Tracking SARS-CoV-2 genomic variants in wastewater sequencing data with *LolliPop*

Authors: David Dreifuss[1,2], Ivan Topolsky[1,2], Pelin Icer Baykal[1,2], Niko Beerenwinkel[1,2,*]

Affiliations:
[1]Department of Biosystems Science and Engineering, ETH Zurich, CH-4058 Basel, Switzerland;
[2]SIB Swiss Institute of Bioinformatics, CH-1015 Lausanne, Switzerland;

***Corresponding Author**

Abstract

During the COVID-19 pandemic, wastewater-based epidemiology has progressively taken a central role as a pathogen surveillance tool. Tracking viral loads and variant outbreaks in sewage offers advantages over clinical surveillance methods by providing unbiased estimates and enabling early detection. However, wastewater-based epidemiology poses new computational research questions that need to be solved in order for this approach to be implemented broadly and successfully. Here, we address the variant deconvolution problem, where we aim to estimate the relative abundances of genomic variants from next-generation sequencing data of a mixed wastewater sample. We introduce *LolliPop,* a computational method to solve the variant deconvolution problem by simultaneously solving least squares problems and kernel-based smoothing of relative variant abundances from wastewater time series sequencing data. We derive multiple approaches to compute confidence bands, and demonstrate the application of our method to data from the Swiss wastewater surveillance efforts.

Introduction

During the COVID-19 pandemic, genomic surveillance has been applied at an unprecedented scale to support various national efforts in containing outbreaks [1]. In this context, wastewater monitoring has seen its broadest and most successful application: PCR-based surveillance of total viral load has been deployed successfully in a large number of surveillance campaigns, some of which[1] are now complementing this piece of information with next-generation sequencing (NGS) data to distinguish different genomic variants [2]. As genomic analysis is extended from clinical samples to samples from wastewater treatment plants (WWTPs), new statistical and computational research questions arise [1]. Beyond SARS-CoV-2, and in the contemporary context of the emergence and resurgence of viral threats, such as, for example, monkeypox, wastewater-based epidemiology (WBE) is becoming a central tool for pathogen surveillance in general [2]. It is therefore pressing that the computational challenges relating to WBE be addressed.

Most of the existing viral genomic data analysis pipelines and tools were designed for clinical samples and rely on classifying the majority variant of each sample from the consensus sequence of the read alignment [3–5]. As wastewater samples consist by nature of a heterogeneous mix of genomic variants, this approach is ill-advised and can at best result in a major information loss and at worst produce false and misleading results. Detecting minor variants in a wastewater sample corresponds to detecting minor variants in a population and is therefore of great interest because it can enable early detection of an emerging lineage before it becomes dominant. One of the main challenges in the analysis of wastewater-derived NGS data is the loss of mutation phasing information, which can result both from fragmentation of the genetic material and from the (usually tiling-amplicon PCR-based) sequencing protocols used. In addition, the sequencing data exhibits very high levels of overdispersed

---

[1] see for example:
https://bsse.ethz.ch/cbg/research/computational-virology/sarscov2-variants-wastewater-surveillance.html
https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1043807/technical-briefing-33.pdf

noise. From a large pool of raw wastewater, extreme downsampling steps (grab or composite sampling, filtering, random reverse transcription, etc.) are followed by extreme amplification steps (PCR). Some tools have been developed to increase sensitivity in the detection of variants, for example, by looking at co-occurring mutations on the same read, which has been shown to improve early detection of variants [6].

Beyond early detection, quantitative estimation of the relative abundances of different variants from wastewater is a very important endeavor. Tracking the relative abundance through time of a new variant can inform on its fitness advantage relative to the dominating strain [7], and hence on its predicted impact on the infection dynamics. It has been shown that this epidemiologically important parameter can be estimated accurately from wastewater samples using either specific PCR-based assays [8,9] or NGS data [6], while using far fewer samples as compared to using clinical data. For practical planning and policy making, it is therefore crucial to have accurate and time-efficient methods for estimation of the relative abundances through time of variants from wastewater NGS data, including reliable measures of uncertainty.

Prior research has shown that the relative abundance of an emerging variant can be accurately quantified by averaging over a set of mutations that is unique for this variant [6]. However, as the number of variants grows, shared mutations cannot generally be discarded as finding sets of unique, characteristic mutations for each variant quickly becomes impossible. To address this limitation, some methods that take into account the correlation structure of mutations between different variants have been developed [10–13]. However, with these methods, quantification of uncertainty is done on the basis of the computationally intensive bootstrap, which can be prohibitive when faced with large amounts of data or limited computational resources.

Here, we introduce a new method for solving the variant deconvolution problem, named *LolliPop*. Tailored to time-series data, our method implements simultaneous deconvolution and kernel smoothing of the variant relative abundances. In addition to confidence bands based on the bootstrap, we derive analytical methods to compute asymptotic confidence bands, which provide a 30-fold speedup. We evaluate our method by comparing to clinical data from [6]. *LolliPop* is currently used for the Swiss wastewater monitoring program[2]. It is available as a Bioconda package.

Methods

Variant deconvolution

We consider an ordered collection of samples, taken at (not necessarily evenly spaced) timepoints $t \in \{1, \ldots, T\}$. Each studied variant $v \in \{1,\ldots, V\}$ carries a subset of the mutations $m \in \{1,\ldots, M\}$ relative to a fixed reference strain. Let $X \in \{0,1\}^{M \times V}$ be the design matrix of variant definitions, i.e., $X_{m,v} = 1$ if the variant v bears mutation m, and $X_{m,v} = 0$ otherwise. Let $y_t \in [0,1]^M$ be the observed mutation frequency vector at time t, where the entries are the observed proportions of reads from the wastewater sequencing experiment supporting a certain mutation m. We are interested in $b_t \in [0,1]^V$, $||b_t||_1 = 1$, the relative variant abundances vector at each time point t. We further make the assumption of a linear probability model, such that at time t, the expected proportion of reads with a given mutation is a linear combination of the relative variant abundances:

$$E[y_t \mid b_t] = Xb_t$$

Definition: For given variant definitions X and a time series of mutation frequencies $y_1, \ldots, y_T$, the variant deconvolution problem is to find the relative variant abundances $b_1, \ldots, b_T$ in the population of the catchment area of the WWTP, such that $y_t = Xb_t$ for all time points *t*.

2 https://bsse.ethz.ch/cbg/research/computational-virology/sarscov2-variants-wastewater-surveillance.html

As finding the exact relative variant abundances in the population is generally not possible due to the randomness of the data-generating process and model misspecification, we relax the problem to finding a best estimate. Solving the variant deconvolution problem is then performed by choosing a loss function L and optimizing:

$$\widehat{b_t} = argmin_{b_t : b_{t,i} \geq 0 \, \forall i, \, ||b_t||_1 = 1} L(y_t - Xb_t)$$

Here, we choose two different loss functions, namely least squares (LS) $L_{LS}(z) = ||z||_2^2$ and the robust soft l_1 loss (SL1) $L_{SL1}(z, \alpha) = \sum_i \left(\frac{1}{\alpha} z_i^2 \cdot 1(|z_i| \leq \alpha) + |z_i| \cdot 1(|z_i| > \alpha) \right)$, but other choices are readily implementable.

Simultaneous smoothing

To deal with the high levels of noise in the observed data, we assume temporal continuity of the variant abundances. Thus, we smooth across time by introducing the kernel values $k(t, t')$ to link all $y_{t'}$ to all b_t, where k is a non-negative, non-decreasing function of $t - t'$,

$$k(t, t')E[y_{t'}|b_t] = k(t, t')Xb_t$$

Summing to account for all contributions implies that

$$\sum_{t' \in T} k(t, t')E[y_{t'}|b_t] = \sum_{t' \in T} k(t, t')Xb_t$$

and hence

$$\frac{1}{Z(t)} E[Y|b_t]k_t = Xb_t$$

where $Y \in [0, 1]^{M \times T} = \left[y_1, ..., y_T\right]$, $k_t = \left[k\left(t, t_1\right), ..., k\left(t, t_T\right)\right]^{\mathrm{T}}$ and $Z(t) = \sum_{t' \in T} k(t, t')$.

We use the box kernel $k_{box}(t, t', \kappa) = 1(|t - t'| \leq \kappa/2)$ and the Gaussian kernel $k_{Gaussian}(t, t', \kappa) = \exp\left\{\frac{-(t - t')^2}{2\kappa}\right\}$; other choices are also possible.

Solving the deconvolution problem

With observed mutation frequencies Y and known variant definitions X as input, we solve the deconvolution problem for a kernel $k(t, t')$ and loss function $L(z)$ by finding $\widehat{b_t}$ as

$$\widehat{b_t} = argmin_{b_t : b_{t,i} \geq 0 \, \forall i, \, ||b_t||_1 = 1} L(\frac{1}{Z(t)} Yk_t - Xb_t)$$

To solve this optimization problem, we use routines from the Python scientific computing library *Scipy* [14]. When using the LS loss function, this involves the non-negative least squares solver [15]. When using the SL1 loss, we use the Trust Region Reflective method [16].

Confidence intervals

To use WBE for robust decision making, it is essential to provide estimates of the uncertainty in the prediction of relative abundances. We pursue two different strategies for computing confidence intervals: one based on an analytical approximation to the standard errors, and one simulation-based.

Asymptotic confidence intervals

We assume that at a given time t, the proportion $y_{t,m}$ of reads supporting a certain mutation m follows a $Binomial/n$ distribution with parameter π_m, i.e.,

$$p(y_m|\pi_m) = \pi_m^{y_m}(1 - \pi_m)^{1-y_m}$$

Using here the assumption that the contributions of each variant to the expected proportion of mutated reads is additive, we have that $\pi_m = [Xb]_m$. We additionally assume conditional independence of the mutation proportions such that $p(y|\pi) = \prod_{m\in M} p(y_m|\pi_m)$. Differentiating twice the log-likelihood, we find the Fisher information matrix (see Supplementary 1)

$$I_b(b) = E\left[-\frac{\partial^2}{\partial b^2}\ell(\pi)\right] = E\left[\left(\frac{\partial}{\partial b}\pi\right)^{\top}\left(-\frac{\partial^2}{\partial \pi^2}\ell(\pi)\right)^{\top}\frac{\partial}{\partial b}\pi\right] = X^{\top} diag\left\{\pi_m(1 - \pi)_m\right\}_{m=1,\dots,M}^{-1} X$$

from which we can extract asymptotic standard errors, which can be used to construct Wald confidence intervals:

$$se^2\left(\hat{b}_v\right) \approx \left[I\left(\hat{b}\right)^{-1}\right]_{vv}$$

Here, a pseudofraction must be added to the entries of b to avoid division by zero when computing the asymptotic standard errors.

Logit reparametrization

To ensure that the confidence bands stay confined to the [0,1] interval, we can also compute the asymptotic standard errors and Wald confidence intervals on the logit scale $\phi_v = logit\left(b_v\right) = log\left(\frac{b_v}{1-b_v}\right)$, before projecting them back to the linear scale. We compute the inverse of the Fisher information matrix of ϕ by using the Delta method:

$$I_\phi(\phi)^{-1} = D_\phi(b) I_b(b)^{-1} D_\phi(b)^{\top}$$

where $D_\phi(b)$ is the Jacobian of the transformation, such that

$$D_\phi(b)_{ij} = \frac{\partial \phi(b)_i}{\partial b_j} = (-1)^{1(i\neq j)}\left(b_i\left(1 - b_i\right)\right)^{-1}$$

Here again, a pseudofraction must be added to the entries of b to avoid division by zero.

Overdispersion

In both Wald type confidence intervals, overdispersion (and underdispersion) is accounted for by following a quasilikelihood approach [17]. In this approach, we are computing the deviations of the data from the model fitted values to estimate to which extent the variability in the data is greater (or smaller) than what is expected from the fitted model. At a given time *t*, the asymptotic standard errors are adjusted for either b_t or ϕ_t:

$$se_{adj.}^2\left(\hat{b}_{t,v}\right) = \sigma_{t,v}^2\, se^2\left(\hat{b}_{t,v}\right),\ se_{adj.}^2\left(\hat{\phi}_{t,v}\right) = \sigma_{t,v}^2\, se^2\left(\hat{\phi}_{t,v}\right)$$

and the dispersion factor $\sigma_{t,v}^2$ is computed as:

$$\sigma_{t,v}^2 = \frac{1}{\sum_{t'\in T}\kappa(t,t')}\sum_{t'\in T}\frac{\kappa(t,t')}{\sum_{m\in M} X_{m,v}}\sum_{m\in M}\frac{\left(y_{t',m} - \hat{y}_{t,m}\right)^2}{\hat{y}_{t,m}\left(1-\hat{y}_{t,m}\right)} X_{m,v}$$

Bootstrapping

Another strategy to compute confidence bands is to use the non-parametric bootstrap approach [18]. Here, we construct B bootstrap samples of the whole time series, by resampling M mutation indices from $m \in 1, \ldots, M$ with replacement. Each bootstrap sample is then processed by deconvolution and smoothing, resulting in B time series of dimensionality V. For each relative variant abundance $v \in 1, \ldots, V$, confidence intervals are constructed at each timepoint t from the empirical quantiles of the bootstrap samples.

Implementation and availability

The methods we present here are implemented in the Python package *LolliPop,* which takes as input a tabular file of observed mutation frequencies and variant definitions, performs simultaneous kernel smoothing and deconvolution using numerical optimization, and produces confidence intervals. *LolliPop* is available on Github[3] and as a Bioconda package.

Wastewater sequencing data

We used the wastewater sequencing data from the Swiss surveillance project reported in [6]. The dataset contains 1295 NGS datasets from longitudinal samples collected at six major WWTPs, sampled daily between January 2021 and September 2021. We defined the variants of concern (VOCs) B.1.1.7 (Alpha), B.1.351 (Beta), P.1 (Gamma), B.1.617.1 (Kappa), and B.1.617.2 (Delta) by querying the mutations present in ≥80% of the sequences defining these variants on Cov-Spectrum [19] and supported by at least 100 sequences. We then counted these mutations from pileups of the read alignments. We deconvoluted using both LS and SL1 losses, using the Gaussian kernel function. We computed Wald confidence intervals with and without logit reparametrization, as well as bootstrap-based confidence intervals (1000 resamples).

Comparison to clinical data

Using the LAPIS API of Cov-Spectrum [19], we retrieved counts of sequenced SARS-CoV-2 PCR-positive clinical samples for Switzerland, stratified by submitting lab, canton, and inferred variant. We restricted the data to samples from the large clinical testing company Viollier, where the PCR-positive samples are randomly subsampled before being sent for sequencing. We compare each WWTP to the clinical data from the canton it is located in. For the Berne WWTP of Laupen, we compare to an aggregate of the clinical sequences from both the cantons of Bern and Fribourg, as the catchment area is split between those two cantons [6].

Hyperparameters

To assess the sensitivity of our deconvolution method to hyperparameter choice, we analyzed the data from the two biggest WWTPs (Zurich and Vaud) for an array of smoothing and robustness hyperparameters. For the deconvolution based on the MSE loss, we explored values of the smoothing bandwidth parameter $\kappa \in \{0, 1, \ldots, 60\}$ (in days). For the deconvolution based on the SL1 loss, we tested pairs of values of the smoothing bandwidth parameter and of the f_scale parameter (controlling the breakpoint between l_2 and l_1 loss) $(\kappa, \alpha) \in \{0, 5, 10, \ldots, 30\} \times \{0.01, 0.05, 0.09, 0.13, \ldots, 1\}$.

For each deconvolved dataset, we linearly regressed the relative abundances of the different variants in clinical sequencing on the relative abundances in wastewater inferred by the deconvolution, using the statistical software R [20], and we reported the R^2 value and the average bias. The regressions and average bias were computed with data points weighted by the square root of the clinical sample sizes.

Results

We have developed *LolliPop,* a statistical method to solve the variant deconvolution problem. Our method uses as input the mutation frequencies observed in wastewater samples, which are weighted

[3] https://github.com/cbg-ethz/lollipop

according to a kernel function and deconvolved according to a variant definition matrix and a loss function (Figure 1 a). The results are vectors of relative variant abundance estimates and their confidence intervals. *LolliPop* is implemented as a Python package available on Bioconda. In the results section, we present a comparison of deconvolved relative abundance of variants from the Swiss wastewater surveillance project to clinical data from populations connected to the respective treatment plants. We assess the effect of hyperparameters, and present confidence bands computed using the different methods introduced.

Hyperparameters

We ran *LolliPop* using a grid of different hyperparameter values. For both loss functions, increasing the smoothing bandwidth κ increased goodness of fit with clinical values (Supplementary Figures S1 a, S2 a). However, increasing the bandwidth generally resulted in a slight increase of the bias (Supplementary Figures S1 b, S2 b). For the deconvolution based on the SL1 loss, the R^2 was highest for scale parameter $\alpha = 0.13$, independently of the value of κ used (Supplementary Figure S2 a).

Comparison to clinical data

We compared time series of relative variant abundances inferred from wastewater sequencing data using *LolliPop* to those estimated using clinical data. We found that the wastewater-based infection dynamics closely follow those derived from clinical sequencing (Figure 1 b). The Pearson *r* (weighted by root clinical sample size) between the deconvolved and clinical estimates was 0.923 and 0.920 for the LS-based and SL1-based deconvolutions, respectively (Figure 1 d,e). The highest cross-correlation value between wastewater-derived estimates and clinical estimates was obtained for a time lag of $\Delta t = 4.0$ days and $\Delta t = 3.3$ days for the LS and SL1 based deconvolutions, respectively (Figure 1 c).

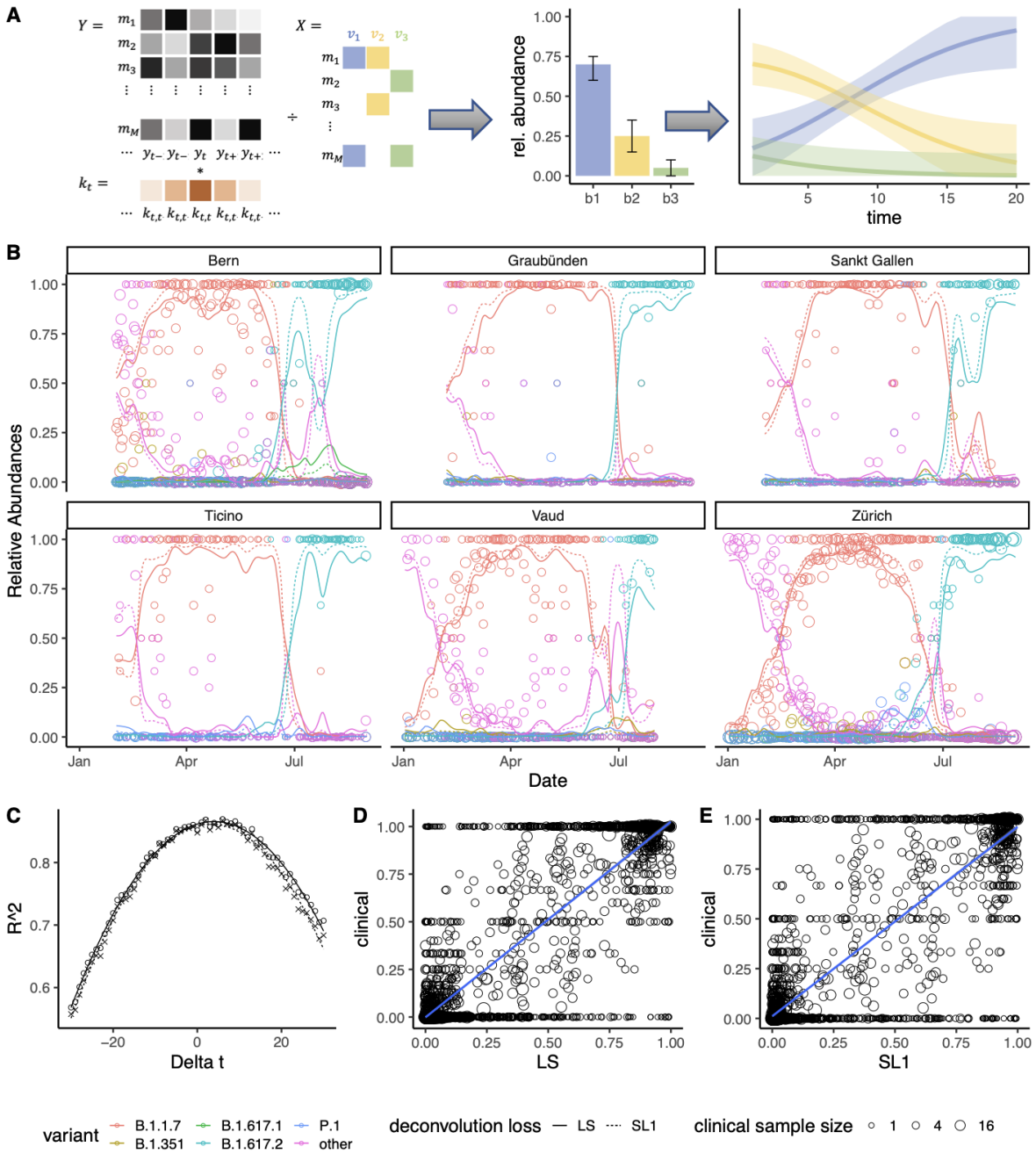


Figure 1: Overview of the method: the vectors ..., y_{t-1}, y_t, y_{t+1}, ... of mutation frequencies in wastewater samples are weighted according to kernel k and deconvolved using the variant definition matrix X, producing estimates of relative variant abundances along with confidence intervals. Repeating the operation for each timepoint tracks the relative abundances b_t of the variants through time (A). Estimates of variant relative abundances obtained from deconvolution of wastewater NGS data (solid and dotted lines), compared to relative abundances of variants in clinical samples of the cantons surrounding the WWTPs (circles) (B). Different colors correspond to the different genomic variants studied. Weighted cross correlation between the deconvolved values of relative abundances and the clinical data estimates peaks at $\Delta t = 4.0$ days (when using LS) and $\Delta t = 3.3$ days (when using SL1), and decreases quadratically (C). Pearson *r* between the deconvolved values and the clinical data estimates (weighted by $\sqrt{n_{clinical}}$) is 92.2% for the LS based estimates (D) and 92.0% for the SL1 based estimates (E) at lag $\Delta t = 0$. LS and SL1 deconvolution was performed with a Gaussian smoothing kernel with $\kappa = 30$. For the SL1 loss, $\alpha = 0.135$.

Confidence intervals

Confidence intervals were different depending on whether they were computed using the Wald method on the linear scale or logit scale or using the bootstrap. Wald confidence intervals computed on the linear scale could, as expected, leave the [0,1] range and thus need to be clipped in a post-processing step (Figure 2 a). They also were in general substantially narrower compared to the other two approaches. Wald confidence intervals computed on the logit scale and then back transformed (Figure 2 b) more closely resembled the confidence intervals computed using bootstrapping (Figure 2 c). Especially in the Wald confidence intervals, uncertainty was consistently higher during the months in which wastewater samples contained low concentrations of SARS-CoV-2 RNA due to low incidence of the virus[4].

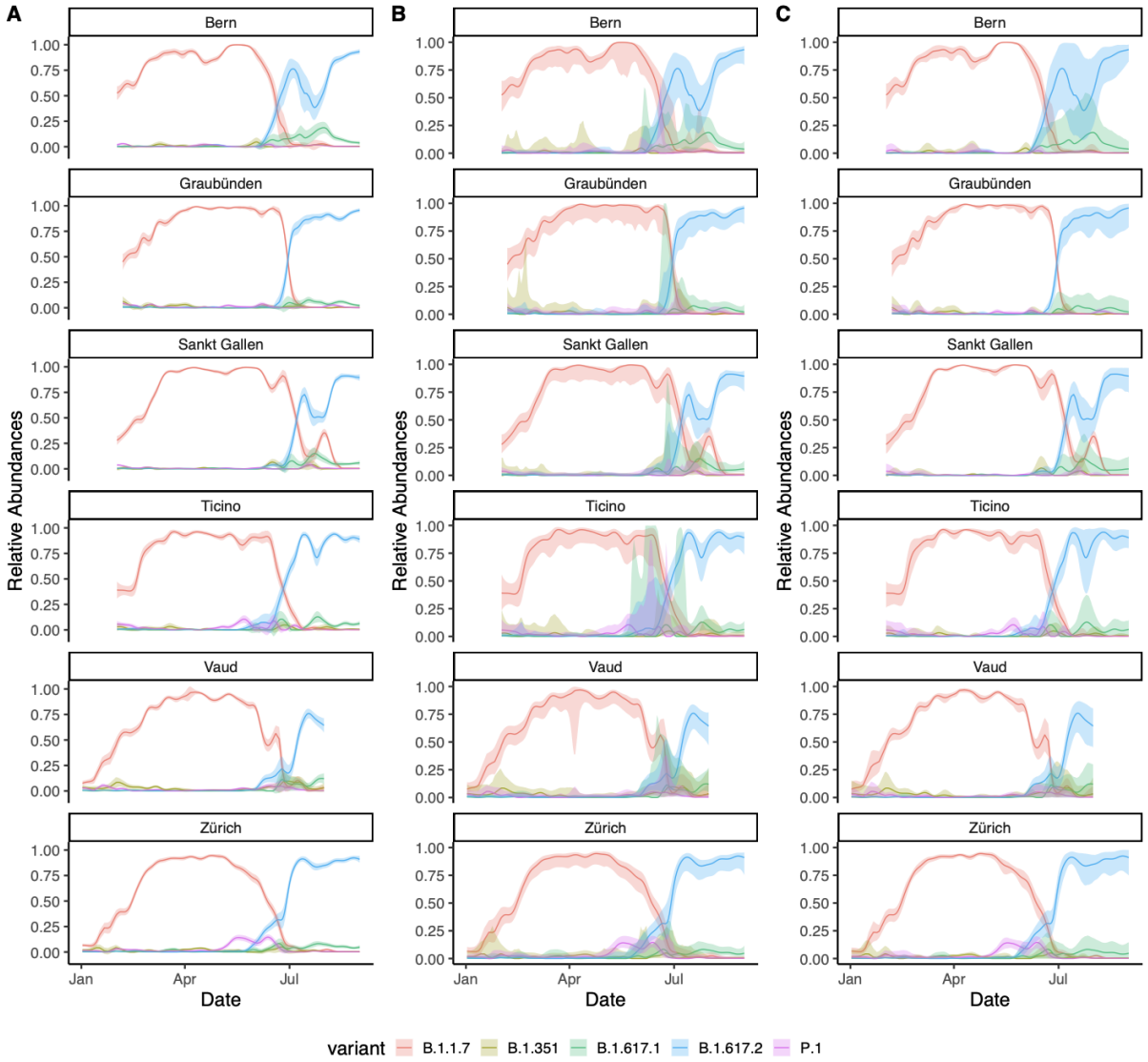


Figure 2: Confidence bands of the deconvolved relative abundance values, at 95% level. Wald confidence bands were computed using the asymptotic standard error on the linear scale (A), or on the logit scale and then back-transformed to the linear scale (B). Bootstrap confidence intervals (C) were computed using 1000 resamples.

[4] https://sensors-eawag.ch/sars/overview.html

Runtime

Wall time on a MacBook Air M1 2020 was ~2s to deconvolve all 1295 samples using the LS loss function. Using the SL1 loss function, wall time was ~1min for the same task. Wall time was ~1min to construct Wald confidence intervals, and ~1min for reparameterized Wald confidence intervals. Constructing confidence intervals by bootstrapping (1000 resamples) had a wall time of ~30min. All computations were performed using a single CPU core.

Discussion

We introduced *LolliPop*, a method for solving the variant deconvolution problem. We showed its application to data from the Swiss wastewater monitoring program, extending over eight months at six different locations around the country. We found that the deconvolved values of relative abundance closely follow the dynamics seen in the clinical sequencing effort.

Regarding runtime, we observed that *LolliPop* can be easily used on a standard laptop computer to process thousands of samples in a matter of seconds or minutes, using a single computing core. This emphasizes how the program can readily be used in analyses that scale up to national monitoring programs. More substantial speedups could easily be obtained by using multiple cores, as the task is naturally parallelizable.

Modifications to *LolliPop* can be easily implemented. First, the behavior of the deconvolution under different types of loss functions could be investigated. We have assessed here Least Squares and Soft l_1 losses, and the results were not very strongly affected by the choice of loss function, indicating robustness of the method to this choice. Other types of losses could certainly have interesting properties for the problem at hand. For example, if the number of candidate lineages to deconvolve grows, building in a sparsity assumption by adding an l_1 regularization term of the relative variant abundances could lower the variance of their estimates.

Another component that can readily be modified is the type of kernel used for smoothing. We have used a Gaussian kernel, but other choices could be relevant depending on the application. For example ad-hoc kernels could be used for more precisely relating the relative abundances in samples to the relative incidences in the population. The temporal dynamics of viral shedding in wastewater are generally described by a shedding load distribution [21], which could be accounted for by an asymmetric and non zero-centered kernel.

To assess the uncertainty in the estimates of relative abundances, we have derived two analytical and one bootstrap-based approaches to produce confidence intervals. Bootstrap confidence intervals are by design computationally intensive, and having an analytical approach can provide substantial speedup. The Wald confidence intervals we derived here include three components of the variance of the estimates: one to account for the mutation overlap between variant definitions, one to account for the binomial nature of the sampling, and one to account for overdispersion. The Wald confidence interval computed on the linear scale suffers from known shortcomings, as it can exit the [0,1] range. They are also substantially narrower than the others and hence we would advise against using these. In our results, the Wald confidence intervals computed on the logit scale very closely resemble the bootstrap confidence intervals, while providing almost ~30x speedup in the computation time. However, they could be subject to numerical stability issues in case of an ill-conditioned Fisher information matrix, which could happen in instances where low coverage leads to high levels of missing values.

We have here built in the assumption of temporal continuity of the variant relative abundances in our model, in the form of kernel smoothing simultaneous to the deconvolution. As in a monitoring program multiple locations are generally being monitored, another useful assumption to build in the deconvolution could be that of spatial continuity if the spatial resolution allows for it. Jointly smoothing the different locations might offer increased robustness to the estimates by partial pooling of the information.

To summarize, *LolliPop* solves the variant deconvolution problem, taking into account the time series nature of wastewater sequencing datasets and mitigates the high levels of noise these experiments

typically display. Our method can estimate uncertainty using different approaches, including analytical confidence bands with short computation times. *LolliPop* enables genomic variant tracking in large-scale wastewater-based epidemiology projects.

References

1. Knyazev, S., Chhugani, K., Sarwal, V., Ayyala, R., Singh, H., Karthikeyan, S., Deshpande, D., Baykal, P.I., Comarova, Z., Lu, A., *et al.* (2022). Unlocking capacities of genomics for the COVID-19 response and future pandemics. Nat. Methods *19*, 374–380.

2. Wastewater monitoring comes of age. (2022). Nat. Microbiol. *7*, 1101–1102.

3. Hadfield, J., Megill, C., Bell, S.M., Huddleston, J., Potter, B., Callender, C., Sagulenko, P., Bedford, T., and Neher, R.A. (2018). Nextstrain: real-time tracking of pathogen evolution. Bioinformatics *34*, 4121–4123.

4. O'Toole, Á., Scher, E., Underwood, A., Jackson, B., Hill, V., McCrone, J.T., Colquhoun, R., Ruis, C., Abu-Dahab, K., Taylor, B., *et al.* (2021). Assignment of epidemiological lineages in an emerging pandemic using the pangolin tool. Virus Evol. *7*, veab064.

5. Turakhia, Y., Thornlow, B., Hinrichs, A.S., De Maio, N., Gozashti, L., Lanfear, R., Haussler, D., and Corbett-Detig, R. (2021). Ultrafast Sample placement on Existing tRees (UShER) enables real-time phylogenetics for the SARS-CoV-2 pandemic. Nat. Genet. *53*, 809–816.

6. Jahn, K., Dreifuss, D., Topolsky, I., Kull, A., Ganesanandamoorthy, P., Fernandez-Cassi, X., Bänziger, C., Devaux, A.J., Stachler, E., Caduff, L., *et al.* (2022). Early detection and surveillance of SARS-CoV-2 genomic variants in wastewater using COJAC. Nat. Microbiol. *7*, 1151–1160.

7. Chen, C., Nadeau, S.A., Topolsky, I., Manceau, M., Huisman, J.S., Jablonski, K.P., Fuhrmann, L., Dreifuss, D., Jahn, K., Beckmann, C., *et al.* (2021). Quantification of the spread of SARS-CoV-2 variant B.1.1.7 in Switzerland. Epidemics *37*, 100480.

8. Caduff, L., Dreifuss, D., Schindler, T., Devaux, A.J., Ganesanandamoorthy, P., Kull, A., Stachler, E., Fernandez-Cassi, X., Beerenwinkel, N., Kohn, T., *et al.* (2021). Inferring Transmission Fitness Advantage of SARS-CoV-2 Variants of Concern in Wastewater Using Digital PCR. medRxiv.

9. Caduff, L., Dreifuss, D., Schindler, T., Devaux, A.J., Ganesanandamoorthy, P., Kull, A., Stachler, E., Fernandez-Cassi, X., Beerenwinkel, N., Kohn, T., *et al.* (2022). Inferring transmission fitness advantage of SARS-CoV-2 variants of concern from wastewater samples using digital PCR, Switzerland, December 2020 through March 2021. Euro Surveill. *27*.

10. Baaijens, J.A., Zulli, A., Ott, I.M., Petrone, M.E., Alpert, T., Fauver, J.R., Kalinich, C.C., Vogels, C.B.F., Breban, M.I., Duvallet, C., *et al.* (2021). Variant abundance estimation for SARS-CoV-2 in wastewater using RNA-Seq quantification. medRxiv.

11. Valieris, R., Drummond, R.D., Defelicibus, A., Dias-Neto, E., Rosales, R.A., and Tojal da Silva, I. (2022). A mixture model for determining SARS-Cov-2 variant composition in pooled samples. Bioinformatics.

12. Karthikeyan, S., Levy, J.I., De Hoff, P., Humphrey, G., Birmingham, A., Jepsen, K., Farmer, S., Tubb, H.M., Valles, T., Tribelhorn, C.E., *et al.* (2021). Wastewater sequencing uncovers early, cryptic SARS-CoV-2 variant transmission. medRxiv.

13. Amman, F., Markt, R., Endler, L., Hupfauf, S., Agerer, B., Schedl, A., Richter, L., Zechmeister, M., Bicher, M., Heiler, G., *et al.* (2022). National-scale surveillance of emerging SARS-CoV-2 variants in wastewater. medRxiv.

14. Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., *et al.* (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat. Methods *17*, 261–272.

15. Ling, R.F., Lawson, C.L., and Hanson, R.J. (1977). Solving least squares problems. J. Am. Stat. Assoc. *72*, 930.

16. Branch, M.A., Coleman, T.F., and Li, Y. (1999). A Subspace, Interior, and Conjugate Gradient Method for Large-Scale Bound-Constrained Minimization Problems. SIAM J. Sci. Comput. *21*, 1–23.

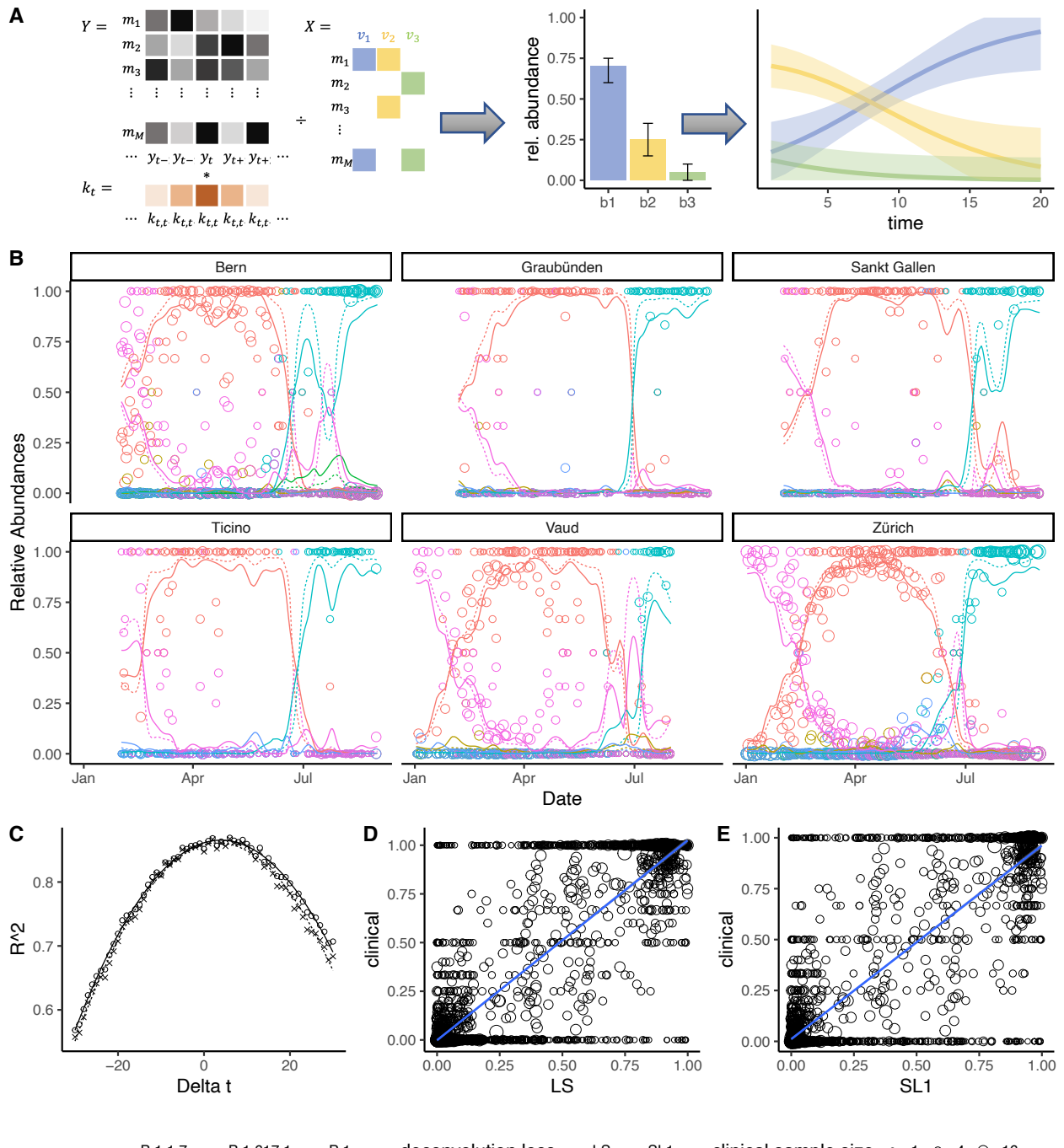
17. McCullagh, P., and Nelder, J.A. (2019). Generalized Linear Models (Routledge).

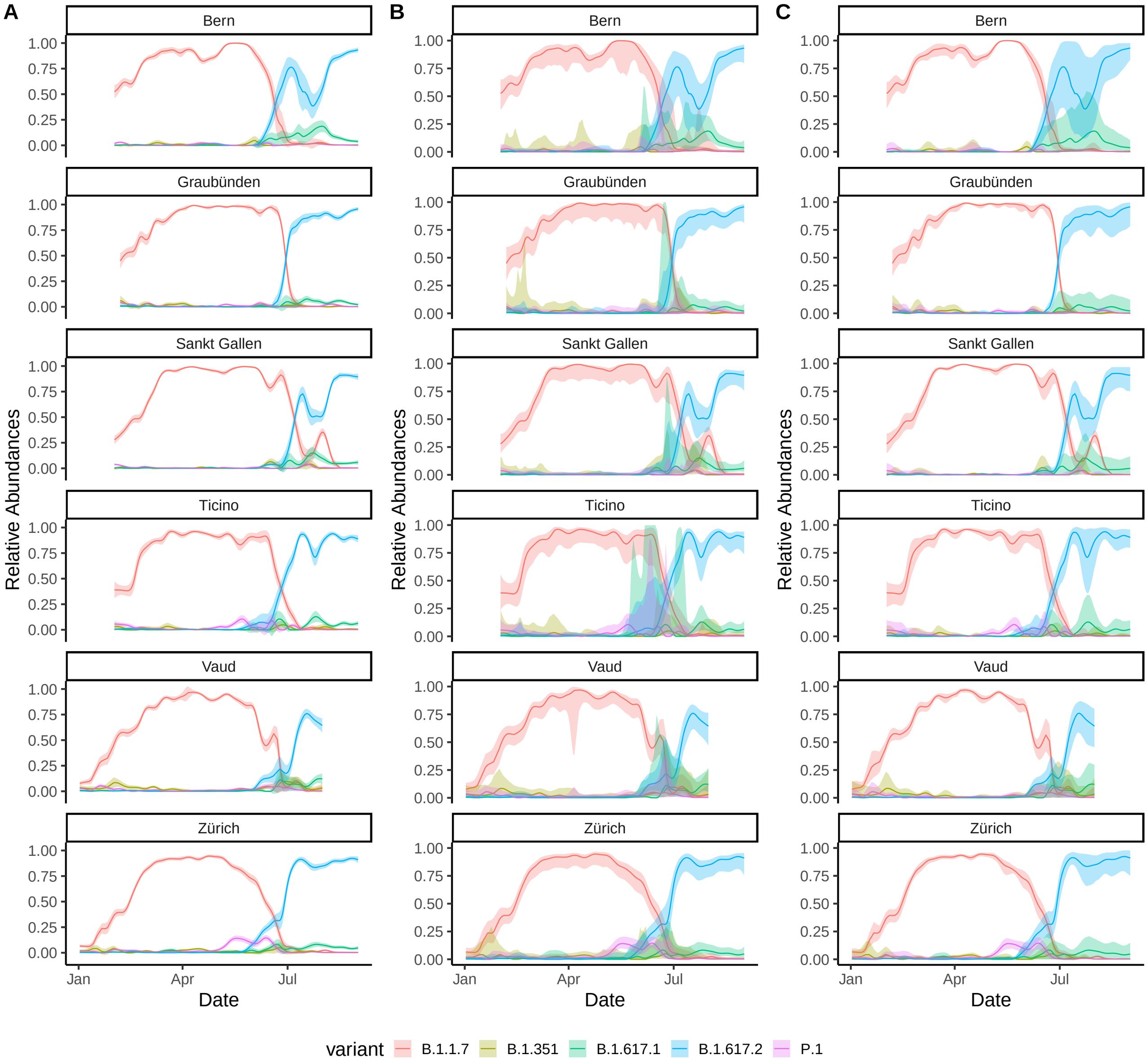
18. Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. Ann. Statist. *7*, 1–26.

19. Chen, C., Nadeau, S., Yared, M., Voinov, P., Xie, N., Roemer, C., and Stadler, T. (2021). CoV-Spectrum: Analysis of Globally Shared SARS-CoV-2 Data to Identify and Characterize New Variants. Bioinformatics.

20. Team, R.C. (2019). R: A Language and Environment for Statistical Computing.

21. Huisman, J.S., Scire, J., Caduff, L., Fernandez-Cassi, X., Ganesanandamoorthy, P., Kull, A., Scheidegger, A., Stachler, E., Boehm, A.B., Hughes, B., *et al.* (2021). Wastewater-based estimation of the effective reproductive number of SARS-CoV-2. medRxiv.


A
Y =
X =
÷
*
rel. abundance
time
b1
b2
b3
B
Bern
Graubünden
Sankt Gallen
Ticino
Vaud
Zürich
Relative Abundances
Date
Jan
Apr
Jul
C
R^2
Delta t
D
clinical
LS
E
clinical
SL1
variant
B.1.1.7
B.1.351
B.1.617.1
B.1.617.2
P.1
other
deconvolution loss
LS
SL1
clinical sample size
1
4
16


A
B
C
Bern
Graubünden
Sankt Gallen
Ticino
Vaud
Zürich
Relative Abundances
1.00
0.75
0.50
0.25
0.00
Jan
Apr
Jul
Date
variant
B.1.1.7
B.1.351
B.1.617.1
B.1.617.2
P.1