SARS-CoV-2 VARIANT PREVALENCE ESTIMATION USING WASTEWATER SAMPLES

I. López-de-Ullibarri[a], L. Tomás[b,c], N. Trigo-Tasende[d], B. Freire[e], M. Vaamonde[a], P. Gallego-García[b,c], I. Barbeito[a], J.A. Vallejo[d], J. Tarrío-Saavedra[a], P. Alvariño[b,c], E. Beade[e], N. Estévez[b,c], S. Rumbo-Feal[d], K. Conde-Pérez[d], L. de Chiara[b,c], I. Iglesias-Corrás[e], M. Poza[d], S. Ladra[e], D. Posada[b,c,f], R. Cao[a]

[a]Research Group MODES, Research Center for Information and Communication Technologies (CITIC), University of A Coruña (UDC), Campus de Elviña, 15071 A Coruña, Spain.

[b]CINBIO, Universidade de Vigo, 36310 Vigo, Spain.

[c]Galicia Sur Health Research Institute (IIS Galicia Sur), SERGAS-UVIGO, 36312 Vigo, Spain.

[d]Microbiology Research Group: meiGAbiome-Biomedical Research Institute (INIBIC)-Center for Advanced Research (CICA)-University of A Coruña (UDC)-CIBER of Infectious Diseases (CIBERINF), Servicio de Microbiología, 3ª planta, Edificio Sur, Hospital Universitario, As Xubias, 15006 A Coruña, Spain.

[e]University of A Coruña (UDC), Research Center for Information and Communication Technologies (CITIC), Database Laboratory, Campus de Elviña, 15071 A Coruña, Spain.

[f]Department of Biochemistry, Genetics, and Immunology, Universidade de Vigo, 36310 Vigo, Spain.

Abstract

The present work describes a statistical model to account for sequencing information of SARS-CoV-2 variants in wastewater samples. The model expresses the joint probability distribution of the number of genomic reads corresponding to mutations and non-mutations in every locus in terms of the variant proportions and the joint mutation distribution within every variant. Since the variant joint mutation distribution can be estimated using GISAID data, the only unknown parameters in the model are the variant proportions. These are estimated using maximum likelihood. The method is applied to monitor the evolution of variant proportions using genomic data coming from wastewater samples collected in A Coruña (NW Spain) in the period May 2021 – March 2022. Although the procedure is applied assuming independence among the number of

reads along the genome, it is also extended to account for Markovian dependence of counts along loci in the aggregated information coming from wastewater samples.

Motivation and background

During the last decade, wastewater-based epidemiological surveillance has emerged as a highly relevant discipline, with the potential to provide information by combining the use of analytical methods with the development of ad hoc modelling approaches. This surveillance has been widely used in recent years to accurately predict consumption patterns for numerous substances (EMCDDA, 2020). During the COVID-19 pandemic, processes for monitoring the viral load of SARS-CoV-2 in wastewater were developed for the first time in the Netherlands (Medema et al., 2020).

Around a third of the people primarily infected with SARS-CoV-2 in Spain were asymptomatic (Pollán et al., 2020). However, the percentage of asymptomatic cases depends on many factors, such as the average age and the degree of natural or artificial immunity in each population. In addition, a significant proportion of people infected with COVID-19, including symptomatic and asymptomatic, who were tested for fecal viral RNA tested positive from the initial steps of infection (Gupta et al., 2020) and tested positive persistently in rectal swabs even after nasopharyngeal testing was negative (Chen et al., 2020; Xing et al., 2020; Xu et al., 2020; Zhang et al., 2020; Cevik et al., 2021; Miura et al., 2021).

Due to all of the above, the genetic material of SARS-CoV-2 can be found in wastewater (Lodder and de Roda Husman, 2020), which has made the monitoring of the RNA viral load in wastewater an excellent tool for the epidemiological monitoring of the COVID-19 pandemic, as well as an efficient early warning method for the detection of outbreaks (Randazzo et al., 2020; Ahmed et al., 2020; Medema et al., 2020; Peccia et al., 2020; F Wu et al., 2020; Wurtzer et al., 2020). Likewise, the methods of massive sequencing of aggregate samples collected in wastewater treatment plants or in the sanitation network itself make it possible to obtain readings that include the mutations observed in the SARS-CoV-2 genome. With the help of appropriate statistical models and methods, estimates of the number of active cases of patients with COVID-19 can be obtained from the viral load quantification data at Wastewater Treatment Plants (WWTPs) (Vallejo et al. 2022).

On the other hand, as a result of the proliferation of SARS-CoV-2 variants, specific mutations have been monitored to study the evolution of variants (Bar-Or et al. 2021) and the total SARS-CoV-2 concentration (Radu et al. 2022). Recently, statistical methods have been proposed that make it possible to analyze the readings of mutation frequencies in the virus genome in order to obtain precise estimates of the proportions of variants (Barbeito et al. 2022, Gafurov et al. 2022, Karthikeyan et al. 2022, Radu et al. 2022, Valieris et al. 2022). In this paper, the joint mutation distribution is estimated using GISAID data and the variant proportions are estimated using maximum likelihood. The model can be formulated either assuming independence among the number of reads along the genome or allowing for Markovian dependence of counts along loci.

Methodology

Since the genetic material of the samples collected at the WWTP is degraded as a consequence of the passage of wastewater through the sanitation network, the genomes collected are remarkably fragmented. On the other hand, each sample corresponds to the genetic material of the thousands of infected human beings among the almost 400,000 inhabitants of the metropolitan area of A Coruña. As a consequence of all this and of the amplicon technology used for massive sequencing (see Section 4), the available information corresponds to counts of mutation reads throughout a number of positions (loci) in the virus genome.

In the case in which clinical samples could be taken from individual patients, it would be possible to observe the complete RNA strand (or at least very large fragments of it that could be juxtaposed), which means having observations of the vector variable that considers which type of mutation has occurred at each locus. However, for the samples obtained at the WWTP, it is only possible to observe the frequencies of mutations in each of these loci in an aggregated manner on the set of individuals that have excreted that genetic material. As a consequence, the statistical methods for estimating the proportions of variants have to be designed for the data-generating process, aggregated, in individuals, and marginal, in loci, that occurs in this setup. We will now formulate this data-generating process.

A viral haplotype can be expressed as a vector $x = (x_1, \dots, x_l)$, l being the number of genomic positions or loci. The set of feasible values for locus x_i is $A_i = \{0, \dots, a_i\}$, where 0 refers to the reference allele and $1, \dots, a_i$ are indices identifying the alternative alleles (i.e. different types of mutations at locus i=1,…,l). As a consequence, $x \in H$, H being the Cartesian product $A_1 \times \cdots \times A_l$. We denote by X and V, respectively, discrete random variables modeling a haplotype and a viral variant sampled at random from the viral genomes in wastewater. For r viral variants $v_1, \dots, v_r$, the quantities $p_x^{(j)}$, for $j = 1, \dots, r$, are defined as $P(X = x \mid V = v_j)$. So $p_x^{(j)}$, when $x \in H$, is just the haplotype distribution of variant v_j. By the total probability law, $p_x = P(X = x) = \sum_{j=1}^{r} \pi_j p_x^{(j)}$, where $\pi_j = P(V = v_j)$ is the unknown probability of the j-th variant. It is important to remark that, although the $p_x^{(j)}$ are also unknown, they can be estimated very easily without using the wastewater samples, e.g., from the viral genomes available at GISAID's EpiCoV database.

If the viral genomic sequences could be fully observed in wastewater, the data would consist of a sample of n haplotype vectors $X_1, \dots, X_n$. Given this "ideal sample" (not observable in wastewater, just for clinical patients), the observed sample can be modeled as follows. Consider, for each locus k, for $k = 1, \dots, l$, the probability α_k that the k-th locus of a viral genome selected at random is observed in the sample. The number of observations for locus k is $N_k = \sum_{i=1}^{n} I_k^{(i)}$, where $I_k^{(i)}$ is a binary random variable indicating whether the i-th "ideally observed" haplotype has been actually observed at locus k. It is natural to model N_k as a random variable with binomial distribution, $\mathrm{B}(n, \alpha_k)$, n being the expected number of reads at locus k. Its mean $n\,\alpha_k$ depends on the α_k probabilities, which are strongly determined by the sequencing technology and may greatly differ across loci. Since the N_k are observable, in the following we condition on their observed values.

Given N_k, for $k = 1, \dots, l$, and assuming that the sequencing technology does not affect the marginal distribution of X, it is possible to derive the distribution of the observed allele frequencies for each locus in the sample, $Y = (Y_1, \dots, Y_l)$, where, for $k = 1, \dots, l$, $Y_k = \left(Y_{k,0}, \dots, Y_{k,a_k}\right)$ and $Y_{k,s} = \sum_{i=1}^{n} I_k^{(i)}\, \mathbf{1}\left(X_i^{(k)} = s\right)$. In the last expression, to avoid ambiguity, the superscript (k) is used to refer to the k-th component of X_i, and $\mathbf{1}(A)$ is the indicator of event A. Clearly, $\sum_{s=1}^{l} Y_{k,s} = N_k$ and, conditionally on N_k, Y_k has

multinomial distribution $\mathrm{M}(N_k, q_k)$, where $q_k = \left(q_{k,0}, \dots, q_{k,a_k}\right)$ is a vector whose s-th component is $q_{k,s} = P\left(X^{(k)} = s\right) = \sum_{x \in H} \mathbf{1}(x_k = s) \sum_{j=1}^{r} \pi_j p_x^{(j)}$.

Thus, the distribution of Y depends on the "known" haplotype probabilities within every viral variant estimated from available data ($p_x^{(j)}$, $j = 1, \dots, r$), the number of reads at every locus (N_k, $k = 1, \dots, l$), and the unknown variant probabilities (π_j, $j = 1, \dots, r$) in the population of viral genomes sampled. The π_j can be estimated using available information and the observed allele frequencies in the wastewater sample. Assuming independence of the random variables Y_k, $k = 1, \dots, l$, and having observed the allele mutation frequencies collected in the vector $y = (y_1, \dots, y_l)$, the likelihood (conditional on N_k, $k = 1, \dots, l$) is:

$$L(\pi_1, \dots, \pi_r) = P(Y = y) = \prod_{k=1}^{l} \frac{N_k}{y_{k,0}! \cdots y_{k,a_k}!} \prod_{s=0}^{a_k} \left(\sum_{x \in H} \mathbf{1}(x_k = s) \sum_{j=1}^{r} \pi_j p_x^{(j)} \right)^{y_{k,s}}.$$

Maximum conditional likelihood estimates of $(\pi_1, \dots, \pi_r)$ are obtained by maximizing $L(\pi_1, \dots, \pi_r)$ constrained to $\pi_1 \geq 0, \dots, \pi_r \geq 0, \sum_{i=1}^{r} \pi_i = 1$, e.g., using an augmented Lagrangian method.

Markovian dependence among loci

The independence assumption among the random variables Y_k, $k = 1, \dots, l$, can be relaxed by just assuming a Markovian condition for the random vector Y:

$$P(Y_k = y_k \mid Y_{k-1} = y_{k-1}, \dots, Y_1 = y_1) = P(Y_k = y_k \mid Y_{k-1} = y_{k-1}), k = 1, \dots, l.$$

By assuming this condition, the likelihood becomes:

$$L(\pi_1, \dots, \pi_r) = P(Y = y) = P(Y_1 = y_1) \prod_{k=2}^{l} P(Y_k = y_k \mid Y_{k-1} = y_{k-1}),$$

which just requires to deal with the conditional probabilities of the form $P(Y_k = y_k \mid Y_{k-1} = y_{k-1})$, for $k = 2, \dots, l$. Without loss of generality and for simplifying the notation, we consider $P(Y_2 = y_2 \mid Y_1 = y_1)$ and assume that $a_1 = a_2 = 1$, i.e. just one type of possible mutation at loci $k = 1, 2$. As a consequence, the joint distribution of

$(Y_1, Y_2) = (Y_{1,0}, Y_{1,1}, Y_{2,0}, Y_{2,1})$ can be expressed in terms of the random vector $Z = (Z_{0,0}, Z_{0,1}, Z_{1,0}, Z_{1,1})$, where the random variable $Z_{i,j}$ denotes the number of co-occurrences of mutation *i* in locus 1 and mutation *j* in locus 2. Indeed

$$Y_{1,0} = Z_{0,0} + Z_{0,1}, Y_{1,1} = Z_{1,0} + Z_{1,1}, Y_{2,0} = Z_{0,0} + Z_{1,0}, Y_{2,1} = Z_{0,1} + Z_{1,1}.$$

Now, since the random vector Z has a multinomial distribution: $\mathrm{M}\left(N_{1,2}, (p_{0,0}, p_{0,1}, p_{1,0}, p_{1,1})\right)$, where $N_{1,2}$ is the number of joint reads at loci 1 and 2 and $(p_{0,0}, p_{0,1}, p_{1,0}, p_{1,1})$ is the vector with the probability mass corresponding to mutations (0 or 1) at loci 1 and 2, the joint probability mass of (Y_1, Y_2) is then straightforward:

$$P(Y_{1,0} = y_{1,0}, Y_{1,1} = y_{1,1}, Y_{2,0} = y_{2,0}, Y_{2,1} = y_{2,1}) = \sum_{z \in C(y)} \frac{N_{1,2}!\, p_{0,0}^{z_{0,0}}\, p_{0,1}^{z_{0,1}}\, p_{1,0}^{z_{1,0}}\, p_{1,1}^{z_{1,1}}}{z_{0,0}!\, z_{0,1}!\, z_{1,0}!\, z_{1,1}!},$$

where $z \in C(y)$ in the sum means that the values of z ranges over all possibilities such that $y_{1,0} = z_{0,0} + z_{0,1}, y_{1,1} = z_{1,0} + z_{1,1}, y_{2,0} = z_{0,0} + z_{1,0}, y_{2,1} = z_{0,1} + y_{1,1}$. The marginal probability mass of Y_1 is even simpler:

$$P(Y_{1,0} = y_{1,0}, Y_{1,1} = y_{1,1}) = \frac{N_{1,2}!\, (p_{0,0} + p_{0,1})^{y_{1,0}} (p_{1,0} + p_{1,1})^{y_{2,0}}}{y_{1,0}!\, y_{2,0}!}.$$

Using the definition of conditional probability, the conditional distribution becomes:

$$P(Y_2 = y_2 \mid Y_1 = y_1) = \frac{\sum_{z \in C(y)} \frac{N_{1,2}!\, p_{0,0}^{z_{0,0}}\, p_{0,1}^{z_{0,1}}\, p_{1,0}^{z_{1,0}}\, p_{1,1}^{z_{1,1}}}{z_{0,0}!\, z_{0,1}!\, z_{1,0}!\, z_{1,1}!}}{\frac{N_{1,2}!\, (p_{0,0} + p_{0,1})^{y_{1,0}} (p_{1,0} + p_{1,1})^{y_{2,0}}}{y_{1,0}!\, y_{2,0}!}},$$

where the co-occurrence probabilities can be easily expressed in terms of the variants bivariate haplotype distributions, $p_{(i,j)}^{2,(m)}$, and the variant marginal distribution:

$$p_{i,j} = P(X_1 = i, X_2 = j) = \sum_{m=1}^{r} \pi_m \sum_{x \in H} \mathbf{1}(x_1 = i, x_2 = j) p_x^{(m)} = \sum_{m=1}^{r} \pi_m\, p_{(i,j)}^{2,(m)}.$$

As a consequence, the likelihood in the Markovian dependence case can be written just in terms of the variants bivariate haplotype distributions and the unknown variant probabilities.

Simulations

Simulated data, as well as synthetic data coming from in vitro experiments, where the proportion of every variant is known, have been used to assess the quality of the method. We considered four scenarios.

Dataset #1 consists of simulated reads of 1 genome per variant without sequencing errors. The data were created from four different genomes from GISAID (consensus sequences), each genome corresponding to a different variant. A simulator of amplicon reads (with no sequencing errors) is applied based on the real coverage/depth profiles of ARCTIC protocol (obtained from real reads) and then those simulated reads are mixed in the percentages included in Table 1, which also contains the estimated percentages.

Variant	B.1.1.7	B.1.617.2	B.1.621	C.37
True percentages	64%	21%	12%	3%
Estimated percentages	61.70%	22.35%	13.70%	2.25%

Table 1: Mixing variant percentages and their estimations for Dataset #1.

Dataset #2 also contains simulated reads without sequencing errors but of multiple genomes per variant. The data were created from four different genomes from GISAID (consensus sequences), each genome corresponding to a different variant. As for the previous dataset, a simulator of amplicon reads is applied based on the real coverage/depth profiles of ARCTIC protocol, obtained from real reads. The simulated reads are mixed in the percentages included in Table 2, which also contains the estimated percentages.

Variant	B.1.1.318	B.1.1.7	B.1.351	B.1.617.1	B.1.617.2	B.1.621	C.37
Real percentages	0.00%	53.00%	0.00%	0.00%	27.00%	7.00%	13.00%
Estimated percentages	0.95%	52.17%	0.91%	1.17%	24.88%	7.05%	12.86%

Table 2: Mixing variant percentages and their estimations for Dataset #2.

Dataset #3 consists of mixing clinical samples created from real genomes reads obtained in the project EPICOVIGAL. For each variant, just one dataset is used and then the reads were mixed according to the percentages presented in Table 3. This table also includes the estimated percentages.

Variant	A.28	B.1.1.7	B.1.525	B.1.617.2	B.1.620	B.1.621	C.37	P.3
Real percentages	0.00%	45.00%	0.00%	35.00%	0.00%	15.00%	5.00%	0.00%
Estimated percentages	0.04%	39.68%	0.10%	38.55%	1.54%	14.57%	5.47%	0.05%

Table 3: Mixing variant percentages and their estimations for Dataset #3.

Dataset #4 was also constructed by mixing clinical samples. It was created from real genomes reads obtained in the project EPICOVIGAL mixed in the percentages collected in Table 4, which also includes the estimated percentages.

Variant	B.1.1.7	B.1.351	B.1.617.2	P.1
Real percentages	40.00%	10.00%	30.00%	20.00%
Estimated percentages	39.41%	9.47%	32.63%	18.49%

Table 4: Mixing variant percentages and their estimations for Dataset #4.

The results in Tables 1-4 show that the estimation error of the variant percentages is always below 2.7% for all the variants in Datasets #1, #2 and #4. For Dataset #3, the largest estimation error is around 5.3%. This happens for B.1.1.7, with a real percentage of 45%. This implies a relative estimation error of around 1/9.

Monitoring the evolution of variant proportions

The method presented is applied to monitoring the evolution of variant proportions using genomic data coming from weekly wastewater samples collected in A Coruña (NW Spain) in the period May 2021 – March 2022. This monitoring was part of the COVIDBENS project. It was an initiative carried out from April 2020 to March 2022 and financed by the public company WWTP Bens S.A., responsible for managing the WWTP in charge of purifying wastewater from the municipalities of A Coruña, Arteixo, Cambre, Culleredo and Oleiros, which comprise a population of nearly 400,000 inhabitants of the metropolitan area of A Coruña (NW Spain). The main objective of the project was to monitor the SARS-CoV-2 coronavirus epidemic in the metropolitan area of A Coruña.

COVIDBENS served as an early warning against possible outbreaks, since it proved to be able to anticipate between 2 and 3 weeks in the beginning of the pandemic waves with respect to the data on active cases reported by the health system (Trigo-Tasende et al. 2022). In addition, using the amount of genetic material of the virus present in the wastewater, nonparametric statistical models were used to estimate the number of infected people in the population (Vallejo et al. 2022).

Since December 2020, complying with the recommendation of the European Commission (https://ec.europa.eu/environment/pdf/water/recommendation_covid19_monitoring_wastewaters.pdf), the COVIDBENS team has been in charge of monitoring the emergence of new mutations and variants of SARS-CoV-2 in the wastewater arriving at the Bens WWTP using massive sequencing technologies. With the collaboration of Aguas de Galicia and EDAR Bens S.A., this challenge was tackled using two different strategies: 1) amplicon sequencing and 2) shotgun sequencing with enrichment of human respiratory viruses. The results obtained by the COVIDBENS team showed that both

technologies are effective for the detection of SARS-CoV-2 mutations. Amplicon sequencing works very effectively to specifically detect SARS-CoV-2 mutations and variants, while shotgun sequencing should be oriented towards the epidemiological monitoring of respiratory viruses in general (SARS-CoV-2, influenza, RSV, etc.). It should be noted that these techniques made it possible to retrospectively detect mutations of the Alfa variant in samples from the metropolitan area of A Coruña at the beginning of December, a month before that variant was detected in clinical samples, demonstrating the great potential of genome analysis of SARS-CoV-2 in wastewater for early epidemiological detection of variants. Once the methodology was fine-tuned and contrasted, it was decided to implement amplicon sequencing as a routine mutation tracking method. The genetic material was extracted and sequenced from samples obtained weekly. Data were analysed for surveillance mutations recommended by ECDC (European Center for Disease Prevention and Control), guidelines updated on March 11, 2022 (https://www.ecdc.europa.eu/en/covid-19/variants-concern).

In the period May 2021 – March 2020, the SARS-CoV-2 sequencing work in wastewater carried out by COVIDBENS enabled reporting on the evolution in the presence of mutations and variants in the metropolitan area of A Coruña on a weekly basis. The data obtained through sequencing and analysis of mutations and variants of the virus can be viewed at the link http://www.edarbens.es/covid19.

The statistical methods presented in the second section were used to estimate weekly the proportions of SARS-CoV-2 variants in the metropolitan area of A Coruña. For facilitating visual interpretation, the estimates of the proportions along time were smoothed with a local polynomial regression estimator. The smoothing parameters were selected using plug-in methods (see Loader, 1999).

Figure 1 contains the smoothed estimates of the SARS-CoV-2 variant proportions along time in the period May 2021 – March 2022. The decrease of the Alpha variant (B.1.1.7) is shown at the beginning of the time period under study. The irruption of the Delta variant (B.1.617.2), its subsequent predominance and final vanishing are observed during this period. In the time interval December 2021 – January 2022, the Omicron variant (B.1.1.529) appeared and abruptly increased, which was parallel to a sudden decrease of the Delta variant. The BA.2 Omicron subvariant also exhibits a sudden increase in February 2022.

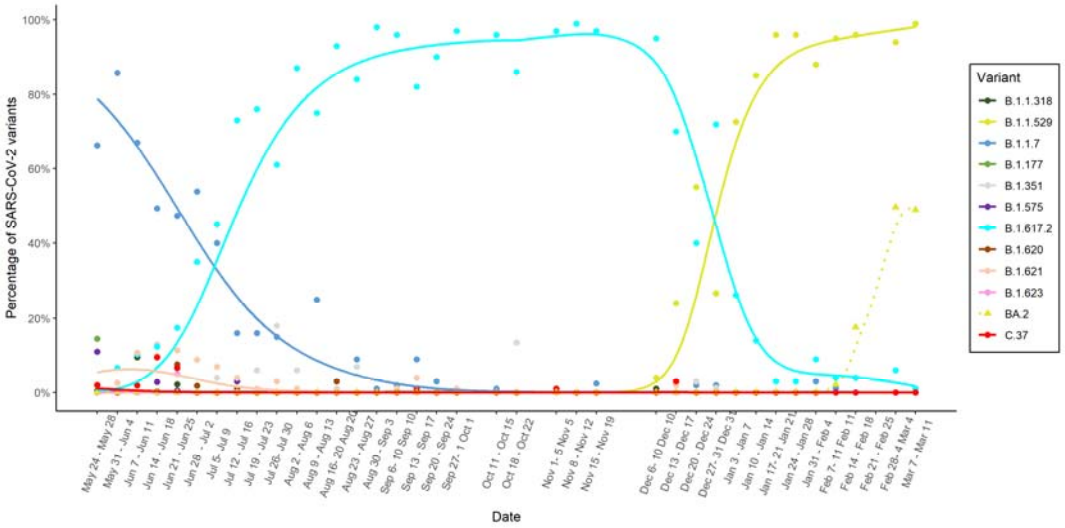

Figure 1: Smooth estimation of the SARS-CoV-2 variant proportions along time in the metropolitan area of A Coruña in the period May 2021 – March 2020.

References

Ahmed, W., Angel, N., Edson, J., Bibby, K., Bivins, A., O'Brien, J.W., Choi, P.M., Kitajima, M., Simpson, S.L., Li, J., et al., 2020. First confirmed detection of SARS-CoV-2 in untreated wastewater in Australia: a proof of concept for the wastewater surveillance of COVID-19 in the community. Sci. Total Environ. 728, 138764.

Barbeito, I., Cao, R., Ladra, S., López de Ullibarri, I., Posada, D., Poza, M., Tarrío, J., Vaamonde, M., Vallejo, J.A., Freire, B., Gallego, P., Iglesias, I., Rumbo, S., Tomás, L., Trigo, N., Alvariño, P., Beade, E., de Chiara, L., Estévez, N., 2022. Wastewater-based epidemiological modelling of SARS-CoV-2 viral load and monitorization of genomic variants in urban metropolitan areas. 40th Annual Meeting of the Spanish Society for Epidemiology.

Bar-Or, I., Weil, M., Indenbaum, V., Bucris, E., Bar-Ilan, D., Elul, M., Levi, N., Aguvaev, I., Cohen, Z., Shirazi, R., Erster, O., Sela-Brown, A., Sofer, D., Mor, O., Mendelson, E., Zuckerman, N.S., 2021. Detection of SARS-CoV-2 variants by genomic analysis of wastewater samples in Israel. Sci. Total Environ., 789, 148002.

Cevik, M., Tate, M., Lloyd, O., Maraolo, A.E., Schafers, J., Ho, A., 2021. SARS–CoV–2, SARS–CoV, and MERS–CoV viral load dynamics, duration of viral shedding, and infectiousness: a systematic review and meta-analysis. Lancet Microbe 2 (1), 13–22.

Chen, Y., Chen, L., Deng, Q., Zhang, G., Wu, K., Ni, L., Yang, Y., Liu, B., Wang, W., Wei, C., et al., 2020. The presence of SARS-CoV-2 RNA in the feces of COVID-19 patients. J. Med. Virol. 92 (7), 833–840.

EMCDDA, E.B., 2020.Wastewater Analysis and Drugs: A EuropeanMulti-city Study. European Monitoring Center for Drugs and Drug Addiction, pp. 1–14.

Gafurov, A., Baláž, A., Amman, F., Boršová, K., Čabanová, V., Klempa, B., Bergthaler, A., Vinař, T., Brejová, B., 2022. VirPool: model-based estimation of SARS-CoV-2 variant proportions in wastewater samples. BMC Bioinformatics, 19, 23 (1), 551.

Gupta, S., Parker, J., Smits, S., Underwood, J., Dolwani, S., 2020. Persistent viral shedding of SARS-CoV-2 in faeces–a rapid review. Color. Dis. 22 (6), 611–620.

Karthikeyan, S., Levy, J.I, De Hoff, P., Humphrey, G., Birmingham, A., Jepsen, K., Farmer, S., Tubb, H.M., Valles, T., Tribelhorn, C.E., Tsai, R., Aigner, S., Sathe, S., Moshiri, N., Henson, B., Mark,A.M., Hakim, A., Baer, N.A., Barber, T., B elda-Ferre, P., Chacón, M., Cheung, W., Cresini, E.S., Eisner, E.R., Lastrella, A.L., Lawrence, E.S., Marotz, C.A., Ngo, T.T., Ostrander, T., Plascencia, A., Salido, R.A., Seaver, Ph., Smoot, E.W., McDonald, D., Neuhard, R.M., Scioscia, A.L., Satterlund, A.M., Simmons, E.H., Abelman, D.B., Brenner, D., Bruner, J.C., Buckley, A., Ellison, M., Gattas, J., Gonias, S.L, Hale, M., Hawkins, F., Ikeda, L., Jhaveri, H., Johnson, T., Kellen, V., Kremer, B., Matthews, G., McLawhon, R.W, Ouillet, P., Park, D., Pradenas, A., Reed, S., Riggs, L., Sanders, A., Sollenberger, B., Song, A., White, B., Winbush, T., Aceves, C.M., Anderson, C., Gangavarapu, K., Hufbauer, E., Kurzban, E., Lee, J., Matteson N.L., Parker, E., Perkins, S.A., Ramesh, K.S., Robles-Sikisaka, R., Schwab, M.A., Spencer, E., Wohl, S., Nicholson, L., McHardy, I.H., Dimmock, D.P., Hobbs, C.A., Bakhtar, O., Harding, A., Mendoza, A., Bolze, A., Becker, D., Cirulli, E.T., Isaksson, M., Barrett, K.M.S., Washington, N.L., Malone, J.D., Schafer, A.M., Gurfield, N., Stous, S., Fielding-Miller, R., Garfein, R.S., Gaines, T., Anderson, C., Martin, N.K., Schooley, R., Austin, B., MacCannell,

D.R., Kingsmore, S.F., Lee, W., Shah, S., McDonald, E., Yu, A.T., Zeller, M., Fisch, K.M., Longhurst, C., Maysent, P., Pride, D., Khosla, P.K., Laurent, L.C., Yeo, G.W., Andersen, K.G., Knight, R., 2022. Wastewater sequencing reveals early cryptic SARS-CoV-2 variant transmission. Nature, 609 (7925), 101-108.

Loader, C.R., 1999. Bandwidth selection: classical or plug-in?. Ann. Stat., 27 (2), 415-438.

Lodder, W., de Roda Husman, A.M., 2020. SARS-CoV-2 in wastewater: potential health risk, but also data source. Lancet Gastroenterol. Hepatol. 5 (6), 533-534.

Medema, G., Heijnen, L., Elsinga, G., Italiaander, R., Brouwer, A., 2020. Presence of SARSCoronavirus-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. 7 (7), 511–516.

Miura, F., Kitajima, M., Omori, R., 2021. Duration of SARS–CoV–2 viral shedding in faeces as a parameter for wastewater-based epidemiology: re-analysis of patient data using a shedding dynamics model. Sci. Total Environ. 769, 144549.

Peccia, J., Zulli, A., Brackney, D.E., Grubaugh, N.D., Kaplan, E.H., Casanovas-Massana, A., Ko, A.I., Malik, A.A., Wang, D., Wang, M., et al., 2020. Measurement of SARS-CoV-2 RNA in wastewater tracks community infection dynamics. Nat. Biotechnol. 38 (10), 1164–1167.

Pollán, M., Pérez-Gómez, B., Pastor-Barriuso, R., Oteo, J., Hernán, M.A., Pérez-Olmeda, M., Sanmartín, J.L., Fernández-García, A., Cruz, I., de Larrea, N.F., et al., 2020. Prevalence of SARS–CoV–2 in Spain (ENE–COVID): a nationwide, population–based seroepidemiological study. Lancet 396 (10250), 535–544.

Radu, E., Masseron, A., Amman, F., Schedl, A., Agerer, B., Endler, L., Penz, T., Bock, C., Bergthaler, A., Vierheilig, J., Hufnagl, P., Korschineck, I., Krampe, J., Kreuzinger, N., 2022. Emergence of SARS-CoV-2 Alpha lineage and its correlation with quantitative wastewater-based epidemiology data. Water Research 215, 118257.

Randazzo,W., Cuevas-Ferrando, E., Sanjuán, R., Domingo-Calap, P., Sánchez, G., 2020a. Metropolitan wastewater analysis for COVID-19 epidemiological surveillance. Int. J. Hyg. Environ. Health 230, 113621.

Trigo-Tasende, N. Vallejo, J.A., Rumbo-Feal, S., Conde-Pérez, K. Vaamonde, M. López-Oriona, A., Barbeito, I., Nasser-Ali, M., Reif, R., Rodiño-Janeiro, B.K., Fernández-Álvarez, E., Iglesias-Corrás, I., Freire, B. Tarrío-Saavedra, J., Tomás, L., Gallego-García, P., Posada, D., Bou, G., López-de-Ullibarri, I., Cao, R., Ladra, S., Poza, M., 2022. COVIDBENS: a multidisciplinary surveillance program for SARS-CoV-2 in wastewater in A Coruña, Spain. Submitted for possible publication.

Valieris, R., Drummond, R.D., Defelicibus, A., Dias-Neto, E., Rosales, R.A., Tojal da Silva, I., 2022. A mixture model for determining SARS-Cov-2 variant composition in pooled samples. Bioinformatics. 38 (7), 1809-1815.

Vallejo, J.A., Trigo-Tasende, N., Rumbo-Feal, S., Conde-Pérez, K., López-Oriona, Á., Barbeito, I., Vaamonde, M., Tarrío-Saavedra, J. . Reif, R., Ladra, S. , Rodiño-Janeiro, B.K., Nasser-Alia, M., Cid, Á., Veiga, M.C., Acevedo, A., Lamora, C., Bou, G., Cao, R., Poza, M. 2022. Modeling the number of people infected with SARS-COV-2 from wastewater viral load in Northwest Spain. Science of the Total Environment, 811, 152334.

Wu, F., Zhang, J., Xiao, A., Gu, X., Lee, W.L., Armas, F., Kauffman, K., Hanage, W., Matus, M., Ghaeli, N., et al., 2020a. SARS-CoV-2 titers in wastewater are higher than expected from clinically confirmed cases. Msystems 5 (4).

Wurtzer, S., Marechal, V., Mouchel, J.M., Maday, Y., Teyssou, R., Richard, E., Almayrac, J.L., Moulin, L., 2020. Evaluation of lockdown impact on SARS-CoV-2 dynamics through viral genome quantification in Paris wastewaters. MedRxiv https://doi.org/10.1101/2020.04.12. 20062679.

Xing, Y.H., Ni,W.,Wu, Q., Li,W.J., Li, G.J., Wang,W.D., Tong, J.N., Song, X.F.,Wong, G.W.K., Xing, Q.S., 2020. Prolonged viral shedding in feces of pediatric patients with coronavirus disease 2019. J. Microbiol. Immunol. Infect. 53 (3), 473–480.

Xu, Y., Li, X., Zhu, B., Liang, H., Fang, C., Gong, Y., Guo, Q., Sun, X., Zhao, D., Shen, J., et al., 2020. Characteristics of pediatric SARS-CoV-2 infection and potential evidence for persistent fecal viral shedding. Nat. Med. 26 (4), 502–505.

Zhang, T., Cui, X., Zhao, X., Wang, J., Zheng, J., Zheng, G., Guo, W., Cai, C., He, S., Xu, Y., 2020. Detectable SARS-CoV-2 viral RNA in feces of three children during recovery period of COVID-19 pneumonia. J. Med. Virol. 92 (7), 909–914.