PiGx SARS-CoV-2 pipeline

The PiGx SARS-CoV-2 pipeline provides end-to-end data processing and analysis for wastewater RNA sequencing. The pipeline takes the input of targeted sequencing of SARS-CoV-2 RNA with geo-tagged samples. The pipeline takes a set of raw fastq read files, additional processing information for the reads and information about the lineages that should be tracked. After quality check and alignment, the variants are called and annotated. The samples from different timepoints are used to produce time-series reports that track trending mutations over time. We use a particular deconvolution step to also track the proportions of lineages representing Variants of Concern over time. Overall, the pipeline returns a set of reports that provide overviews over lineage and single-mutation abundance in each sample, a taxonomic classification analysis of unaligned reads, and detailed quality control information. Furthermore, all per-sample results are summarized as tables and also combined to visualize time-series and geo-location plots, making the pipeline suitable for continuous sampling.

The pipeline needs local databases (downloaded by the user) for some of the annotation and alignment tools, such as *Ensembl VEP, Kraken2,* and *Krona tools,* while the tools themselves are automatically installed. Furthermore, the user needs to provide: (i) a sample sheet (CSV format) containing information about sampling date and location; (ii) a settings file (YAML format) for specifying the experimental setup and optional custom parameter adjustments, (iii) a mutation sheet containing the lineages of interest and their signature mutations in nucleotide notation and and BED file containing their genomic coordinates; (iv) the reference genome of

the target species (see Methods for a detailed description); (v) BED file containing the PCR primer locations (provided with the pipeline for ARTIC protocol).

To ensure reliable variant calling and robust lineage abundance prediction, the sample has to match stringent quality control measures. For this, information about the sequencing primers, adapters, and also a BED file containing the sites of the signature mutations is necessary. Specifically the latter is important to ensure comparability of the called variants across all processed samples.

Given these input files, the pipeline executes a series of quality check, alignment, variant calling, deconvolution and mutation trend analysis steps. In the end, it provides interactive reports with quality check, geospatial and time-series information for mutations and lineages, as well as downloadable files for the downstream analysis.

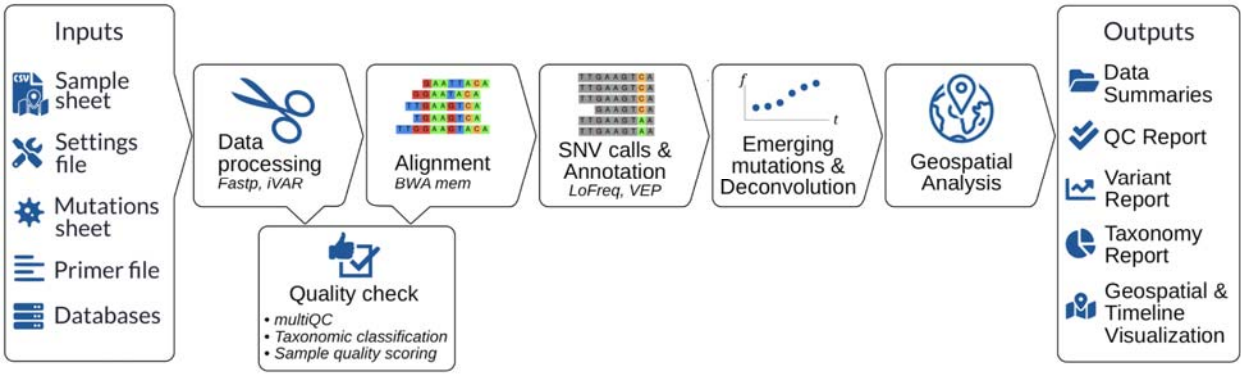


Figure 1: Flowchart of PiGx SARS-CoV-2 pipeline describing required input files, the analysis workflow and used tools and output files.

Berlin wastewater SARS-CoV-2 sequencing and analysis

We sequenced a total of 67,783,582 reads from 38 samples collected at four different wastewater treatment plants in Berlin operated by “Berliner Wasserbetriebe" during a 125 days interval from 06th of February to 10th of June 2021. We analyzed these RNA sequencing results using our pipeline.

The average number of covered signature mutation sites per sample was 81.3 (from a total of 94 tracked signature mutations, see mutations sheet in the Supplementary Table S1). From the 38 samples, 13 samples did not pass the defined quality control threshold (< 90% of the signature mutation sites covered). The Supplementary Table S2 shows the quality control results for each sample.

We were able to align from 22.3% to 99% of the reads to the reference SARS-CoV-2 genome, and the resulting alignments were used for variant calling. We were able to detect a total of 1,907 mutations (from those, 55 are signature mutations) across all the samples (See methods for details on alignment and variant calling). The overall frequency of mutations per sample is

shown on Supplementary Table S3. All counts for the found signature mutations (per sample and pooled) can be found in Supplementary Table S4.

Relationship between genome coverage from wastewater sequencing and case numbers

In order to obtain comparable results for the lineage and mutation analysis over time, a minimum reference genome coverage despite varying infection dynamics is needed. To test the minimal coverage necessary, we compared how much genome coverage we get from our samples over time with the number of COVID-19 cases in Berlin shown in Figure 2. Different sampling locations or their characteristics were not taken into consideration.

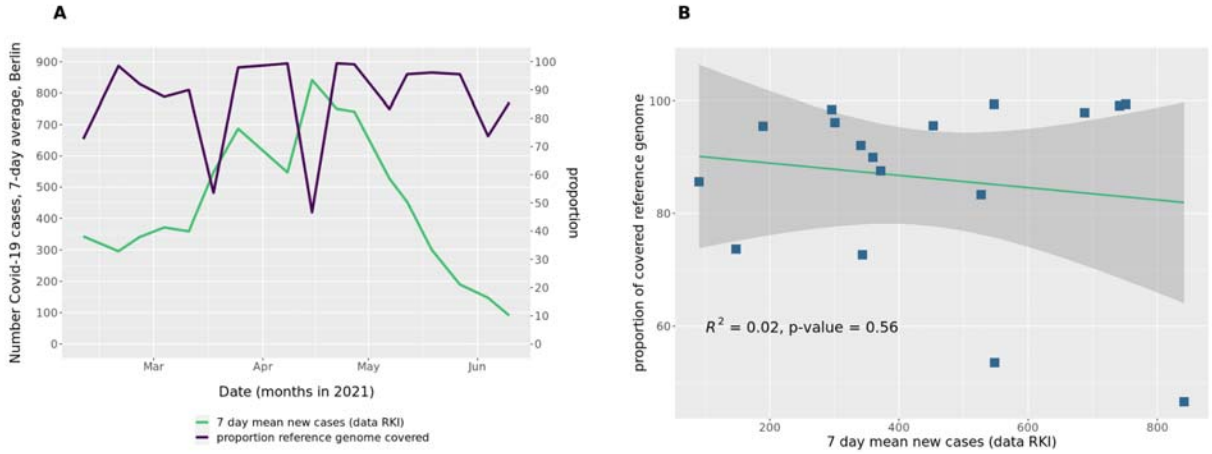


Figure 2: Covid-19 cases in Berlin rolling 7-day average and proportion of covered reference genome. Samples with genome coverage below 40% were discarded because they were technical outliers. R2 was calculated using Pearson correlation. Samples from four different wastewater treatment plants were pooled by day.

Figure 2 B shows that the genome coverage (samples pooled by day) does not correlate (R^2 = 0.023) with the infection dynamic. The genome coverage matches our minimal requirement for the analysis (>=90%) in 56% of the days (see Supplementary Table S5). It is also >=90% (with few exceptions) for most of the sample time (see Supplementary Figure 1) and consistently drops below 90% when cases are fewer than ~150 (Supplementary Table S5). From the data, it is not clear whether the dips in mid-March and mid-April are related to the following decreases in case numbers or sample quality.

Emerging mutations can be teased out from time-series analysis

The time-series nature of the data can be also used to identify trending mutations for SARS-CoV-2 in wastewater. By tracking the frequencies of mutations over time we were able to highlight any mutation which shows strong increasing trends. We applied a linear regression model for each mutation using the date of sampling as the independent variable to identify mutations with strong increasing or decreasing trends over time (see methods for details). We considered significant mutations where t-test p-value was smaller than 0.05 (Supplementary Table S6). Overall, nine mutations were significantly changing over time (Figure 3).

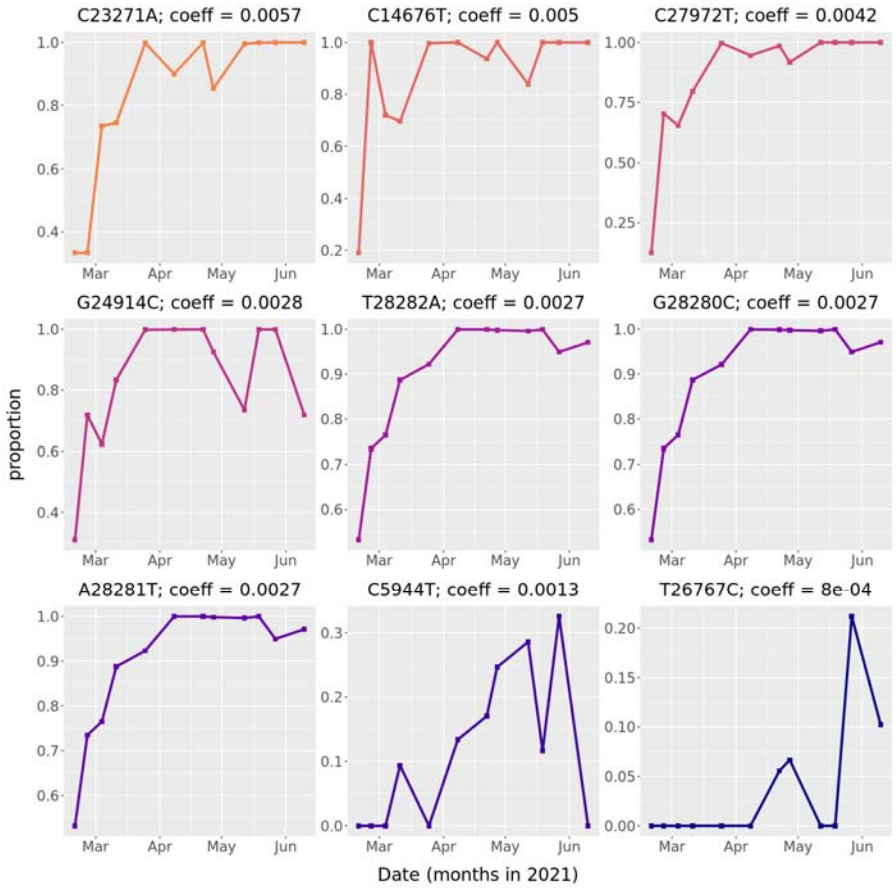


Figure 3: Mutations that significantly increase over time. The mutations were pooled over locations of four different wastewater treatment plants and daytime and sorted by decreasing coefficients from linear models. Statistical significance was evaluated by a t-test using p <= 0.05 as cutoff. Only samples passing the sample quality scoring (> 90% mutation coverage) were used.

Seven of those mutations (denoted here with the resulting protein mutation as translated by *Ensembl VEP* in the following pattern gene:protein-mutation::nt-mutation) S:A570D::C23271A, ORF1ab:P4804-::C14676T, ORF8:Q27*::C27972T, S:D1118H::G24914C, N:D3E::T28282A, N:D3H::G28280C, N:D3V::A28281T are mutations characterizing uniquely for the alpha variant.

The mutation M:I82T::T26767C is characterized uniquely for delta and ORF1ab:D1893-::C5944T is a mutation that is not tracked as a signature mutation for any Variant of Concern. However it is reported as a mutation characterizing a subclade of B.1.1.7 .

Deconvolution of mutation frequencies infers SARS-CoV-2 lineage frequencies

We have also developed the capability to deconvolve the frequencies of VoCs from pooled sequencing reads. Briefly, the deconvolution method uses signature mutations for each VoC and tries to discern the proportions of these variants making up the observed mutation frequencies in the pooled (bulk) sequencing reads obtained from the wastewater. In this study, we tracked four lineages which are currently classified as VoCs: B.1.1.7 (alpha), B.1.351 (beta), P1 (gamma) and B.1.1617.2 (delta). We characterized the lineages with a mutation sheet (Supplementary Table S1) containing signature nucleotide mutations from covidCG [20]. We took a list of mutations with a sequence consensus threshold of 70%. We included mutations that are unique for each lineage as well as mutations that are shared by two or more lineages. However, the pipeline is flexible and can track more variants if the signature mutations are provided in nucleotide format.

Next, we applied this deconvolution method (based on the frequencies of the signature mutations) to infer the proportions of each lineage on each sample (Supplementary Table S7). The lineage frequencies are predicted using a regression model based on the observed frequencies of the signature mutations for each lineage. Additionally, during the deconvolution process, we weighted the tracked lineages differently based on how many signature mutations were found for each of them for a given sample. This step is necessary in order to get more precise predictions of lineages with low abundance and for which only few or only shared mutations were found (See Methods for details).

Figure 4A shows VoC proportion changes over time for each wastewater treatment plant in Berlin. Overall we can see an increase in B.1.1.7 (alpha) that had 29% on February 19th (beginning of sampling) and increased to 92% on June 10th (end of sampling) with a peak of 99% on May 25. Also B.1.351 (beta) increased from zero detection in February to 2% on June 10 peaking on May 27 with 6.8%. The B.1.617.2 (delta) lineage was barely detected with 3% over the sampling time increasing to 4% on June 10. For P1 we can detect a decrease from 8.6 % on February 19 to zero detection in June. Similarly the proportion of the calculated reference strain (labeled as “WT”, see Methods “*Signature matrix construction”* for details) decreased from 55% on February 19th to 3% on June 10 with an intermediate peak of 16% on May 12th. Unpooled results for single locations are attached as Supplemental material.

In order to see if the predicted results can represent the reality we compared the deconvolution results with lineage analysis data published by the Robert Koch-Institute (RKI) for Germany (Figure 4B). Hereby, lineage dynamics for Germany are very comparable to the dynamics within the city of Berlin only. We can see that our predicted lineage frequencies are very similar to the reported frequencies. Only B.1.1.7 shows mostly higher predicted values, but with very similar

trends. Most importantly also the lineages with very low abundance and for which only very few signature mutations where found (data shown in Supplementary Table S4) could be predicted accurately.

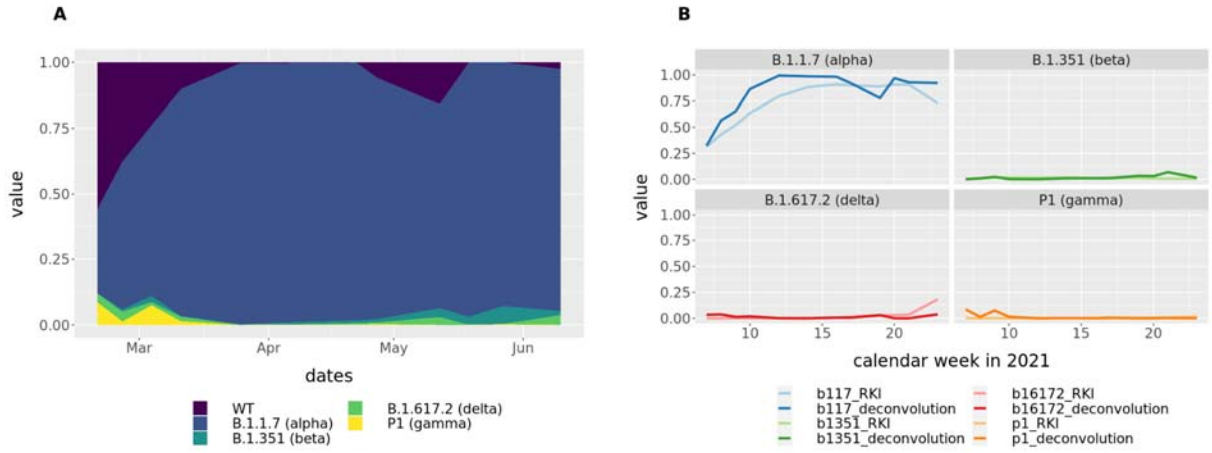


Figure 4: A) Proportion of tracked lineages over time. The proportions were calculated with a deconvolution model based on the signature mutation frequencies. "WT" denotes a set of reference mutations derived from the deconvolution matrix. Sample results were pooled from four different wastewater treatment plants using weighted mean with read number as weights. In case of undistinguishable lineages the proportion derived for the group was distributed equally for the affected lineages. Only samples passing the sample quality scoring (>=90% mutation coverage) were considered. B) Comparison of deconvolution results (dark color) with lineage frequency analysis data from RKI (light color). Deconvolution results were pooled by weeks using weighted mean using sample read numbers as weights. Only samples passing the sample quality scoring (>=90% mutation coverage) were used.

SARS-CoV-2 mutation load in wastewater is correlated with the incidence rate

We hypothesized that more infections would lead to more mutations in the SARS-CoV-2 genome and therefore these two quantities will be correlated even though we calculate the mutation load from wastewater samples rather than the genetic material obtained from patients. In order to test that, we calculated mutation load as the number of non-signature mutations obtained from SARS-CoV-2 wastewater sequencing experiments. We correlated Berlin/Germany case numbers as a measure of incidence rate and mutation load and found a significant association between incidence rate and mutation load obtained from wastewater (adjusted R^2 = 0.35, Pearson's correlation coefficient = 0.63, t-test p-value = 0.03, see Figure 5 and Supplementary Table S8). We have also performed a cross-correlation analysis between case numbers and mutation load with different time lags. The highest correlation with different time lags were not better than the ones without the lag (See Supplementary Figure 1).

We also checked if RT-qPCR results behave the same way using the correlation analysis. For RT-qPCR, we used four pairs of primers for SARS-CoV-2 detection (RT-qPCR) on the wastewater samples. Due to the very low amount of viral particles present overall, we decided for a semi-quantitative approach, instead of using the Ct values, calculating the number of positive detections divided by the number of total reactions carried, grouping all the samples for each day (See Methods for details). The daily percentage of positive qPCR reactions ranges from 0 to 62.5% (Supplementary Table S9). We used the same approach as with the mutation load analysis. We also found positive but no significant correlation with RT-qPCR results and incidence rates (adjusted R^2 = 0.15, coefficient =0.46, t-test p-value = 0.07, See Figure 5 and Supplementary Table S8). In addition, we have also repeated the cross-correlation analysis between incidence rate and RT-qPCR results with different time lags. In this case, lag= -1 week also had positive correlation with the incidence rate (adjusted R2 = 0.25, coefficient = 0.5, t-test p-value = 0.03, See Supplementary Figure 1).

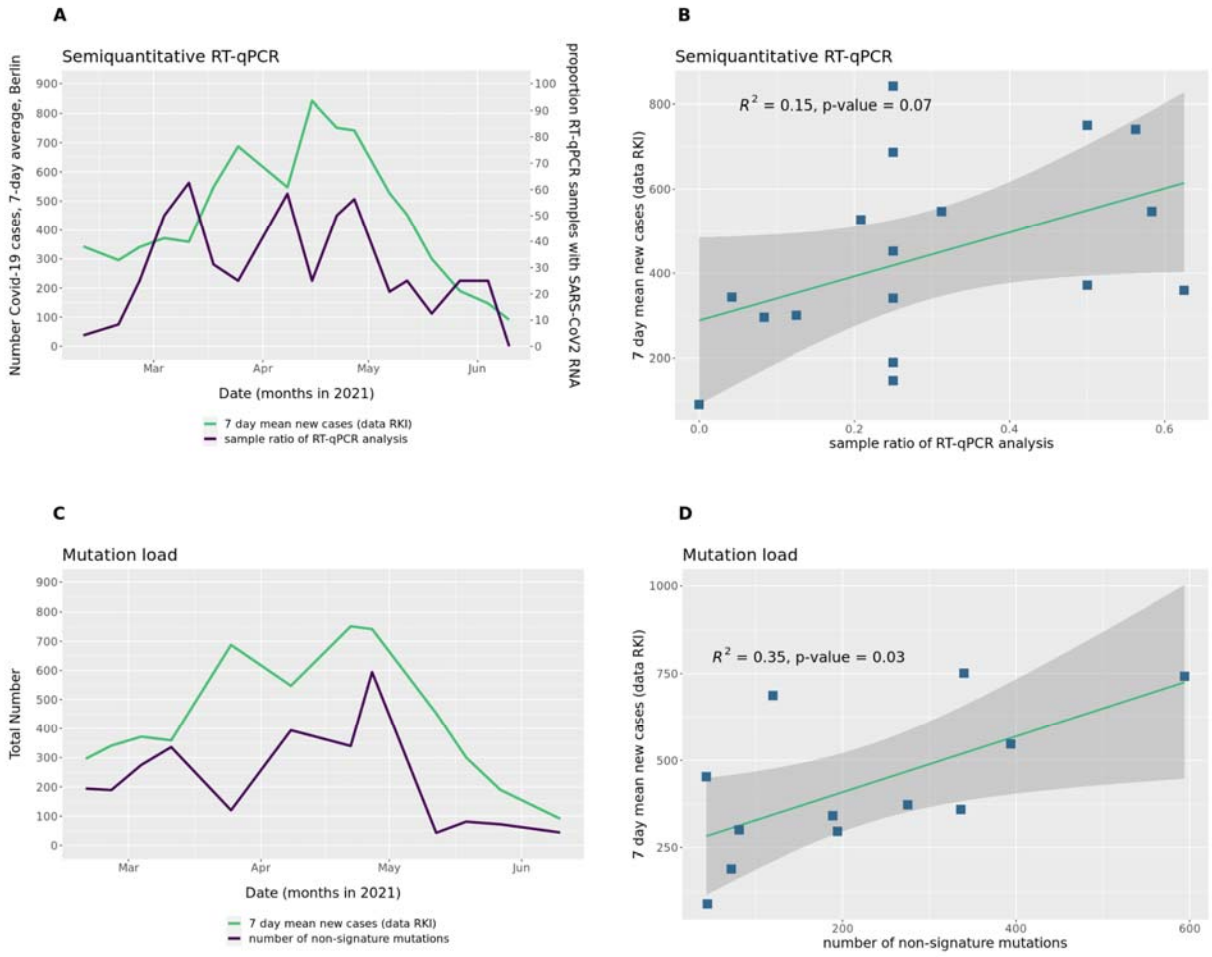


Figure 5: A) 7 days average of COVID-19 cases in Berlin, data from Robert Koch-Institute (RKI) (light green, left axis) and proportion of samples positively determined SARS-CoV-2 RNA by RT-qPCR (dark violet, right axis) over Feb - Jun 2021. B) Correlation of 7 days average of COVID-19 cases in Berlin and proportion of samples with positively determined SARS-CoV-2 RNA by RT-qPCR. C) 7 days average of COVID-19 cases in Berlin (RKI) (light green) and mutation load (number of non-signature mutations detected) (dark violet) over Feb - Jun 2021. D) Correlation of 7 days average of COVID-19 cases in Berlin and mutation load (number of non-signature mutations detected). For all figures samples from 4 different wastewater treatment plants were used and pooled by day.

The proportion of samples with SARS-CoV-2 RNA starts to strongly increase to 25% on February 25th from 8% on February 19th. The number of cases however are only starting to increase from ~360 on March 11th to ~548 on March 18th. This is an offset of nearly three weeks. The second increase in proportion was from 25% on March 25th to 58% on April 8th. And the following case increase was shown from ~547 on April 8th to ~841 cases on April 15th which is an offset of one week. So on average the increase of SARS-CoV-2 RNA in wastewater was shown two weeks earlier than the increase in detected COVID-19 cases (Supplementary Table S9).

Discussion

In many countries like Germany, epidemiological monitoring of SARS-CoV-2 is largely dependent on PCR-based methods without sequencing which is applied on patient samples. These techniques can be used for variant detection only after a concerning new lineage is detected and an appropriate assay was developed. In order to discover new lineages, we need to be able to call mutations of the SARS-CoV-2 genome which can be done using sequencing. However, sequencing-based techniques are deployed on only a fraction of the patient population. Wastewater monitoring emerged as a viable option to track the prevalence of COVID-19 and also for the emergence of different lineages [21] at the population level not only because it is faster and cheaper than sequencing of samples derived from testing but it can also be more representative (without bias through the choice of which samples are going to be sequenced). Furthermore it can also be used to track emerging mutations or lineages of SARS-CoV-2. However, sequencing of SARS-CoV-2 material obtained from wastewater presents data analysis challenges as the samples are potentially from numerous patients, and have lower quality than material obtained directly from patients. In addition, the analytical workflows should be able to deal with samples from multiple locations and time points and combine the information in an easily accessible manner.

In order to address these challenges, we have built a reproducible analytics pipeline that takes in raw sequencing reads and provides sharable interactive reports with geospatial information, and mutation and lineage tracking features over time. In comparison to other commonly used

pipelines for variant analysis like V-pipe [22] or the recommended ARTIC bioinformatics pipeline [23], PiGx SARS-CoV-2 additional features (discussed below) improved usability, reproducibility, and application for environmental samples like wastewater. In addition, the geospatial tracking allows to compare and monitor infection dynamics from different locations (See example reports in Data access section). In terms of usability, the novelty with PiGx SARS-CoV-2 is that the output reports include result visualization for each sample individually and also for the overview and summary of all samples with a choice of visualization methods that are straightforward to interpret. Furthermore, all outputs relevant for the assessment for lineages, quality control and mutations are produced in human-readable format such as HTML reports from which CSV files can be extracted. That makes further data analysis easier by providing formatted tables. Last but not least, PiGx SARS-CoV-2 offers state-of-the-art software reproducibility thanks to GNU Guix [19].

The pipeline comes with built-in flexible quality control metrics since samples from wastewater can have more frequent quality issues. In our analysis, we applied a strict cutoff for genome coverage (>=90%) to reduce noise in our predictions. Our pipeline also allows the user to input their own reference genome and their own set of signature mutations and lineages. As an additional step for QC, we implemented a taxonomic classification of reads that did not align to the SARS-CoV-2 reference genome. Since we used a PCR based protocol, we expect some degree of nonspecific amplifications, so it is of great help to have an additional control by means of the taxonomic classification to assess potential biases [24]. Also since *Kraken2* is a k-mer classifier, this method can reveal reads that match SARS-CoV-2 but are not aligned by stringent alignment tools. This is important to know because it provides insights about potential loss of new mutations missed on the alignment. This step allows the user to investigate potential issues and, if necessary, to adjust the alignment stringency. In our results, the taxonomy report shows a large amount of fragments (~2 million reads across all samples) is mapped to SARS-CoV-2 (Supplement Table S10) and we will investigate this further.

One of the primary features of our approach is built-in tracking of emerging mutations. This feature allowed early prediction of lineages such as B.1.617.2 from a single signature mutation M:I82T::T26767C (Figure 3) in our dataset. We were able to detect the lineage before it was detected in the population (Figure 4 B). This specific mutation was described to be associated with critically increased viral fitness [25]. This analysis and results are also visualized without the need for any additional steps directly in the summarizing report. We showed that our pipeline and its reports can be a valuable tool for early warning predictions and to guide targeted further analysis.

Another key feature of our approach is the deconvolution method that helps us identify the proportion of lineages present in a pooled sample such as wastewater samples. By making use of a weighted regression method we were able to provide accurate estimates of lineage proportions for our samples over time. For the four VoCs that we tracked with signature mutations, we showed in Figure 4 that our model can accurately predict the composition of lineages when comparing with the number of cases reported during the same time frame, even with very low frequencies. It is important to note that the mutations commonly used for tracking

B.1.1.7 in other studies, S:N501Y::A23063T and del69/70 [26, 27] were rare or not found, respectively. However, mutations S:A570D::C23271A, ORF1ab:P4804-::C14676T, ORF8:Q27*::C27972T, S:D1118H::G24914C, N:D3E::T28282A, N:D3H::G28280C, and N:D3V::A28281T were found and mostly important for our predictions. The method predicts the increase of the VoC B.1.617.2 (delta) in June 2021. Supplement Figure 2 shows that the single mutation M:I82T::T26767C—a unique signature mutation of B.1.617.2—can predict the following increase in delta-related case numbers two weeks earlier (see Supplementary Table 11). This increase cannot be seen with similar scale with the results from the deconvolution. Possible reasons are that the deconvolution is influenced by PCR bias and inaccuracies because only few (at most four) signature mutations for the B.1.617.2 lineage were found in June. Conclusively it can be said that there is room for improvement for the model's accuracy on lineages with low abundance, but having the individual mutation detection in place can still guide early detection of increasing appearance of potential new lineages of concerning mutations.

As reported in previous studies in other cities around the globe [28], we showed that also for Berlin the quantification from wastewater can reveal increasing but also decreasing infection dynamics potentially earlier than it is possible from clinical testing. Although RT-qPCR results are not fully quantitative, observing this expected trend was important and paved the way for more robust lineage and mutation trend analysis using sequencing.

Interestingly, we have also shown that mutation load calculated from SARS-CoV-2 wastewater sequencing is predictive of incidence rates. The fact that mutation load is associated with case numbers (incidence rate) is also shown previously on mutation analysis of patient-derived SARS-CoV-2 samples [29]. However, to our knowledge this is the first time we are able to show the link between mutation load derived from wastewater samples and the incidence rates. In addition, in our dataset, mutation load has a slight advantage over RT-qPCR results when we compare the variance explained by the two models. We showed that mutation load from wastewater SARS-CoV-2 samples is at least as predictive as RT-qPCR for incidence rates. However, to make more conclusive statements we need to sample wastewater for a much longer duration and compare the results. Regardless of the methods used on wastewater, as previously published reports also indicate, wastewater monitoring may provide early warning for future case numbers and emerging mutations before these patients hit the healthcare system.

In conclusion, we present a reproducible, comprehensive workflow with a high level of usability that has features for tracking mutations and Variants of Concern over time and geographical locations. We highlight these points using real-world data from Berlin wastewater sequencing samples, and demonstrate the potential to provide more detailed and conclusive insights for SARS-CoV-2 wastewater sequencing efforts.

Methods

Experimental methods

Enrichment of viral particles from raw wastewater and RNA extraction

Raw wastewater samples were collected from four different wastewater treatment plants across Berlin, serving a population of approximately 3.4 mio in total. They were collected as composite two hour samples (8-10pm and 10-12pm) at the primary influent collector at the indicated wastewater treatment plants. Typical characteristics of Berlin wastewater treatment plant influents are 500-1500 mg/L chemical oxygen demand, 200-600 mg/L suspended solids, 40-80 mg/L ammonium-N, 2-8 mg/L orthophosphate-P, 1500-2000 µS/cm electrical conductivity.

Samples were stored and transported at four degrees, and processed about 12 hours after collection. Virus particles were enriched as previously described [30]. About 100ml sample was filtered through 2 glass fiber and 0.2 µM PVDF filters (Millipore, cat# AP2007500 and S2GVU02RE). Of this filtrate, 60 ml was transferred to a 10 kDa cutoff centricon unit, that was previously pre-conditioned with 50 ml ultrapure water and centrifugation with 3000 g for 15 minutes at 4 °C. After centrifugation of the samples for 30 minutes at 4 °C and again 3000 g, the unit was inverted and about 400 µl concentrate was collected by centrifugation for 1000g at 4 °C for 3 minutes. The concentrate was mixed with 3 volumes of Trizol LS (ThermoFisher cat# 10296-010), and the RNA extracted using the Direct-zol RNA miniprep kit (Zymo cat# R2052) including the DNase treatment and elution with 50 µl ultrapure water according to the manufacturer's instruction. Absence of PCR inhibitors was confirmed by mixing the sample 1:1 with total RNA from human cells followed by amplification of a human transcript by RT-qPCR.

Reverse transcription / quantitative polymerase chain reaction (RT-qPCR)

The extracted RNA was amplified using the LunaScript reverse transcription mix (NEB cat# E3010L), with 16 µl RNA and 4 µl reaction master mix according to the manufacturer's instructions, except for a 20 minutes incubation at 55 °C instead of 10 minutes. Afterwards, the cDNA was diluted 1:10 with ultrapure water, and 3.75 µl diluted cDNA used per qPCR reaction, using a SYBR green master mix (ThermoFisher cat# 43-643-46), and final concentrations of 250 nM of the primers on Supplementary Table S12. The presence of the proper amplicon was verified using a 2.5% TAE agarose gel.

ARTIC-seq of the SARS-CoV-2 genome

Amplicon sequencing libraries of the SARS-CoV-2 genome were generated using the ARTIC v3 protocol [18], using 6 µl of the cDNA generated as described above as an input. The primer pools were obtained from IDT. Amplicon libraries were sequenced on an Illumina Miseq or Novaseq device with 2x250 paired-end sequencing and 20% phiX spike-in.

Computational Methods

General Pipeline description

In the first step the pipeline takes the raw reads and the additional information about used primers and adapters to perform extensive quality control. Primer trimming is done with *iVAR* [31] and *fastp* [32] is used for adapter trimming and filtering. In order to make the read quality process comprehensible, *fastQC* reports are generated after each step and summarized with additional *MultiQC* reports. The processed reads are aligned to the reference genome by *BWA Mem* [33] and various coverage statistics are taken by *SAMtools coverage/bedcov* [34]. The alignment is used further for single nucleotide variant (SNV) calling using *LoFreq* [35]. For predicting the lineage abundances, a deconvolution matrix is generated by matching the set of mutations called by *LoFreq* against the provided mutation sheet. The SNVs are translated to protein mutations by *Ensemble VEP* [36]. *Kraken2* [37] is used to get taxonomic classification of the unaligned reads as an additional quality measure and further insight in the samples. A deconvolution method is used to calculate the proportion of lineages (more details in the section *Deconvolution analysis*) for each sample. For summarizing and visualizing the deconvolution results as a time series, samples with SARS-CoV-2 genome coverage below 90% are discarded. For each mutation, linear regression models are used (more details in the section *Regression analysis for mutations*) to detect if any mutation is significantly increasing over time. Here discarded samples were also not included.
For each sample a set of four reports (multiQC, general qc report, taxonomic classification report, lineage report) is generated using *R - markdown* and *knitr.* The R-package of *plotly* is used for generating interactive visualizations. The relevant results across all provided samples are summarized by an extra report that provides insightful visualizations and accessible navigation linking to all the single reports. In this way the pipeline output provides both - an easily accessible overview about lineage and mutation dynamics in a communicable format but also enables extensive data exploration and access to sample wise tables and summaries without the need for running extra scripts. PiGx SARS-CoV-2 uses snakemake [38] for automatic workflow management.

Pipeline accessibility

The pipeline can be installed over GNU Guix and runs with the command [pigx-sars-cov2-ww -s {sample_sheet} {settings_file}]. A cloud version that does not require any installation is also already under development. Alternatively, the pipeline will be available through Docker packages. However, to ensure reproducibility using GNU Guix is recommended [39].

Deconvolution analysis

Model description: With **m** being a system of linear equations build by using **B** being a signature matrix constructed from the signature mutations provided as input and **f** being the proportions for the lineages the deconvolution approach can be represented as **m = f** x **B.** Similar to what has been shown before for deconvolution of cell types from gene expression profiles or methylation profiles [40], we follow the assumption that the frequency of signature mutations corresponds with the frequency of the actual lineage which is characterized by it. The difference in our approach is that we use sequence mutations and apply weights to the signature matrix in order to get more realistic prediction results.

Signature matrix construction: The signature matrix is obtained by matching the set of mutations found in the sample against the set of signature mutations provided as input. In case the mutation sheet contains mutations that are shared between lineages, it is possible that multiple lineages can not be distinguished from each other. In this case, the signature matrix will be deduplicated leaving only one column of the duplicated lineages which will be renamed with the grouped names of all lineages showing this duplicated signature mutation “pattern”.
To make the matrix more robust additional “reference mutations” are added as well as a reference column denoted as “WT”. Bulk frequencies for the “reference mutations” are the difference between 1 and the value of the related signature mutation.
We propose the assumption that the more signature mutations can be found for a specific lineage the higher the probability that this lineage is present with a higher proportion within the sample. We therefore weigh the signature matrix (without the reference mutations) for each lineage with the proportion of signature mutations that has been found for each specific lineage from the total number of signature mutations that was given to characterize it. Applying weights results into less variation and more accurate predictions.

Regression: To deconvolute the lineage abundance we performed robust regression analysis on the signature matrix and the bulk frequency values of the signature mutations using the “Robust Fitting of Linear Models” - rlm() function from the R library MASS [41] (default settings, maxit = 100). Similar to the deconvolution method CIBERSORT [40], we set negative coefficients to 0 and normalized all coefficients to add up to 1 which then form the output value providing the predicted lineage frequency values for the provided lineages and an additional “WT” (reference strain) estimation.
PCR bias as well as the number of signature mutations found influences the robustness of the results. We therefore added the additional constraint to only perform the deconvolution analysis on samples matching a minimum quality score.

Dealing with indistinguishable variants: After deconvolution grouped indistinguishable lineages have to be split again. There are three possible outcomes for those groups:
Firstly, when no signature mutations for a lineage could be found the group includes the “WT” column and is in fact “WT” only. So the grouped lineages are getting the proportion value 0, “WT” get’s the deconvoluted value. Secondly, the grouped lineages are deconvoluted to 0. In this case both lineages are assigned with the value 0. Thirdly, the grouped lineages are not equal to “WT” and are getting a deconvolution value above 0. In this case the assumption for

normal distribution of the lineage abundances is applied and the deconvolution value is divided by the number of grouped lineages. Each lineage is assigned this adjusted value.

Regression analysis for mutations

For the regression analysis on mutation frequencies we applied a linear regression model using the “Fitting Linear Models” - lm - function of *R base* . The test was only performed on mutations if N(x>0) > 5 being the number of frequency values x that are above 0 across all samples. To get only increasing trends, the coefficient values were filtered for values x > 0 only. P-values were calculated by the lm-function using t-test and were filtered for $p < 0.05$. We report the mutation trend analysis together with and sorted by the regression coefficient as a comparable value for unstandardized effect size.

Pooling of samples for time series analysis and plots

For summarizing across daytime and location, the lineage frequencies are pooled by calculating the weighted average using the total number of reads of each sample as weights. The mutation frequencies are pooled by using the simple mean setting removal of missing values to FALSE. Figures and deconvolution plots are done with *ggplot2* [42] . For the cross-correlation analysis samples were pooled by week and the pooled unique set of non-signature mutations was counted.
All details and code can be found on the pipelines repository: https://github.com/BIMSBbioinfo/pigx_sarscov2_ww

Sample scoring for quality check

With the provided BED file for the signature mutations listed in the mutation sheet a coverage analysis is performed using *BEDtools coverage* [43] within the pipeline. For the regression analysis and time series plots only samples are taken in concern that cover more than 90% of all provided signature mutation sites.

Correlation of mutation load analysis to case numbers

We checked if the number of non-signature mutations (here referred to mutation load) correlates with the number of cases in Berlin. For that, we performed a cross correlation analysis using the “CrossCorrelation Function” - ccf() from R “tseries” package (see Supplementary Figure 1) up to $h = \pm 7$. Later we performed a time series intersection to calculate values for cases and mutations or qPCR for lag = -1. We did a standard linear regression on the intersection results using pearson correlation to get R^2.

Reproducible environment

The presented results were produced using PiGx SARS-CoV-2 version 0.03 commit 2603b275106a2a96a422dcfba61554f4d9c0d780. The manifest and guix-version file to create a reproducible environment are provided as supplemental material.

Data access

The data will be deposited to the European Nucleotide Archive (ENA). The interactive report that was used and produced for this pipeline can be found here: https://bimsbstatic.mdc-berlin.de/akalin/AAkalin_pathogenomics/sarscov2_ww_reports/211104_pub_version/index.html

Acknowledgements

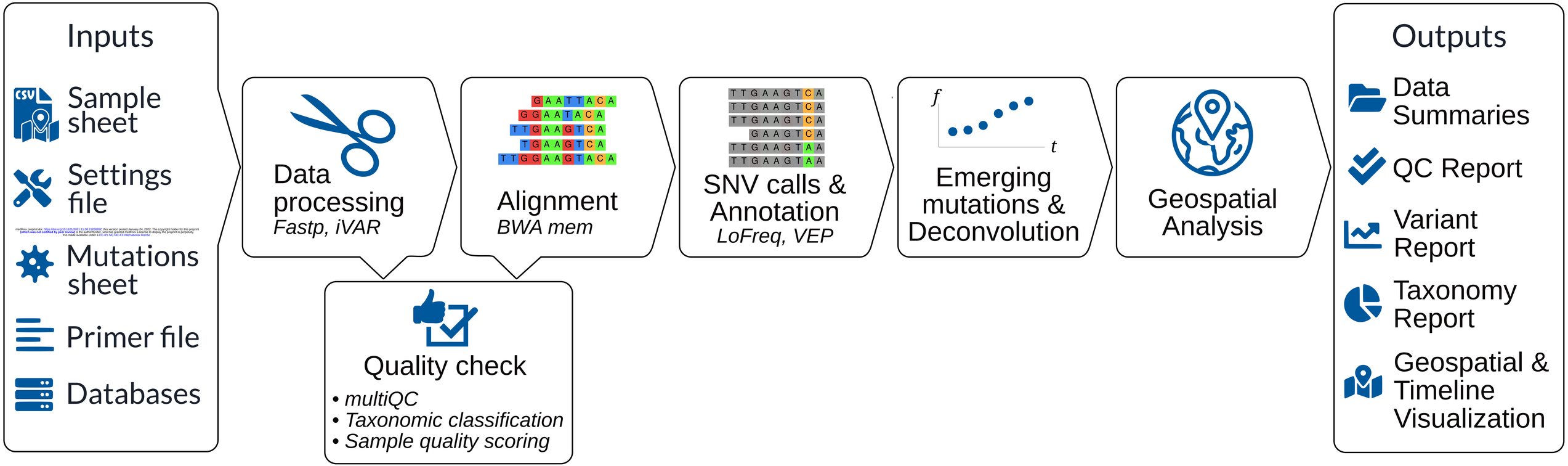
We thank Mrs. Burzyk, Cytner, Darre, Göldner, Grunow, Heinig, Horn, Kapczynski, Klawonn, Koch, Krug, Meyer, Neideck, Schmidt, Schwarzenberg, Stroede, Zühlsdorff, and Messrs. Armbrecht, Dombrowski, Frankenstein, Halatta, Hambarsomian, Linnek, Muss, Flatau, of the Berliner Wasserbetriebe for sampling and logistic support; and Dr. Meixner of amedes as well as Dr. Selinka of the Umweltbundesamt and also Mrs. Schumacher for helpful discussions. Also, we would like to thank Friederike Dündar for consultation on best practices for visualization techniques and Jonas Freimuth for code discussions and support with code development.

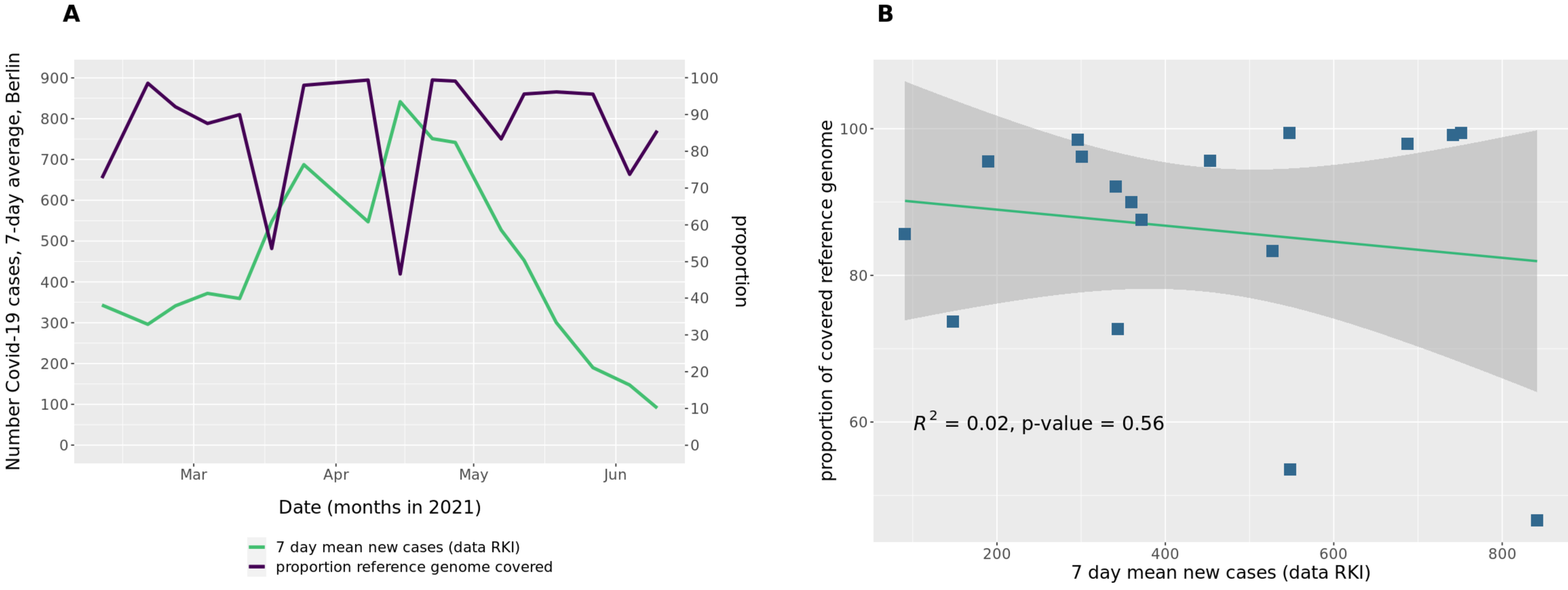
References

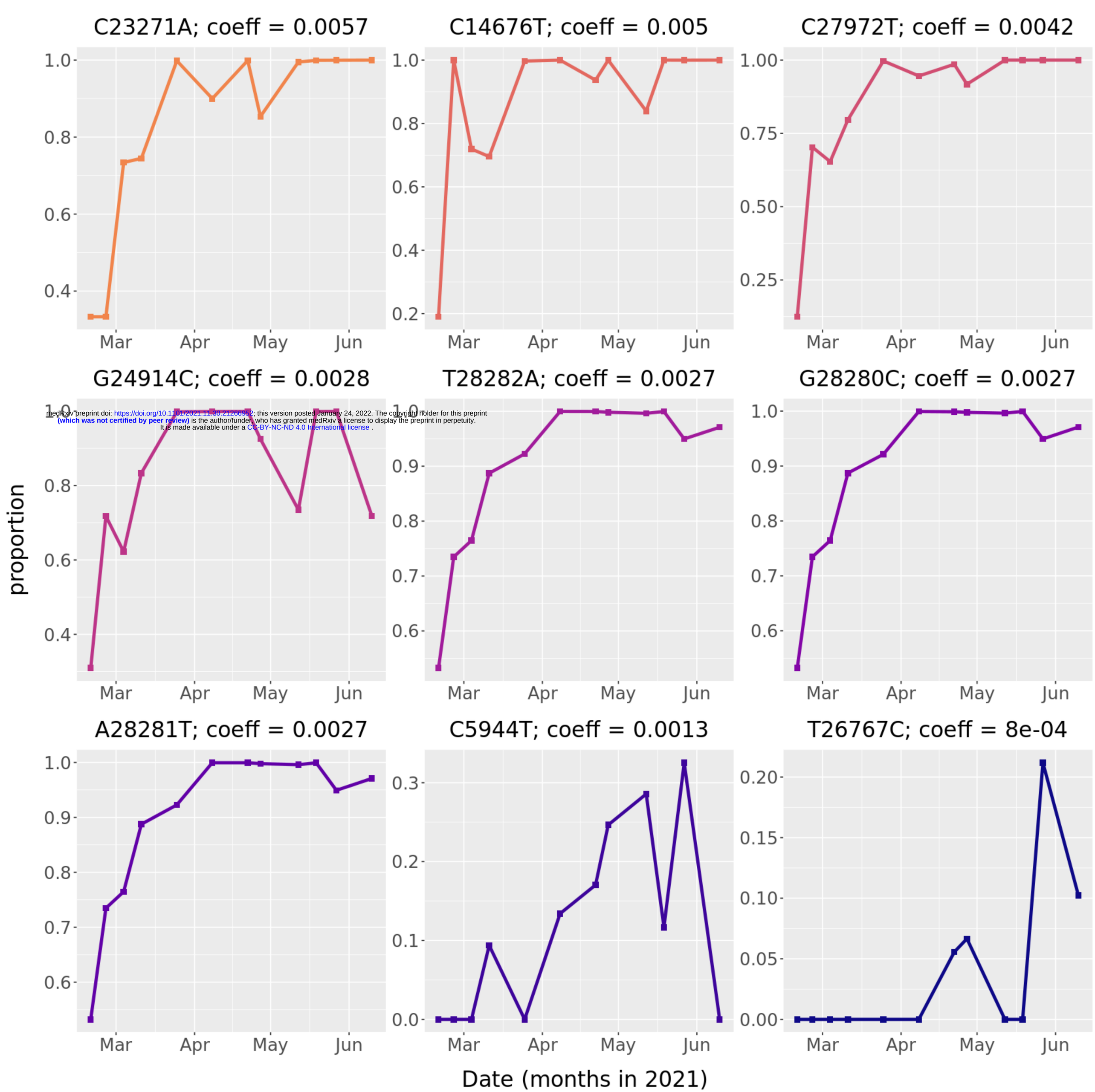
1. European Commission. Commission Recommendation (EU) 2021/472 of 17 March 2021 on a common approach to establish a systematic surveillance of SARS-CoV-2 and its variants in wastewaters in the EU.
2. Fehr AR, Perlman S. Coronaviruses: an overview of their replication and pathogenesis. Methods Mol Biol. 2015;1282:1–23.
3. Zhou P, Yang X-L, Wang X-G, Hu B, Zhang L, Zhang W, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature. 2020;579:270–3.
4. Xiao F, Sun J, Xu Y, Li F, Huang X, Li H, et al. Infectious SARS-CoV-2 in Feces of Patient with Severe COVID-19. Emerg Infect Dis. 2020;26:1920–2.
5. Cevik M, Tate M, Lloyd O, Maraolo AE, Schafers J, Ho A. SARS-CoV-2, SARS-CoV, and MERS-CoV viral load dynamics, duration of viral shedding, and infectiousness: a systematic review and meta-analysis. The Lancet Microbe. 2021;2:e13–22.
6. Boutros J, Benzaquen J, Marquette CH, Ilié M, Labaky M, Benchetrit D, et al. Salivary detection of COVID-19: clinical performance of oral sponge sampling for SARS-CoV-2 testing. ERJ Open Res. 2021;7:00396–2021.
7. Jones DL, Baluja MQ, Graham DW, Corbishley A, McDonald JE, Malham SK, et al. Shedding of SARS-CoV-2 in feces and urine and its potential role in person-to-person transmission and the environment-based spread of COVID-19. Science of The Total Environment. 2020;749:141364.
8. Ranta J, Hovi T, Arjas E. Poliovirus surveillance by examining sewage water specimens: studies on detection probability using simulation models. Risk Anal. 2001;21:1087–96.
9. Petrinca AR, Donia D, Pierangeli A, Gabrieli R, Degener AM, Bonanni E, et al. Presence and environmental circulation of enteric viruses in three different wastewater treatment plants. J Appl Microbiol. 2009;106:1608–17.
10. Wu F, Zhang J, Xiao A, Gu X, Lee WL, Armas F, et al. SARS-CoV-2 Titers in Wastewater Are Higher than Expected from Clinically Confirmed Cases. mSystems. 2020;5.
11. Medema G, Heijnen L, Elsinga G, Italiaander R, Brouwer A. Presence of SARS-Coronavirus-2 RNA in Sewage and Correlation with Reported COVID-19 Prevalence in the Early Stage of the Epidemic in The Netherlands. Environ Sci Technol Lett. 2020;7:511–6.
12. Bar-Or I, Yaniv K, Shagan M, Ozer E, Erster O, Mendelson E, et al. Regressing SARS-CoV-2 sewage measurements onto COVID-19 burden in the population: a proof-of-concept for quantitative environmental surveillance. preprint. Epidemiology; 2020.
13. Xiao A, Wu F, Bushman M, Zhang J, Imakaev M, Chai PR, et al. Metrics to relate COVID-19 wastewater data to clinical testing dynamics. preprint. Infectious Diseases (except HIV/AIDS); 2021.
14. Naughton CC, Roman FA, Alvarado AGF, Tariqi AQ, Deeming MA, Bibby K, et al. Show us the Data: Global COVID-19 Wastewater Monitoring Efforts, Equity, and Gaps. preprint. Public and Global Health; 2021.
15. Crits-Christoph A, Kantor RS, Olm MR, Whitney ON, Al-Shayeb B, Lou YC, et al. Genome Sequencing of Sewage Detects Regionally Prevalent SARS-CoV-2 Variants. mBio. 2021;12.
16. Izquierdo-Lara R, Elsinga G, Heijnen L, Munnink BBO, Schapendonk CME, Nieuwenhuijse D, et al. Monitoring SARS-CoV-2 Circulation and Diversity through Community Wastewater Sequencing, the Netherlands and Belgium. Emerg Infect Dis. 2021;27:1405–15.
17. Landgraff C, Wang LYR, Buchanan C, Wells M, Schonfeld J, Bessonov K, et al. Metagenomic sequencing of municipal wastewater provides a near-complete SARS-CoV-2 genome sequence identified as the B.1.1.7 variant of concern from a Canadian municipality

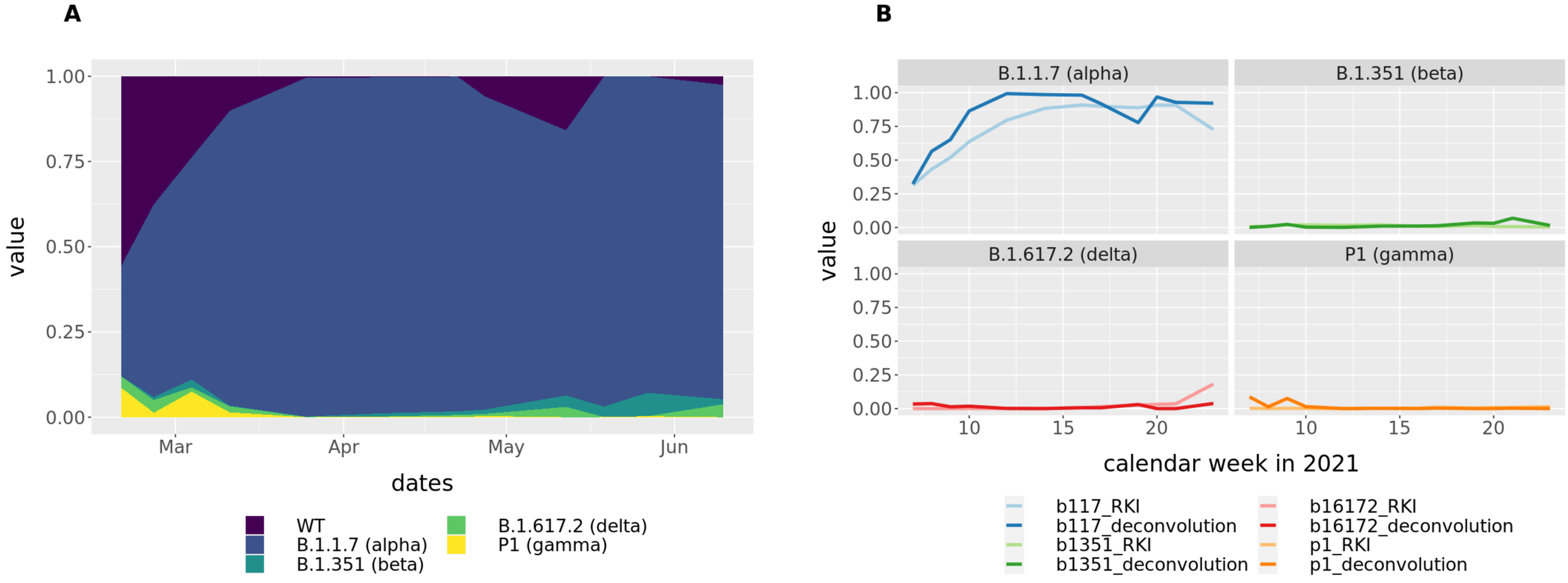
concurrent with an outbreak. preprint. Public and Global Health; 2021.
18. Pipelines R&D D, Farr B, Rajan D, Betteridge E, Shirley L, Quail M, et al. COVID-19 ARTIC v3 Illumina library construction and sequencing protocol v4. preprint. 2020.
19. Wurmus R, Uyar B, Osberg B, Franke V, Gosdschan A, Wreczycka K, et al. PiGx: reproducible genomics analysis pipelines with GNU Guix. Gigascience. 2018;7.
20. Chen AT, Altschuler K, Zhan SH, Chan YA, Deverman BE. COVID-19 CG enables SARS-CoV-2 mutation and lineage tracking by locations and dates of interest. Elife. 2021;10:e63409.
21. Lin X, Glier M, Kuchinski K, Ross-Van Mierlo T, McVea D, Tyson JR, et al. Assessing Multiplex Tiling PCR Sequencing Approaches for Detecting Genomic Variants of SARS-CoV-2 in Municipal Wastewater. mSystems. 2021;6:e01068-21.
22. Posada-Céspedes S, Seifert D, Topolsky I, Jablonski KP, Metzner KJ, Beerenwinkel N. V-pipe: a computational pipeline for assessing viral genetic diversity from high-throughput data. Bioinformatics. 2021;:btab015.
23. Loman N. A quick guide to tiling amplicon sequencing and downstream bioinformatics analysis. 2020.
24. Smyth DS, Trujillo M, Gregory DA, Cheung K, Gao A, Graham M, et al. Tracking Cryptic SARS-CoV-2 Lineages Detected in NYC Wastewater. preprint. Infectious Diseases (except HIV/AIDS); 2021.
25. Shen L, Bard JD, Triche TJ, Judkins AR, Biegel JA, Gai X. Emerging variants of concern in SARS-CoV-2 membrane protein: a highly conserved target with potential pathological and therapeutic implications. Emerg Microbes Infect. 2021;10:885–93.
26. Sandoval Torrientes M, Castelló Abietar C, Boga Riveiro J, Álvarez-Argüelles ME, Rojo-Alba S, Abreu Salinas F, et al. A novel single nucleotide polymorphism assay for the detection of N501Y SARS-CoV-2 variants. Journal of Virological Methods. 2021;294:114143.
27. Vega-Magaña N, Sánchez-Sánchez R, Hernández-Bello J, Venancio-Landeros AA, Peña-Rodríguez M, Vega-Zepeda RA, et al. RT-qPCR Assays for Rapid Detection of the N501Y, 69-70del, K417N, and E484K SARS-CoV-2 Mutations: A Screening Strategy to Identify Variants With Clinical Impact. Front Cell Infect Microbiol. 2021;11:672562.
28. Ahmed W, Bertsch PM, Bivins A, Bibby K, Farkas K, Gathercole A, et al. Comparison of virus concentration methods for the RT-qPCR-based recovery of murine hepatitis virus, a surrogate for SARS-CoV-2 from untreated wastewater. Sci Total Environ. 2020;739:139960.
29. Smith MR, Trofimova M, Weber A, Duport Y, Kühnert D, von Kleist M. Rapid incidence estimation from SARS-CoV-2 genomes reveals decreased case detection in Europe during summer 2020. Nat Commun. 2021;12:6009.
30. Jahn K, Dreifuss D, Topolsky I, Kull A, Ganesanandamoorthy P, Fernandez-Cassi X, et al. Detection and surveillance of SARS-CoV-2 genomic variants in wastewater. preprint. Infectious Diseases (except HIV/AIDS); 2021.
31. Grubaugh ND, Gangavarapu K, Quick J, Matteson NL, De Jesus JG, Main BJ, et al. An amplicon-based sequencing framework for accurately measuring intrahost virus diversity using PrimalSeq and iVar. Genome Biol. 2019;20:8.
32. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 2018;34:i884–90.
33. Li H. Aligning sequence reads, clone sequences and assembly con*gs with BWA-MEM. figshare; 2014.
34. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25:2078–9.
35. Wilm A, Aw PPK, Bertrand D, Yeo GHT, Ong SH, Wong CH, et al. LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. Nucleic Acids Res. 2012;40:11189–201.
36. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The Ensembl Variant Effect Predictor. Genome Biol. 2016;17:122.

37. Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2. Genome Biol. 2019;20:257.
38. Mölder F, Jablonski KP, Letcher B, Hall MB, Tomkins-Tinch CH, Sochat V, et al. Sustainable data analysis with Snakemake. F1000Res. 2021;10:33.
39. Courtès L, Wurmus R. Reproducible and User-Controlled Software Environments in HPC with Guix. arXiv:150602822 [cs]. 2015.
40. Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y, et al. Robust enumeration of cell subsets from tissue expression profiles. Nat Methods. 2015;12:453–7.
41. Venables WN, Ripley BD. Modern Applied Statistics with S. 4. ed., [Nachdr.]. New York: Springer; 2010.
42. Wickham H. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York; 2016.
43. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26:841–2.

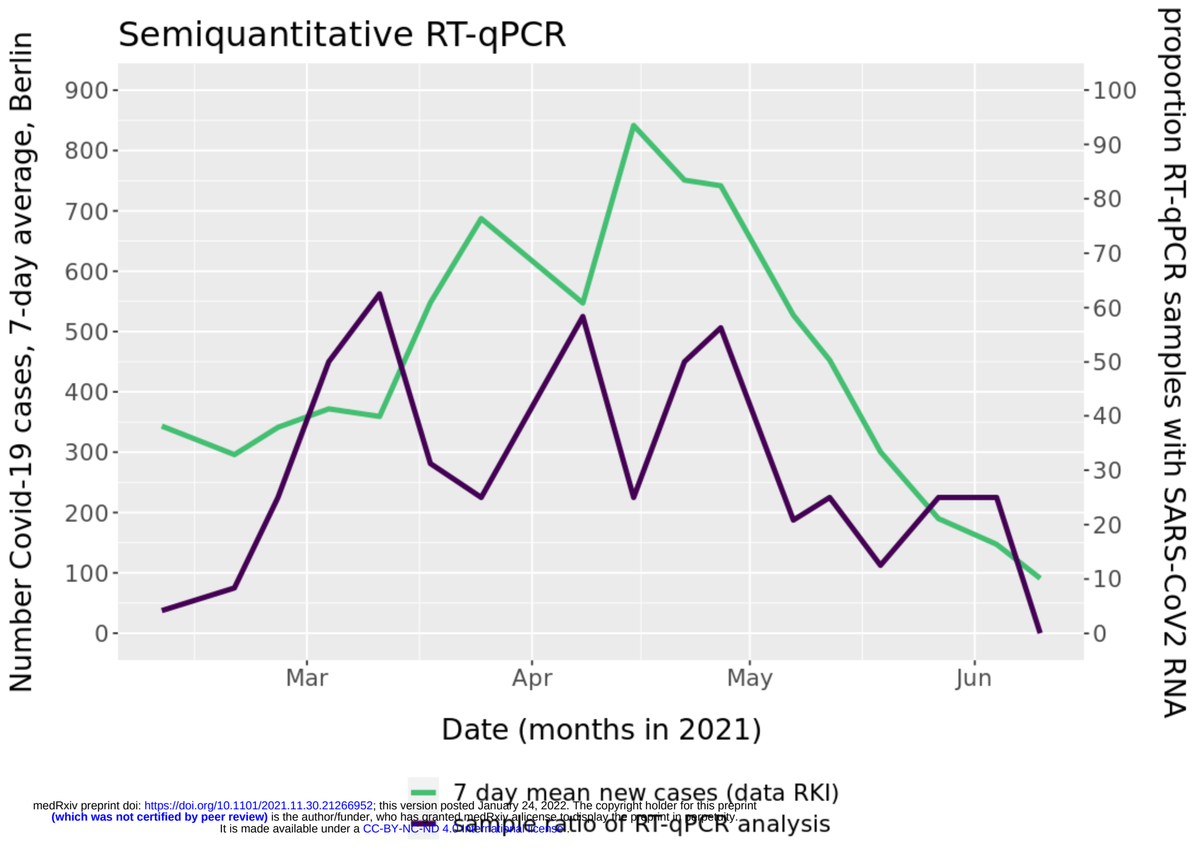



A
Number Covid-19 cases, 7-day average, Berlin
proportion
Mar
Apr
May
Jun
Date (months in 2021)
7 day mean new cases (data RKI)
proportion reference genome covered
B
proportion of covered reference genome
R^2 = 0.02, p-value = 0.56
7 day mean new cases (data RKI)

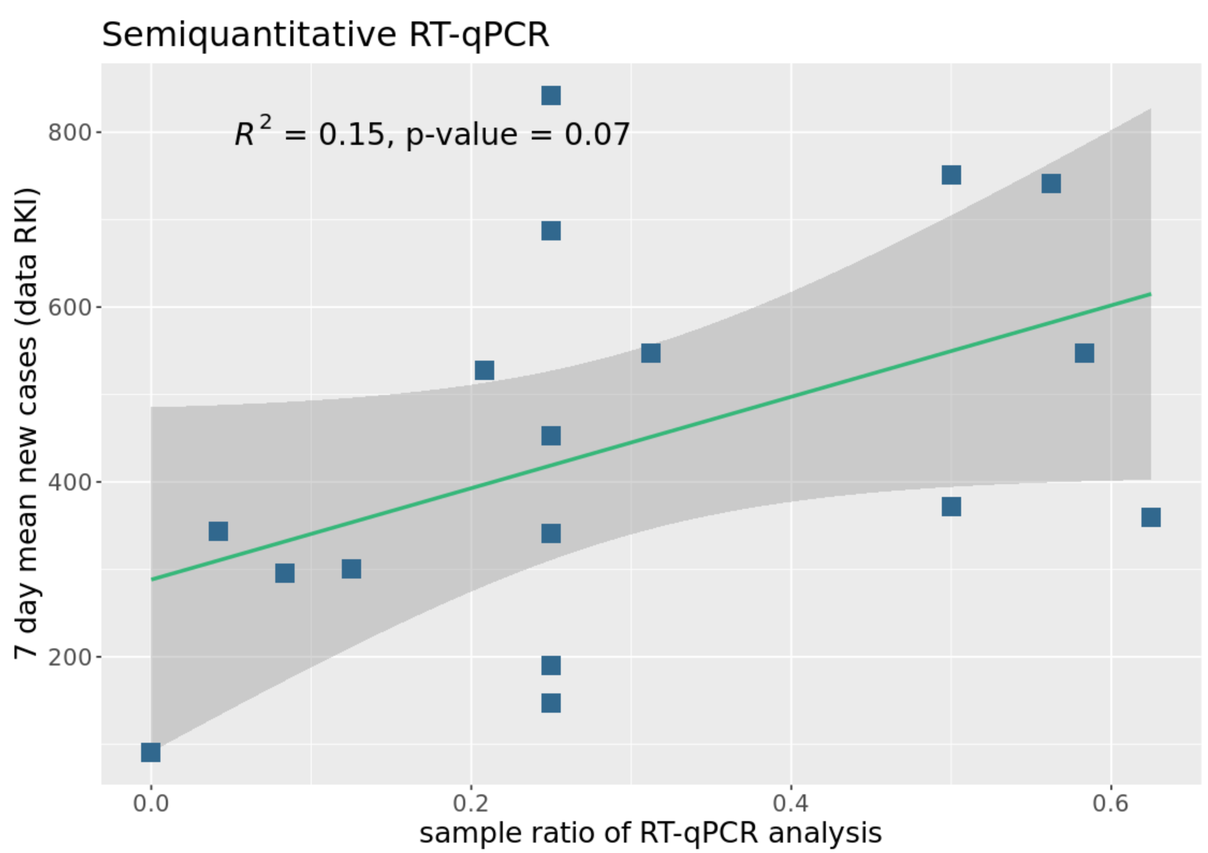



A
1.00
0.75
0.50
0.25
0.00
value
Mar
Apr
May
Jun
dates
WT
B.1.1.7 (alpha)
B.1.351 (beta)
B.1.617.2 (delta)
P1 (gamma)
B
B.1.1.7 (alpha)
B.1.351 (beta)
B.1.617.2 (delta)
P1 (gamma)
10
15
20
calendar week in 2021
b117_RKI
b117_deconvolution
b1351_RKI
b1351_deconvolution
b16172_RKI
b16172_deconvolution
p1_RKI
p1_deconvolution

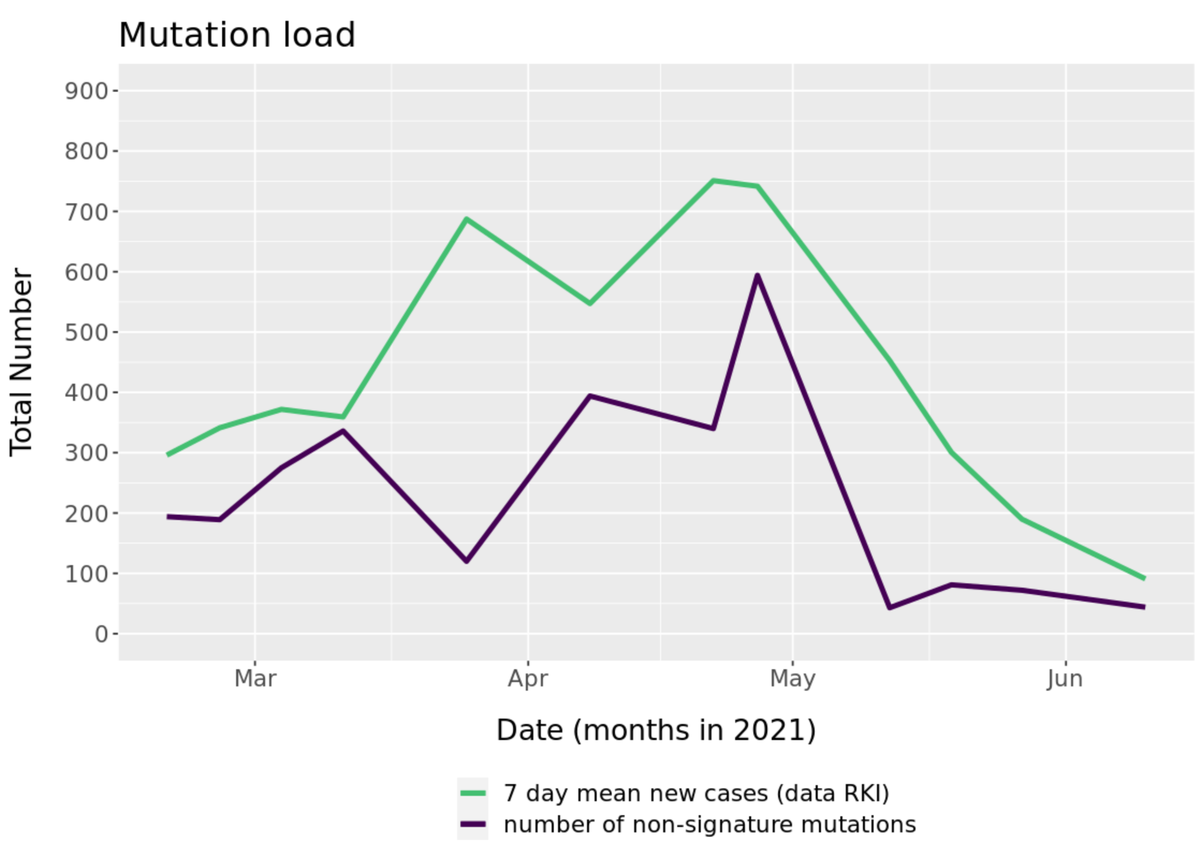
A



B



C



D