Modeling human pollution in water bodies using somatic coliphages and bacteriophages that infect *Bacteroides thetaiotaomicron* strain GA17

Méndez, Javier*[1, 3]; García-Aljaro, Cristina[1]; Muniesa, Maite[1]; Pascual-Benito, Míriam[1]; Ballesté, Elisenda[1], López, Pere[2, 3]; Monleón, Antonio[2, 3]; Blanch, Anicet R.[1] and Lucena, Francisco[1, 3].

1- Section of Microbiology. Department of Genetics, Microbiology and Statistics, Faculty of Biology, University of Barcelona, Av. Diagonal 643, 08028 Barcelona, Spain.

2 - Section of Statistics. Department of Genetics, Microbiology and Statistics, Faculty of Biology, University of Barcelona, Av. Diagonal 643, 08028 Barcelona, Spain.

3 - BIOST3 Group. Section of Statistics. Department of Genetics, Microbiology and Statistics. University of Barcelona, Av. Diagonal 643, 08028 Barcelona, Spain.

JM: jmendez@ub.edu; CGA: crgarcia@ub.edu; MM: mmuniesa@ub.edu; MP: mpascualbenito@ub.edu; EB: eballeste@ub.edu; PL: plopezbr8@alumnes.ub.edu; AM: amonleong@ub.edu; JJ: jjofre@ub.edu; ARB: ablanch@ub.edu; FL: flucena@ub.edu

Running title: *Human pollution modeling in water bodies*

*Corresponding author:

Javier Méndez

Email: jmendez@ub.edu

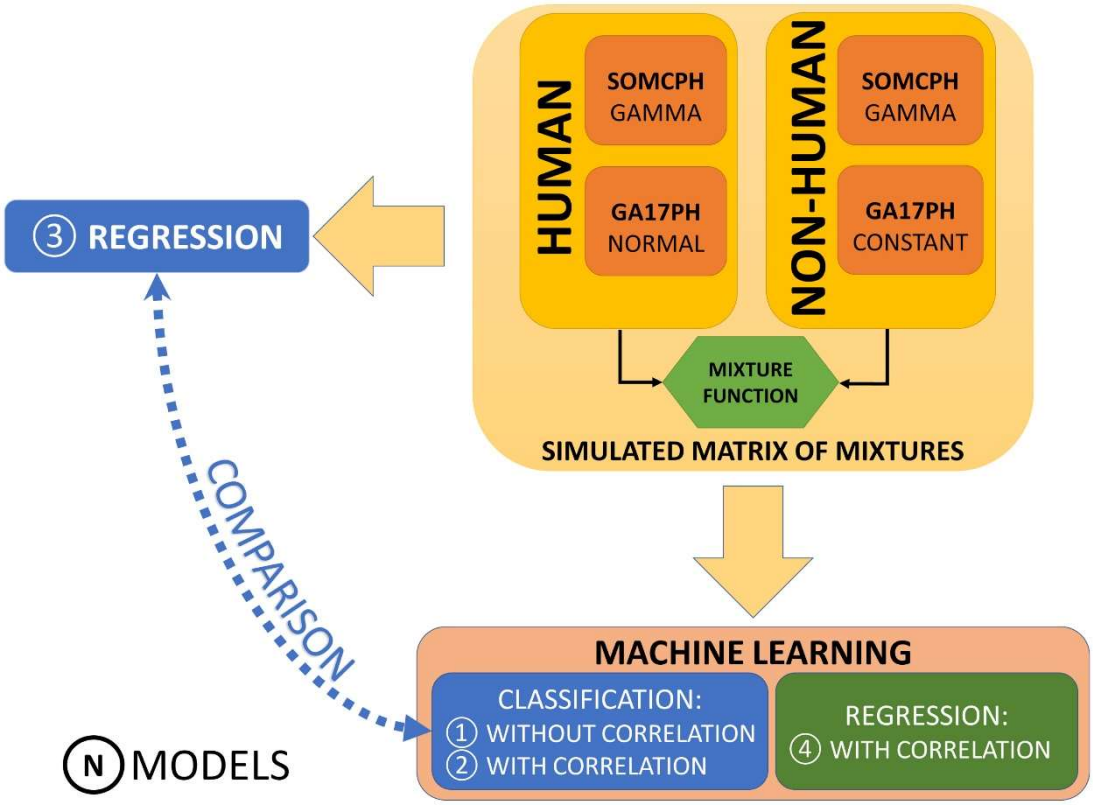
Tel: +34934021503

Abstract

The ability to detect human fecal pollution in water is of great importance when assessing the associated health risks. Many microbial source tracking (MST) markers have been proposed to determine the origin of fecal pollution, but their application remains challenging. A range of factors, not yet sufficiently analyzed, may affect MST markers in the environment, such as dilution and inactivation processes. In this work, a statistical framework based on Monte Carlo simulations and non-linear regression was used to develop a classification procedure for use in MST studies. The predictive model tested uses only two parameters: somatic coliphages (SOMCPH), as an index of general fecal pollution, and human host-specific bacteriophages that infect *Bacteroides thetaiotaomicron* strain GA17 (GA17PH). Taking into account bacteriophage dilution and differential inactivation, the threshold concentration of SOMCPH was calculated to be around 500 PFU/100 mL for a limit of detection of 10 PFU/100 mL. However, this threshold can be lowered by increasing the analyzed volume sample, which in turn lowers the limit of detection. The resulting model is sufficiently accurate for application in practical cases involving MST and could be easily used with markers other than those tested here.

Keywords: fecal pollution, MST, water, somatic coliphages, GA17 bacteriophages

Graphical abstract



Highlights

- A model to predict the content of human fecal pollution in water was developed.
- Human pollution can be predicted using only two microbiological parameters (SOMCPH and GA17PH).
- The effect of natural and artificial inactivation was considered in the predictions.
- The model can be used in MST studies with other markers.

1. Introduction

Waterborne pathogens originating from fecal pollution are major contributors to infectious disease outbreaks around the world. Fecal pollution can reach water bodies from various sources, including wildlife and a direct discharge of human fecal waste during rainfall events (Garcia-Aljaro et al. 2017). The risk to human health depends on the pathogen type and the fecal load. Microbial source tracking (MST) is an emerging area of applied environmental microbiology that describes a suite of methods and investigative strategies that can be used to detect fecal water pollution from different hosts, such as humans, livestock, and wildlife.

Relatively successful MST approaches have been based on the detection of microbial markers (host-specific pathogens or commensal microorganisms) (Harwood *et al.*, 2011; Jofre *et al.*, 2014; Lee *et al.*, 2011; Lee *et al.*, 2009; McMinn *et al.*, 2014; Noble *et al.*, 2003; Stapleton *et al.*, 2007), using phenotypic and molecular-based methods (Caldwell *et al.*, 2007; Dubinsky *et al.*, 2012; Gomez-Donate *et al.*, 2012; Mauffret *et al.*, 2013; Gomez-Donate *et al.*, 2016; Blanch *et al.*, 2016; Shanks *et al.*, 2016; Harwood *et al.*, 2017; Jebri *et al.*, 2017). However, a better discrimination is achieved when MST markers are used in combination rather than individually. Predictive models recently developed with inductive machine learning systems accurately predicted the source of fecal water contamination over a wide European geographical area; 4 main pollution inputs were considered (human, porcine, bovine and poultry), not only at point source but also after dilution, and the environmental decay of the markers was also taken into account (Balleste *et al.* 2020). As these prediction models involve a variable number of parameters, the laboratory and computing resources may be unaffordable for the end-user. However, a previous multi-laboratory study of several chemical and microbiological markers identified a set of only two variables that allowed a correct

86 classification of wastewaters and slurries of human and non-human fecal origin at point
87 source. This set, comprising the ratio between the logarithmic values of somatic
88 coliphages (SOMCPH) and bacteriophages infecting strain GA17 of *Bacteroides*
89 *thetaiotaomicron* (GA17PH), has been proposed as a discriminate marker of human or
90 animal pollution (Blanch *et al.*, 2006, Muniesa *et al*., 2012).
91 Bacteriophages infecting certain strains of *Bacteroides* spp. are mostly detected in fecal
92 pollution of human origin (Ebdon *et al.*, 2007; Jofre *et al.*, 2014; McMinn *et al.*, 2014;
93 Puig *et al.*, 1999; Tartera *et al.*, 1989). As well as host-specificity, these bacteriophages
94 have other characteristics of a practicable fecal marker, such as feasible numerical
95 detection and temporal stability. Previous studies have shown that the prevalence of
96 phages in human-specific bacterial strains of *Bacteroides* might vary among human
97 populations, so the bacterial strain should be chosen on the criteria of human-specificity
98 and the highest bacteriophage count. *Bacteroides thethaiotaomicron* GA17 has been
99 tested for human specificity in several locations and has given an excellent performance
100 in some countries of Europe (Payán et al., 2005; Ballesté et al. 2021), South America
101 and Northern Africa (Venegas et al., 2015; Yahya et al., 2015). Other *Bacteroides* host
102 strains such as *B. fragilis* GB124 have proved suitable for determining human-specific
103 phages in other regions (Edbon *et al.*, 2011).
104 The aim of the current study was to assess whether it is possible to predict the fraction
105 of human fecal pollution in a water body using only two variables, SOMCPH and
106 GA17PH, which have already been successfully used to predict human versus non-
107 human fecal pollution at point source. The effects of dilution and environmental decay
108 in the receiving waters on both selected indicators were evaluated. Advantages and
109 limitations are discussed, including the sampling size, consequences of indicator
110 inactivation, and the dilution effects.

2. Material and methods

2.1. Bacteriophage enumeration

Bacteriophage enumeration was carried out according to ISO standards 10705-2 (ISO, 2000) and 10705-4 (ISO, 2001). Briefly, wastewater samples were filtered through low protein-binding polysulfone membrane filters with a 0.22 μm pore size. Filtered samples were diluted 100-, 1,000- and 10,000-fold. Phages in each diluted sample were enumerated using a double-layer agar plaque assay procedure as described in the ISO standards (ISO, 2000, 2001). For somatic coliphages, the host was *Escherichia coli* strain WG5 whereas for the GA17PH titration the strain GA17 of *Bacteroides thethaiotaomicron* was used. Briefly, DAL procedure was carried out by pouring 2.5 mL of a complete semisolid agar + 1 mL of the sample + 1 mL of a culture of the host strain. After adding the host bacteria, each tube was mixed carefully avoiding bubble formation and the content was poured onto an appropriate agar plate. For somatic coliphages the composition of the agar plates was Modified Scholtens' Agar (MSA) and the semisolid (ssMSA) was made using the half mass agar of the MSA. ssMSA can be supplemented with nalidíxic acid to a final concentration of 250 μg/mL when a high background microbiota is expected.

For titration of GA17PH, the media used for the agar plates was Bacteroides Phage Recovery Medium Agar (BPRMA) and the overlay was performed using ssBPRMA. Once the semisolid was poured, agar plates were allowed to solidify and incubated upside-down at (36 ± 2) °C for (18 ± 2) hours. Plates for GA17PH titration were incubated in anaerobiosis. After incubation plaques were counted.

Quantification of GA17PH in non-human sewage was carried out by titrating 10 replicates of 1 mL of undiluted wastewater. Although the limit of detection in these conditions was 1 PFU/10 mL, it was assumed to be 10 PFU/100 mL.

2.2. Sampling

2.2.1 Samples used to build the models

Human SOMCPH and GA17PH were measured in 53 samples collected from raw influents of four municipal wastewater treatment plants (41°16'36.0"N, 2°02'30.9"E; 41°31'29.9"N, 2°25'30.7"E; 41°48'29.2"N, 3°01'47.0"E; 42°14'39.5"N, 3°06'13.1"E). Non-human bacteriophages were measured in 33 samples collected from sewage of six abattoirs dealing exclusively with pigs, poultry and cattle, in which GA17PH was not detected. The human and non-human SOMCPH and GA17PH distribution functions were determined in these sets of samples.

To determine the degree of variability in a replica, the relative standard deviation (RSD) of human GA17PH was assayed from 43 counts of the same 1,000-fold diluted sample from urban raw wastewater.

The effect of bacteriophage inactivation due to wastewater treatment was determined from two secondary effluents of two wastewater treatment plants ((41°16'36.0"N, 2°02'30.9"E ; 41°31'29.9"N, 2°25'30.7"E; 108 samples) and from one tertiary effluent (41°31'29.9"N, 2°25'30.7"E; 65 samples: 12 chlorinated effluents, 40 UV-treated effluents and 13 samples submitted to both treatments). To model the natural decay of bacteriophages in fresh waters, data were extracted from Durán *et al.* 2002.

Briefly, natural inactivation of somatic coliphages and *Bacteroides* sp host phages was measured in summer and winter seasons. A settled urban raw wastewater was diluted in proportion 1/50 with river water, which content of bacteriophages between 3 and 4

159 decimal logarithmic units lower than counts in wastewater. That was to avoid
160 interference in the detection of the added bacteriophages and minimized the probability
161 of occurrence of bacteriophage replication. The inactivation study was carried out in the
162 site where samples were collected. Samples to measure inactivation, were prepared by
163 placing the spiked water river samples into dialysis tubes (cut-off 14 kDa) which, were
164 conveniently sealed and placed 20–25 cm deep in the river water in the same area where
165 it had been collected. Inactivation was followed by taking samples and titrating them at
166 various time intervals.
167 2.2.2 Samples used to test the models
168 The effect of artificial inactivation in treated wastewaters was evaluated using 54
169 samples from secondary and 36 from tertiary effluents.
170 In addition, 102 samples were collected from two rivers in Spain receiving pollution
171 from different sources (urban and rural). The Llobregat River (41°17'07.8"N,
172 2°03'09.9"E ; 14 samples), which in its lower transect flows through a highly populated
173 zone, served as a model of an anthropogenically contaminated river (Köck *et al.*, 2011).
174 Its fecal pollution comes mainly from secondary treated wastewaters directly discharged
175 into the river, as well as from diffuse pollution and run-off, mostly in its upper course.
176 In contrast, the Riudaura Stream (42°11'53.5"N, 2°21'44.5"E ; 88 samples) was used as
177 a model of a water course with a low level of human pollution, the main fecal source
178 being reclaimed water, farming and potentially some septic overload from scarcely
179 populated areas.
180
181 **2.3 Probability distribution fitting**

The Monte Carlo method (MC) involves the use of probability distribution functions (PDF). Therefore, selecting the most appropriate distributions to characterize the probability is one of the most important steps in MC.

Statistical analyses and modeling were carried out using the statistical software package R version 3.5.3 (R Core Team, 2019). Distribution fitting, distribution simulation and correlation bootstrapping were performed using the R packages fitdistrplus (Delignette-Muller & Dutang, 2015), ExtDist (Wu *et al.*, 2015), mc2d (Pouillot & Delignette-Muller, 2010), RVAideMemoire (Hervé, 2019) and Matching (Sekhon, 2011). Distribution functions were selected according to their Akaike information criterion and their goodness-of-fit using a bootstrap version of the Kolmogorov-Smirnov (KS) test.

2.4. Development of the predictive model

Four predictive models were developed. The ratios between the log-values of SOMCPH and GA17PH and the Spearman's correlation coefficient (*rho* value) of both parameters were used as predictor variables (covariables).

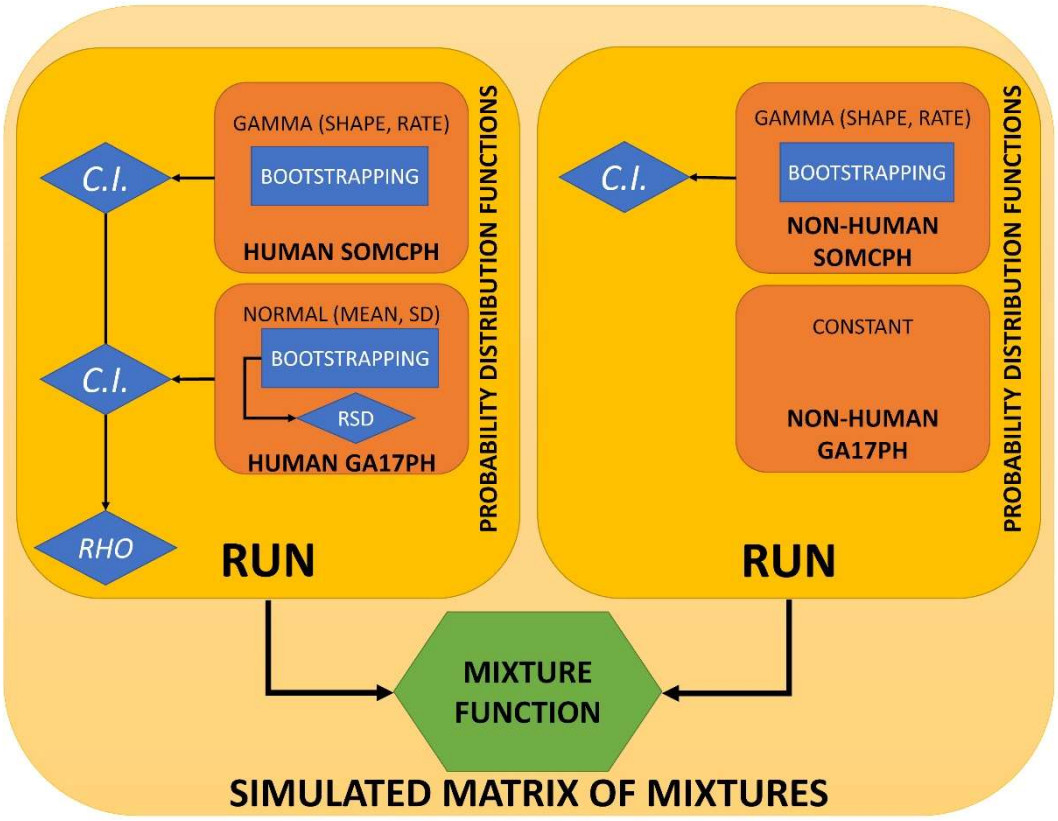
As already mentioned, the ratio between the logarithms of SOMCPH and GA17PH has been proposed as a tool for correctly classifying wastewaters and slurries of human and non-human origin at point source (Belanche-Muñoz & Blanch, 2008). In the case of a simple mixture of two wastewater samples, one from a human and the other from a non-human source, the ratio of logarithms of both bacteriophages can be defined as

$$Ratio=\log10((Fraction_{human}\cdot SOMCPH_{human}) + (Fraction_{non\text{-}human}\cdot SOMCPH_{non\text{-}human}))/\log10((Fraction_{human}\cdot GA17PH_{human})+(Fraction_{non\text{-}human}\cdot GA17PH_{non\text{-}human})$$

where $Fraction_{non\text{-}human} = 1 - Fraction_{human}$.

To develop the models (Figure 1), the first step was to determine the PDF for each marker (SOMCPH and GA17PH) in urban raw wastewater and abattoir sewage, which were considered as the respective point sources of human and non-human pollution.

Figure 1. Diagram of the steps used to obtain the simulated mixture matrix.



Four PDFs were defined: a function for GA17PH in human polluted water samples, which was defined under a normal distribution; a function for GA17PH in non-human polluted water samples, which was defined as a constant value; and two gamma distribution functions for SOMCPH in human and non-human polluted water samples. Distribution functions were truncated at the minimum value, which was set to 10 PFU/ 100 mL. When necessary, values were discretized by rounding. As selecting the most appropriate PDF is the most important step in developing the predictive model, outlier and extreme values were removed using Grubb's test (Grubbs, 1950).

The goodness of the bacteriophage distribution fitting was assessed using a bootstrap version of the Kolmogorov-Smirnov test. When several functions fitted the data, Akaike's information criterion (AIC) was used. AIC, which is based on information theory, provides a "measure" of the relative quality of a statistical model. When several candidate functions can be satisfactorily implemented with the same dataset, the model with the lowest AIC best explains the data.

The resulting PDFs were used in a MC simulation of 101 mixtures of wastewaters, containing a range of human pollution from 0 % to 100 % with a step of 1 %. For each mixture, 100 runs were simulated and a bootstrapping of 5,000 replicas was carried out in each run. Each replica was calculated by resampling each bacteriophage from its own PDF.

The values of *rho* and the ratios between the two bacteriophages were obtained from the simulated cases. To avoid outliers in the simulations, only those values that lay within the 97.5% confidence interval of the PDF were allowed.

Spearman's correlation was determined in human samples. The confidence interval for the *rho* parameter was calculated by bootstrapping (1,000 replicas) from the original human data set. This confidence interval was used as a control step, only allowing simulated mixtures into the simulation if the correlation of the simulated human portion lay within the bootstrapped confidence interval of the original human set.

Moreover, the human GA17PH RSD was used as an additional control step to prevent overdispersion in the bootstrap replicas. In the bootstrapping step with 5,000 replicas, only simulated GA17PH samples with an RSD up to 1.5 times the 97.5^{th} percentile of the previously assessed RSD were allowed.

After the simulation, a matrix containing 10,100 rows of human content percentage, bacteriophage ratios and correlations (101 simulated mixture sets, with each mixture

containing 100 values) was obtained. This simulated matrix of mixtures was used to classify and predict the percentage of human content. The four models were developed as follows. First, a machine learning classifier algorithm was used with only the bacteriophage ratio as a predictor variable (1); the impact of correlation on classification was then taken into account (2). A classification model based only on regression analysis (3) was developed to compare the results with those of the two machine learning classification models. This comparison would serve as a control of the synthetic generation of samples, as an incorrect generation and sample selection would result in an overfitted classification. Finally, a regression machine learning model was trained (4) to predict the numeric percentage of the human content of the samples based on the bacteriophage ratio and correlation. All models were evaluated using independent samples of known pollution source.

2.4.1 Classification of the samples by machine learning

Machine learning classification was carried out using the R package caret (Kuhn *et al.*, 2019). Three classification groups were arbitrarily defined: a "non-human class" for samples containing ≤ 33 % of human content; a "mixed class" for samples containing from 33 % to 66 % of human content; and a "human class" for samples containing ≥ 66 % of human content.

Briefly, the simulated matrix of mixtures was randomly split into two sets: a training set (70 % of the samples), which was used to train the learning algorithm; and a validation set (the remaining 30 %), which was used to validate the predictions. One classification model was developed using the bacteriophage ratio as a predictor variable, whereas in the second classification model the predictor variables were the bacteriophage ratio and Spearman's correlations.

The training was carried out with the KNN classifier, using the three groups against their corresponding predictor variables with a 10-fold cross-validation. The accuracy of the classification was evaluated with a confusion matrix using the validation set.

2.4.2 Classification of the samples by regression analysis

A classification model based on regression was developed using a non-linear mixed effects model of regression analysis based on a linear-quadratic rational model ($(a + b \cdot x)/(1 + c \cdot x + d \cdot x^2)$) . Values of the human content of the simulated matrix of mixtures were fitted against the simulated ratio values.

As in a usual regression, the values of the four parameters (a, b, c, and d) were determined. In addition, two parameters (a and b, specified as random effects) were allowed to vary in order to adapt the curve for the values of each run.

The starting values for the regression were obtained using the library minpack.lm (Elzhov *et al.*, 2016) and the mixed effects analysis was carried out using the library nlme (Pinheiro *et al.*, 2019).

To verify the goodness of the regression model, a linear regression between the predicted human content values versus the simulated human content values was carried out.

The same three classification groups were defined and used to determine the accuracy of the predicted classification groups. Notably, this model involved all the data from the matrix mixture.

2.4.3 Regression with machine learning

As in the machine learning classification, the simulated matrix of mixtures was randomly split into training (70 %) and validation (30 %) sets. Training was carried out

against the percentage of human content in the samples. The bacteriophage ratio and correlation were used as predictor variables.

The KNN learning algorithm was trained with a 10-fold cross-validation. Finally, after the training, the goodness of fit was assessed using the validation set.

2.4.4 Performance metrics used to evaluate machine learning models

To evaluate the performance of this models several metrics were used.

For classification models, the value of the sensitivity, specificity, accuracy and the Cohen's Kappa were determined using a confusion matrix. The three first parameters score between 0 to 1. Sensitivity refers to the true positive rate that is a measure of the proportion of positives that are correctly identified. Specificity refers to the true negative rate, which is a measure of the proportion of negatives that are correctly identified. Accuracy is a measure of how well a binary classification test correctly identifies. In general, if the value gets higher, the better model is.

Concerning to Cohen's Kappa, the Kappa value provides a measure of the degree of agreement. That parameter is based on the accuracy and it varies between -1 to 1. When it scores 1 indicates a perfect agreement in the classification, a score of 0 indicates an agreement not better than chance; and a negative Kappa means that there is less agreement than would be expected by chance.

For classification and regression models, another parameter to determine the performance is the area under the ROC curve (AUC). Although AUC values can be between 0 and 1, usually, it lies between 0.5 to 1. An AUC > 0.75 indicates a good performance, an AUC > 0.90 implies that the performance of model is excellent and an AUC = 1 means that the performance of the model is perfect

Besides AUC, for regression were used the coefficient of determination (R^2), the mean absolute error (MAE) and root mean squared error (RMSE). The coefficient of determination scores between 0 and 1 and how well performances the regression model. MAE and RMSE are two metrics to measure the difference between the real values and the predicted ones.

2.4.5 Non-parametric tests used for comparing two or more samples.

Non-parametric test are used when the data cannot be assumed to be normally distributed. Two samples comparison has been carried out using the Wilcoxon signed rank test with continuity correction, which is a non-parametric alternative to two-sample t-test to determine whether the medians of the samples are equal.

When more than two samples were compared, the Kruskal-Wallis test was used. This test is an alternative to one-way ANOVA test by extending the two-samples Wilcoxon test to more than two groups.

2.5 Evaluation of the models with real samples

Although the machine learning models included a set of samples for their validation, the four models were evaluated using several sets of real samples of different origins.

The variability of bacteriophage counts in the environmental samples was simulated by taking into account different sample sizes (3 to 16 samples).

2.5.1 Estimation of minimum sample size

Regarding phage variability, an important issue arises related to it, and it is how determine the minimum sample size to make appropriate inferences about these bacteriophage populations. Two approaches were used to estimate the minimum sample

size. The first one assumed that the bacteriophage ratios can be defined under a normal distribution; and the sample size for the mean is estimated as $N = (z^2 \cdot \sigma^2)/e^2$, where N is the sample size, z is the abscissa of the normal curve that cuts off an area α at the tails (≈1.96 for an $\alpha = 0.05$, which corresponds to a confidence interval of 95%), e is the desired level of precision and σ^2 is the variance of the population.

In the second approach, which also assumes a normal distribution of the phage ratios, the minimum sample size was determined using library biotools (da Silva *et al.*, 2017). The confidence interval of the relative standard deviation of GA17PH in the human samples and a bootstrap of 10,000 replicas was used.

2.5.2 Minimum bacteriophage concentrations required for predictions

In this study, the minimum bacteriophage concentrations were calculated to ensure the predictions of the machine learning regression model were valid. The following three premises were assumed: 1) SOMCPH and GA17PH are randomly distributed in wastewater matrices. 2) The dilution process equally affects both bacteriophages. As the concentration of GA17PH in wastewater samples was lower, it was used to calculate the dilution limit. 3) The maximum admissible dilution fold provides less than 5 % of GA17PH-negative samples.

Briefly, 100 bootstrapped simulations were run, performing a dilution process with 1,000 replicas. Thus, water matrices contained different concentrations of GA17PH (its distribution function previously calculated), which were randomly sampled and diluted from 100- to 500,000-fold, and for each dilution fold the probability of presence/absence was calculated according to a Poisson likelihood.

2.5.3 Artificial inactivation of SOMCPH and GA17PH in wastewater treatments

Differential inactivation or survival of the two bacteriophages in water matrices may significantly affect the ratio and correlation, thereby altering the model prediction. One way to correct the bias caused by artificial treatments would be to study their effect on the selected bacteriophages and adjust the prediction models according to the level of inactivation. The main drawback of this approach is that it requires knowing the inactivation at one particular point. As the efficiency of different treatments may vary according to the physicochemical properties of the water matrix, an approximation may be achieved by using empirical or probability distribution functions related to the inactivation of each bacteriophage, a strategy used in microbial risk assessment. Briefly, a model with and without taking into account artificial inactivation was evaluated by replacement sampling, in which groups of 3 to 16 samples were randomly assembled to mimic different levels of replicas. The human content was determined twice for each group, once using the data as is, and then by adding the possible inactivated bacteriophages according to their own inactivation distribution function. This evaluation involved 100 replicas, in which the results comprised the effect of the different number of replicas (from 3 to 16) and the effect of the differential inactivation on the prediction of the human pollution content.

2.5.4 Natural inactivation of SOMCPH and GA17PH in river water matrices

To determine the effect of environmental decay on both bacteriophages, data were used from Durán et al. (2002), who assessed the natural decay of SOMCPH and GA17PH in river waters during the winter and summer seasons. Using these data, the bacteriophage decay was adjusted to two models, which were selected according the Akaike Information Criterion (AIC). One of the models was based on a power function ($a \cdot \text{time}^b$), which was used for SOMCPH, and the other based on the Gompertz relation

394 ($a \cdot \exp(-\exp(b - c \cdot \text{time}))$), which was used for GA17PH. As the environmental decay of
395 both bacteriophages differs according to season, a nonlinear mixed effect regression
396 model was assayed for each one. For the sake of simplicity, in both cases only one
397 variable was selected as a random parameter, based on the criteria of overall R^2
398 achieved in the regression, the homoscedasticity and the normality of the residual
399 distribution.
400
401 This evaluation involved 100 replicas, in which the results comprised the effect of
402 natural decay at 0, 72, 120, 160 and 360 hours, and the sampling size (from 3 to 16). For
403 each bacteriophages, two models of prediction were used (summer and winter seasons),
404 which were used to rectify the bacteriophage counts and calculate bacteriophage ratios
405 and correlations. The human pollution content was predicted using the machine learning
406 regression model.
407
408 **2.5.5 Evaluation procedure**
409 Models were evaluated using water samples of known origins: secondary and tertiary
410 treated effluents and river water. The limited number of samples was artificially
411 increased by resampling with replacement. Samples were randomly assembled in
412 groups of 3 to 16 to mimic different levels of replicas, and the human content was
413 determined for each group. This step was carried out 100 times. The goodness of
414 prediction of the classification models (mixed effect regression, and the two machine
415 learning classifiers) was measured using accuracy, and R^2 was used for the machine
416 learning regression. Additionally, the effects of artificial inactivation by wastewater
417 treatments and natural decay were taken into account.
418

3. Results

3.1. Bacteriophage concentration in the analyzed water samples

After the removal of outlier and extreme values, a total of 245 water samples were selected: 41 from urban sewage to determine human fecal pollution and 31 from abattoir sewages for non-human contamination were used to obtain the bacteriophage PDF (Table 1) and, 108 samples from secondary and 65 from tertiary treated urban sewage were assessed to determine the differential inactivation of both bacteriophages in wastewater treatments.

To evaluate the models, a total of 185 water samples of known sources (Table 2) were submitted to resampling with replacement. These samples were used to determine the feasibility of prediction and how this was affected by taking into account inactivation/decay.

To analyze natural decay, 14 samples from an urban human-polluted river (Llobregat River), and 81 samples from a water course considered to be without human pollution (Riudaura Stream) were used.

In all samples, bacteriophage counts ranged from 10^0 to 10^4 PFU/100 mL for GA17PH and from 10^3 to 10^6 PFU/100 mL for SOMCPH.

3.2. Definition of the probability distribution functions of the bacteriophages

Almost all the bacteriophage distributions in the various water types were satisfactorily adjusted to a gamma (SOMCPH) or normal (GA17PH) function, The shape and rate of the adjusted gamma functions, were respectively, of 2.817 and $7.021 \cdot 10^{-7}$ in human samples and 0.959 and $1.105 \cdot 10^{-7}$ in non-human samples. GA17PH in human samples displayed a mean of $8.335 \cdot 10^4$ and a standard deviation of $4.818 \cdot 10^4$. GA17PH in non-human fecal samples had to be defined as constant (Table 3), because, in contrast with

444 other studies (Gomez-Donate *et al.*, 2011; Payán *et al.*, 2005), no GA17PH
445 bacteriophages were detected in non-human samples. It was consequently adjusted to 10
446 PFU/100 mL, which is very close to the value obtained by Gómez-Doñate *et al.* (2011),
447 who detected five GA17PH-positive samples in a non-human set of 125 sewage
448 samples (sampling volume of 10 mL), with an average value of GA17PH in non-human
449 samples of 12 PFU/100 mL.
450 SOMCPH and GA17PH were also tested for correlation and dependency. In non-human
451 wastewater samples, the GA17PH and SOMCPH variables were independent and non-
452 correlated, as GA17PH was adjusted to constant as stated above.
453 SOMCPH and GA17PH from human fecal wastewater samples showed a statistically
454 significant dependency (p-value ≤ 0.05) with a Spearman's correlation coefficient of
455 0.514.
456
457 **3.3 Classification with the machine learning models**
458 Two classifications were carried out with the KNN algorithm and three classes were
459 defined according to the origin of the pollution. The classification based exclusively on
460 the ratio achieved an accuracy of 82.00% with a Kappa-value of 72.98 % for the
461 validation set. The sensitivity and specificity for each class of pollution were,
462 respectively, 86.74 % and 92.17 % for human, 72.06 % and 86.62 % for mixed, and
463 86.50 % and 94.36 % for non-human. The balanced accuracy for the three classes was,
464 respectively, 89.45 %, 79.34 % and 90.43%; and their 95 % confidence intervals of the
465 AUC values were 96.95 % to 97.98 %, 91.85 % to 93.85 % and 97.68 to 98.66 %,
466 respectively.
467 The classification using the ratio and Spearman's correlation achieved an accuracy of
468 94.00% for the validation set, with a Kappa-value of 90.99 %. The respective sensitivity

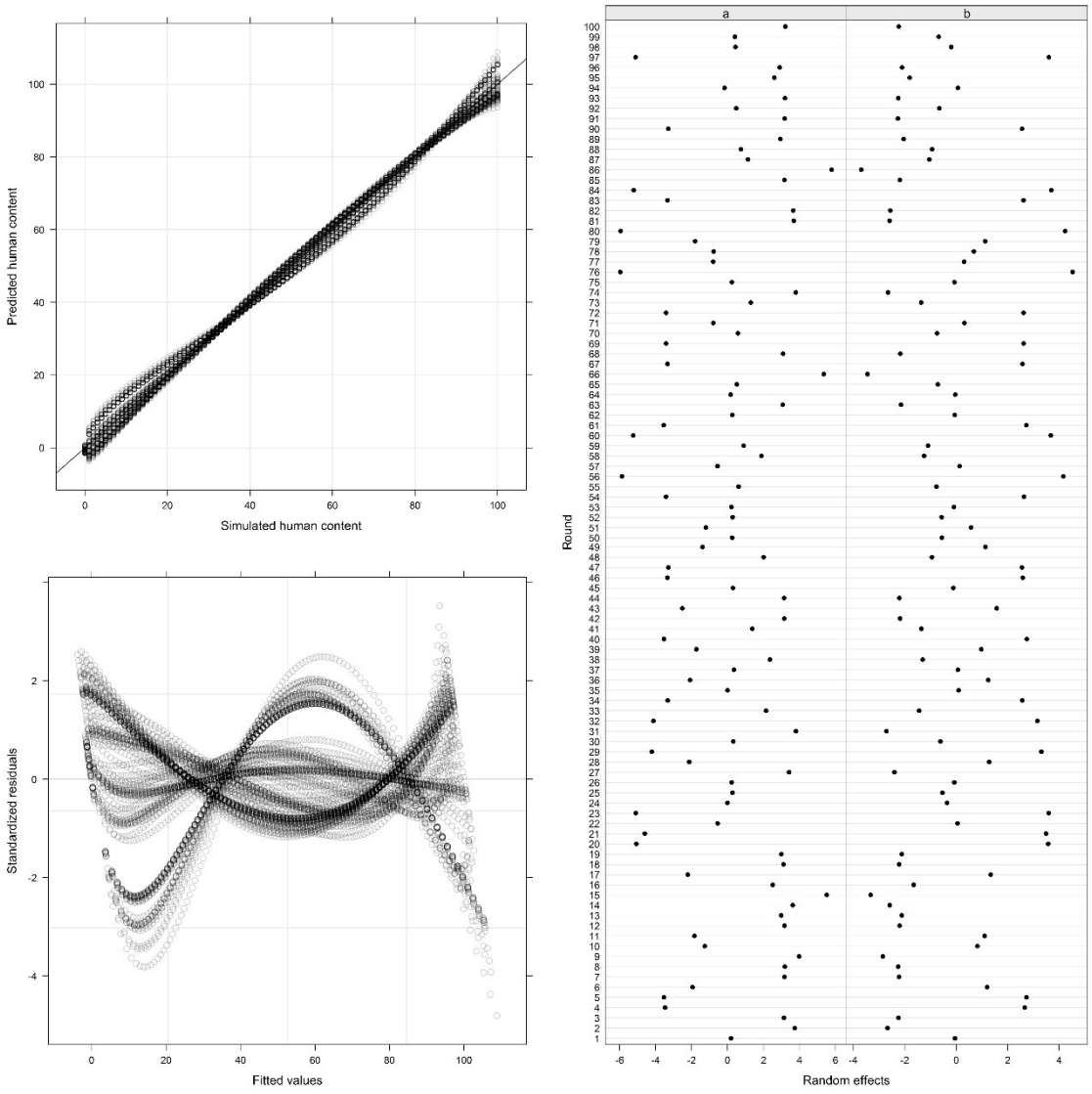
and specificity for each class were 97.70 % and 98.42 % for human, 90.79 % and 95.50 % for mixed, and 93.21 % and 97.15 % for non-human. The balanced accuracy was, respectively, 98.06 %, 93.15 % and 95.18%; and their 95 % confidence intervals of the AUC values were 99.32 % to 99.79 %, 96.87 % to 98.03 % and 98.34 % to 99.10 %, respectively.

3.4. Classification with the non-linear regression model

A non-linear mixed effects regression model was developed. To select the equation that best defines the regression curve, several non-linear models were previously assayed, as shown in Supplementary-Table 1 (only non-linear models that achieved an $R^2 > 80$ % are depicted). The best model was selected according to the AIC, and a linear/quadratic rational model ($(a + b \cdot x)/(1 + c \cdot x + d \cdot x^2)$) was selected for building the mixed effect regression.

The fixed parameters were estimated as $a = 6.0459$, $b = -2.2491$, $c = -1.4788$ and $d = 0.5615$ and the values of the random effects for the a-parameter fluctuated from -9.3183 to 8.3102 and for the b-parameter from -5.6966 to 6.7682. The results of the regression are shown in Supplementary-Figure 1.

Supplementary-Figure 1. In the top left linear regression plot, the simulated values are plotted against the fitted values, the diagonal line representing a perfect fit. In the bottom left, the standardized residuals are plotted against the percentage of human content in the mixture, and the variability of the random effects that impact parameters a and b is depicted.



493

494 In Supplementary-Figure 1, noteworthy the relationship between both random effects

495 (Pearson's correlation of -0.9937, p-value ≤ 0.05). The classification model indicated

496 that samples with a bacteriophage ratio expressed as log(SOMCPH):log(GA17PH) ≤

497 1.487 should be classified as "human", whereas samples with a ratio ≥ 1.640 should be

498 classified as "non-human" (human content ≤ 33.3 %), with the remaining samples

499 belonging to the "mixed class".

500 The overall accuracy of the classification was 82.83 %, with a Kappa-value of 74.24 %.

501 When considering the three pollution classes separately, their respective sensitivity and

502 specificity were 87.36 % and 93.17 % for human, 70.49 % and 89.83 % for mixed, and,

92.82 % and 91.23 % for non-human. The balanced accuracy for the three classes (human, mixed and non-human) was 90.27 %, 80.16 % and 92.03%, respectively. The similarity of these results to those obtained with machine learning suggests that overfitting did not occur in the machine learning classifications.

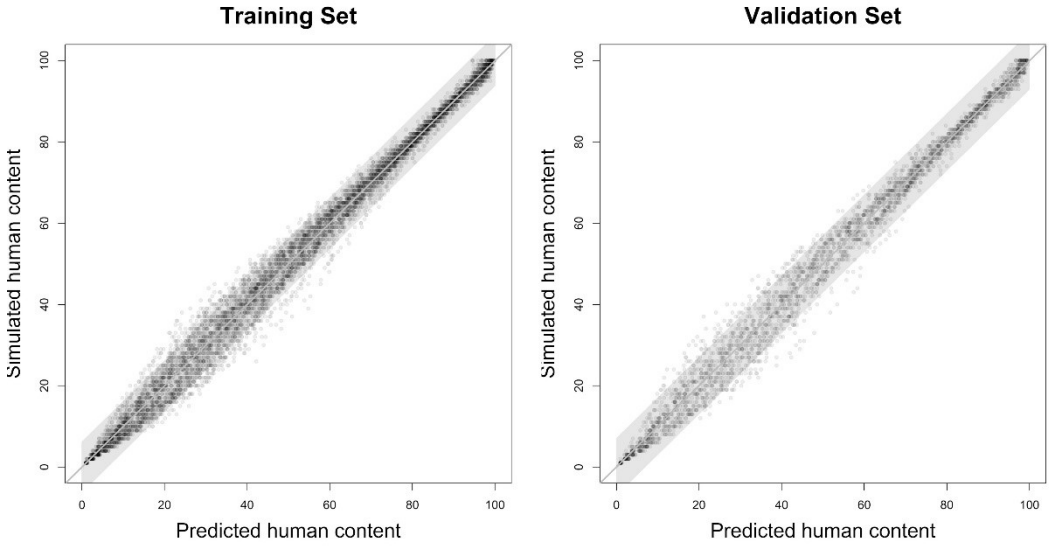
The model developed with real samples of known origin (Table 2) achieved an optimal classification of the treated secondary effluents, with a mean accuracy of 97.43 %. This value increased slightly to 99.57 % when bacteriophage inactivation was included. However, the regression model failed when it was applied to tertiary effluents, providing a mean accuracy of 0.64 %, which increased to 3.29 % when inactivation was considered. Significant differences were observed between the accuracies when inactivation was taken into account (Wilcoxon test, p-value ≤ 0.05). Individual accuracy for every group of samples is shown in Table 4.

3.5 Machine learning regression model

The selection criterion used for the machine learning regression was the root-mean-square error (RMSE), the optimal model having smallest RMSE value. The best result was obtained with 9 neighbors, providing an RMSE of 3.69, an MAE of 2.64 and an R^2 of 98.37 %.

To evaluate the performance of the model, a linear regression (Figure 2) was carried out, plotting the predicted human content (using the validation dataset) against the simulated human content. An RMSE of 3.73, an MAE of 2.68 and an adjusted R^2 of 98.29 % were obtained. The AUC was calculated by taking into account the probabilities of the predictions, the mean AUC was 98.46 % with a 95 % confidence interval of 95.53 % to 100%.

Figure 2. Linear regression of the predicted human content against the simulated human content of the training and validation sets calculated by the machine learning method.



3.6 Minimum sample size

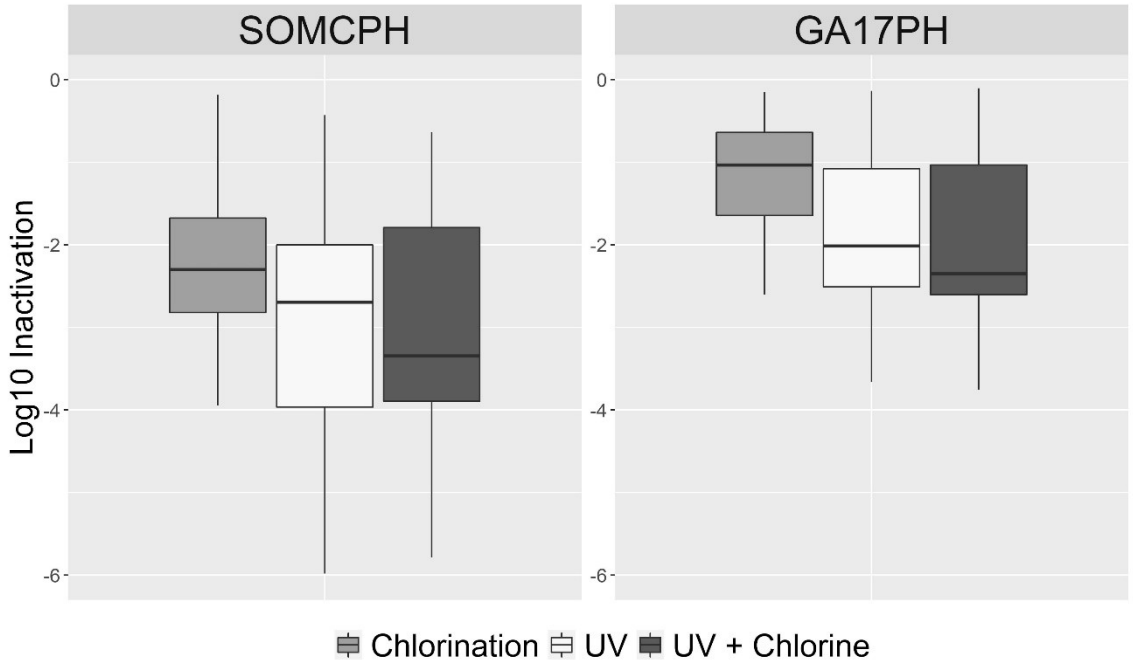
The bacteriophage ratio of the human samples in this study could be fitted to a normal distribution (KS-test, *p*-value > 0.05), which was defined by a mean of 1.3635 and a standard deviation of 0.0982. Assuming a relative error of 5%, the minimum sample size was approximately 15 samples. Additionally, the sample size calculated using the library biotools provided a similar sample size, 15 to 16 samples.

3.7 Minimum bacteriophage concentrations required for predictions

The maximum dilution fold allowed had a mean value of 19,475, with a minimum of 12,897 and a maximum of 41,481. The 97.5th percentiles for the concentrations of SOMCPH and GA17PH were of ≈ 510 PFU/100 mL (13.85 to 835.22) and ≈ 10 PFU/100 mL (0.174 to 14.896), respectively.

It should be noted that the 97.5th percentile for GA17PH coincides with the limit of detection for this bacteriophage, which could change according to its geographical distribution. The percentiles for both bacteriophages can be lowered by increasing the

546 analyzed volume (e.g. by increasing the number of plates or using concentration
547 methods).
548
549 **3.8. Effect of differential inactivation of the markers**
550 **3.8.1 Artificial inactivation in wastewater treatments**
551 In this work, the inactivation of each bacteriophage in the secondary and tertiary
552 effluents (54 and 36 samples, respectively) from an urban wastewater treatment plant
553 was fitted to a triangular function (KS-test, p-value > 0.05).
554 For SOMCPH the parameters of the triangular function were for the secondary a
555 minimum of -5.251, a mode of -1.836 and a maximum of -0.509; and for the tertiary a
556 minimum of -6.350, a mode of -2.176 and a maximum of 0.162.
557 For GA17PH the parameters of the triangular function were for the secondary a
558 minimum of -4.809, a mode of -1.818 and a maximum of -0.900; and for the tertiary, a
559 minimum of -4.384, a mode of -0.720 and a maximum of -0.103.
560 Differences in inactivation between the two phages were only apparent in tertiary
561 effluents (Wilcoxon test, p-value ≤ 0.05). Statistically significant differences were
562 observed for chlorination and UV treatments (Wilcoxon test, p-value ≤ 0.05) but not
563 when both were applied together (Wilcoxon test, p-value > 0.05). The boxplots in
564 Supplementary-Figure 2 depict the inactivation of the two bacteriophages induced by
565 different tertiary treatments.
566
567 Supplementary-Figure 2.-Boxplot on the left depicts the inactivation of somatic
568 coliphages (SOMCPH) and on the right, the human host-specific bacteriophages that
569 infect *Bacteroides thetaiotaomicron* strain GA17 (GA17PH).
570

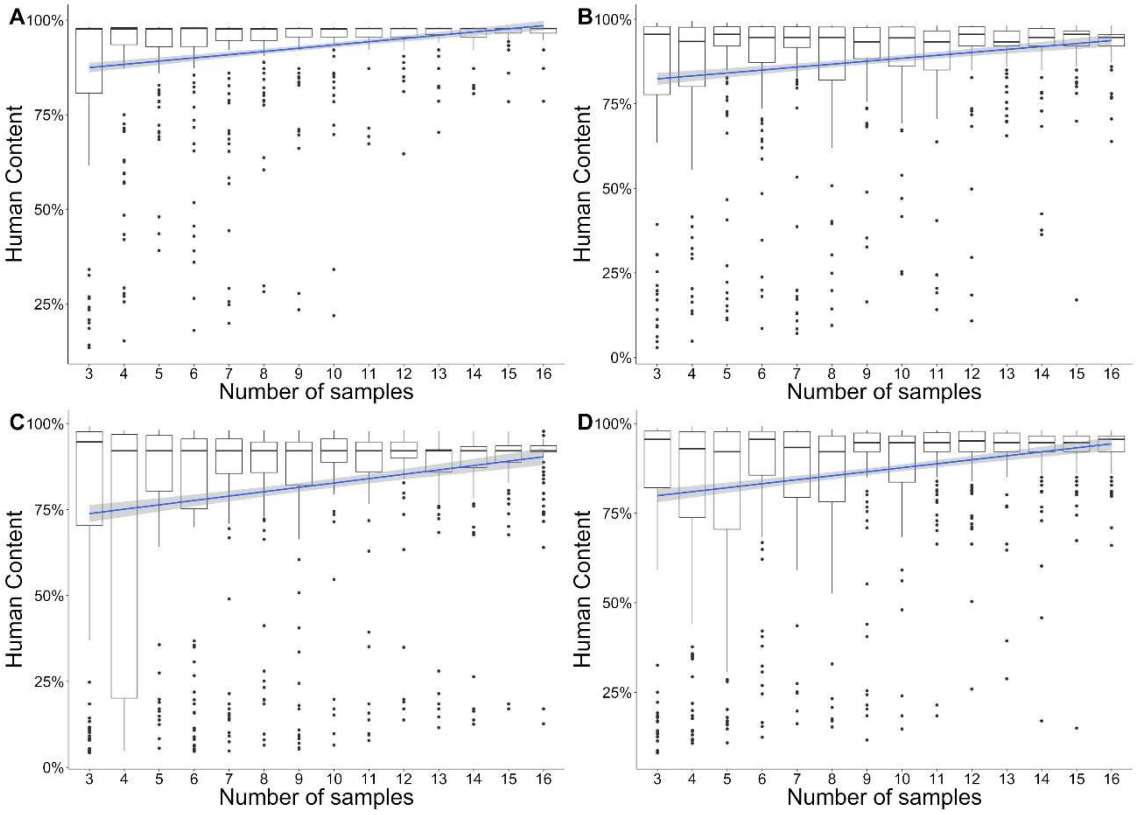


The accuracy of the machine learning classification models using only the bacteriophage ratio as a predictor variable was negligible for secondary (6.36 %) and tertiary effluent (2.43 %) samples. When inactivation was also included, accuracy increased to 16.79 % for secondary and 6.93 % for tertiary effluents. However, when both correlation and ratio were used as predictor variables, the accuracy increased dramatically to 95.36 % and 87.14 % for secondary and tertiary effluents, respectively, and increased still further to 97.71 % and 90.29 % when inactivation was included. Classification accuracy differed significantly between secondary and tertiary effluents (Wilcoxon test, p-value ≤ 0.05). The results of each classification model with the numbers of all tested samples are shown in Table 4.

Similar results to those of the machine learning classification model were obtained when using the mixed effect regression model, which provided a mean accuracy of 97.43 % for real samples from the treated secondary effluents. However, the regression model failed when applied to tertiary effluents, providing a mean accuracy of 0.64 %. When bacteriophage inactivation was taken into account, accuracy increased slightly to 99.57 % for secondary effluents and 3.29 % for tertiary effluents.

588 The machine learning regression model predicted a mean percentage of human pollution
589 in secondary effluents of 93.02 %, with a 95% confidence interval (defined as the
590 interval between the 2.5th % and 97.5th % percentiles obtained in the simulation) from
591 46.33 % to 98.11 %. It should be noted that the fecal pollution in all urban wastewater
592 effluent samples was considered to be 100 % of human origin, and under this
593 assumption neither R^2 nor a statistical test were carried out to compare both results.
594 When inactivation was taken into account, the mean content of human pollution
595 increased to 94.38 % (ranging from 62.11 % to 98.11 %). The correction did not
596 produce any significant differences in predictions for secondary effluents or in the
597 results associated with the sample number within each group (Kruskal-Wallis test, *p*-
598 value > 0.05, df = 13).
599 When the prediction model was applied to tertiary effluents, the mean value of human
600 pollution content was 81.23 % (9.00 % to 98.00 %), which underwent a slight but
601 significant increase to 87.53 % (17.15 % to 98.11 %) when inactivation was included
602 (Wilcoxon test, *p*-value ≤ 0.05). However, when the simulated samples with different
603 samples sizes were compared, no significant differences were observed (Kruskal-Wallis
604 test, *p*-value ≥ 0.05, df = 13).
605 For secondary and tertiary effluents, the fraction of misclassified or poorly predicted
606 samples decreased as the number of replicas increased. In tertiary effluents, statistical
607 differences were observed between fractions containing ≤ 9 replicas and those
608 containing 16 replicas (*p*-value ≤ 0.05, test of equal or given proportions). The results
609 are shown in Figure 3 as boxplots.
610
611 Figure 3. Boxplots of the 100 simulations for each number of samples. A and C show
612 the predictions of human content for secondary and tertiary effluents, respectively. B

and D show the predictions of human content for secondary and tertiary effluents, respectively, taking into account inactivation. Dots state for outlier predictions, blue line refers to the lineal regression of the human content vs the number of samples for which the confidence interval is represented in gray.



3.8.2 Natural inactivation in river water matrices

Models not using correlation as a predictor variable were excluded because of their poor classification performance. The remaining models were applied to the Llobregat River and Riudaura Stream. The values of SOMCPH and GA17PH detected in the human-polluted Llobregat River were at least 18-fold higher than those of secondary treated wastewaters; this suggests that the main source of microbial pollution may be incompletely treated urban wastewaters with a minimal contribution of farming. For contrast, three transects of the Riudaura Stream were used as models of a river with a low level of human fecal pollution.

For SOMCPH, the parameters *a* and *b* of the power function have respective values of -0.0227 and 0.6320 for winter and 0.0227 and 0.8443 for summer. Parameter *b* was defined as the random parameter. The adjusted R^2 for the overall model was 81.03 %, with an MAE of 0.68 and an RMSE of 0.82. For GA17PH, the parameters *a*, *b* and *c* of the Gompertz relation were, respectively, -0.8485, 1.2490 and 0.0280 for winter and -2.4136, 1.2490 and 0.0280 for summer. Parameter *a* was defined as the random parameter. The adjusted R^2 for the overall model was 96.31 % with an MAE of 0.12 and an RMSE of 0.15.

When the machine learning model was applied, a mean classification accuracy of 63.89 % (minimum of 51 % and maximum of 76 %) was achieved for the Llobregat River samples, whereas for the Riudaura Stream the values increased to 96.27 % (79 % to 100 %).

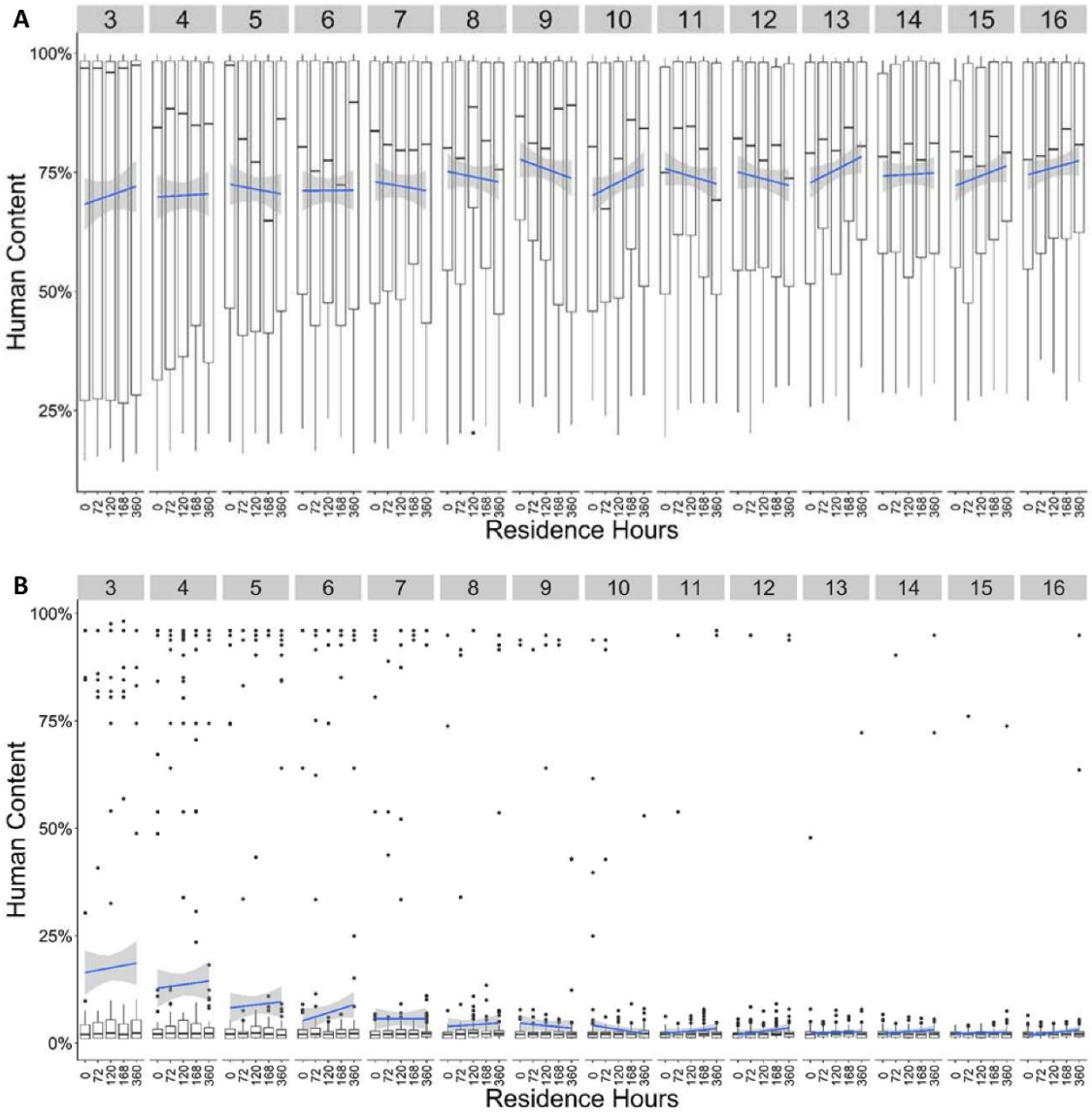
In the prediction for the Llobregat River, it was observed that the percentage of samples classified within the human-class was unaffected by the number of samples, but a higher percentage were misclassified. The percentage of samples identified as non-human decreased as the number of grouped samples increased, falling below 6 % when the sample number was greater than 14.

When the machine learning regression model was applied, statistically significant differences were observed in accuracy for both rivers according to the number of samples considered in the resampling (Supplementary-Table 2), whether or not the correction was applied (Kruskal-Wallis test, p-value ≤ 0.05, df = 13). The application of the environmental decay factor did not significantly alter the results for the Llobregat River (Wilcoxon test, p-value > 0.05), whereas differences were observed in the Riudaura Stream at aging times beyond 120 hours (Wilcoxon test, p-value ≤ 0.05).

When the regression model was applied without inactivation, the Llobregat River

showed a mean human content of 73.70 % (23.78 % to 99.22 %), which increased to 73.96 % (23.78 % to 99.22 %) after taking into account natural decay (residence time of 360 hours). The equivalent values for the Riudaura Stream were 5.48 % (1.00 % to 7.12 %), which decreased to 5.43 % (1.00 % to 7.00 %). The results obtained when correcting for natural inactivation during the summer season are shown in Figure 4.

Figure 4. Boxplots of the 100 simulations for each number of replicas. A and B respectively show the predictions of human content for the Llobregat River and the Riudaura Stream with the summer corrections. Dots state for outlier predictions, blue line refers to the lineal regression of the human content vs the number of samples for which the confidence interval is represented in gray.



664

665

666 **4. Discussion**

667 Among the four models evaluated in this study, those including correlations were more
668 robust in predicting the human pollution content in real samples from water matrices
669 where markers may be submitted to artificial or natural decay. This result attests to the
670 importance of taking into account parameters often missed in MST studies: the
671 minimum sampling size of the water body necessary to obtain a statistically significant
672 result, the dilution effect and natural or artificial differential inactivation of the MST
673 markers.

The use of a reduced set of MST markers clearly has some limitations, i.e., the presence of GA17PH in water samples is an indicator of human fecal pollution, but its absence does not imply an animal source, especially when the concentration of SOMCPH is low. GA17PH is the limiting parameter in the tested models, as its concentration is always lower than that of SOMCPH (Moce-Llivina *et al.*, 2005; Muniesa *et al.*, 2012). In such cases, it may therefore be necessary to use additional bacteriophages for predicting specific animal fecal pollution, which would help to accurately determine the human content of the samples.

Concerning the dilution of MST markers, *a priori* the dilution of wastewaters containing human- and non-human-associated phages should not lead to differences in the logarithm ratio, as long as the post-dilution values of GA17PH are higher than its detection limit. However, the use of predictive models in samples that have received a highly effective treatment might be affected by the low values of SOMCPH and GA17PH and the inversion of the log10 inactivation between bacteriophages, as occurred in the tertiary effluents. In such a case, the inferred model would result in a distorted "humanization" of the fecal pollution. This was observed in the classification by a machine learning predictive model trained with the ratio and correlation when applied to a river with negligible human pollution and low levels of both MST markers. In fact, tertiary effluents and samples from the Riudaura Stream presented dilution/inactivation levels beyond the maximum dilution fold threshold calculated according to the distribution function of GA17PH in urban raw wastewaters.

Under these circumstances, it would be important to lower the limit of detection of the method by using a concentration methodology prior to bacteriophage detection and to amend the results according to the efficiency of the concentration method.

An interesting point is that the machine learning regression model was the only model able to deal with this limitation, satisfactorily predicting all real samples, although the predictions improved as the number of replicas increased. This result supports the need to determine the minimum sample size according to the variability of the MST markers. It also shows that the sample size could be reduced depending on the prediction method, potentially a significant factor when several time-consuming or expensive MST markers are being used.

In summary, the development of predictive models for MST can be influenced by the choice of markers. To ensure an optimum model performance, when two or more markers are selected, they should ideally exhibit similar characteristics and environmental behavior. The concentration of the marker should also be taken into account in order to assess the dilution threshold below which the marker can no longer be detected due to method limitations, environmental decay and inactivation by disinfection treatments. Furthermore, the degree of certainty of the method used to quantify the different markers should be assessed. In this study, the main limitation of the bacteriophage quantification method (double-agar layer plaque assay) concerns the analyzed samples, which represent only part of the water matrix. Moreover, the distribution of the markers in the matrix is likely to be irregular, due to processes such as adsorption and particularization.

5. Conclusions

In this work, the fraction of human fecal pollution in two Mediterranean rivers in north-eastern Spain, which are subjected to different sources of pollution, was predicted using a set of only two microbial parameters and their correlation. Based on the results, the

minimum advisable concentrations of SOMCPH and GA17PH are ≥ 500 PFU/100 mL and 10 PFU/100 mL, respectively.

Although the selection of an appropriate marker is important to correctly predict the human contribution in a fecal point source pollution event, this study reveals the importance of other parameters missed in the majority of MST studies: the minimum sampling size of the water body necessary to obtain a statistically significant result, and marker dilution and inactivation.

The proposed classification procedure involves the following steps:

a) Characterization of marker variability in the point source fecal pollution. This was achieved by fitting the variables to their probability distribution functions, but elementary factors should be taken into account, such as the number of samples according to marker variability, accuracy of the detection method, and any concerns about recovery and imperfect detection procedures used in the quantification of MST variables.

b) Generation of several wastewater mixing models under potential scenarios arising from the possible MST probability distribution functions and different mixes of point source fecal pollution. The best models should be selected based on their goodness-of-fit criterion.

c) Taking into account effects such as dilution and differential/natural inactivation that could modify the results of the classification procedure.

d) Establishment of a sampling plan for the target water body, which depends on the variability of the markers it contains.

e) Additional assessments of the dilution effects or differential inactivation of the markers in the water body should be considered.

We believe this approach to model construction could be used with other markers reported in the literature, and that the model might be improved by including animal-specific markers (e.g. from housed animals).

Conflicts of interest

No conflicts of interest to declare.

Acknowledgements

The authors thank Joan Jofre for the suggestions in the research and preparation of the publication.

Funding

This work was supported by the grant agreement No.: 311846 from the Spanish government, Spain (research projects CGL2014-59977-C3-3-R and AGL2016-75536) and the by Catalan government, Spain (2017 SGR 170).

References

Balleste,E., Belanche-Munoz,L.A., Farnleitner,A.H., Linke,R., Sommer,R., Santos,R. et al. (2020) Improving the identification of the source of fecal pollution in water using a modelling approach: From multi-source to aged and diluted samples. Water Res 171: 115392.

Ballesté, E., Blanch, A. R., Mendez, J., Sala-Comorera, L., Maunula, L., Monteiro, S., Farnleitner, A. H., Tiehm, A., Jofre, J., & García-Aljaro, C. (2021). Bacteriophages Are Good Estimators of Human Viruses Present in Water. Frontiers in microbiology, 12, 619495. https://doi.org/10.3389/fmicb.2021.619495

770 Belanche-Muñoz, L., Blanch, A.R., 2008. Machine learning methods for
771 microbial source tracking. Environ. Model. Softw. 23, 741–750.
772 doi:10.1016/j.envsoft.2007.09.013
773 Blanch A.R., Ballesté E, Weidhaas J, Santo Domingo J, Ryu H. 2016. Methods
774 of Targeting Animal Sources of Fecal Pollution in Water, p 3.4.4-1-3.4.4-28. In Yates
775 M, Nakatsu C, Miller R, Pillai S (ed), Manual of Environmental Microbiology, Fourth
776 Edition. ASM Press, Washington, DC. doi: 10.1128/9781555818821.ch3.4.4
777 Blanch, A.R., Belanche-Munoz, L., Bonjoch, X., Ebdon, J., Gantzer, C., Lucena,
778 F., Ottoson, J., Kourtis, C., Iversen, A., Kuhn, I., Moce, L., Muniesa, M., Schwartzbrod,
779 J., Skraber, S., Papageorgiou, G.T., Taylor, H., Wallis, J., Jofre, J., 2006. Integrated
780 analysis of established and novel microbial and chemical methods for microbial source
781 tracking. Appl Environ Microbiol. 72, 5915-5926.
782 Caldwell, J.M., Raley, M.E., Levine, J.F., 2007. Mitochondrial multiplex real-
783 time PCR as a source tracking method in fecal-contaminated effluents. Environ Sci
784 Technol. 41, 3277-3283.
785 da Silva, A.R.; Malafaia, G.; Menezes, I.P.P. (2017) biotools: an R function to
786 predict spatial gene diversity via an individual-based approach. Genetics and Molecular
787 Research, 16: gmr16029655.
788 Delignette-Muller, M.L., Pouillot, R., Denis, J.B., Dutang, C., 2010. Fitdistrplus:
789 Help to fit of a parametric distribution to non-censored or censored data, R package
790 version 0.1-3, URL http://CRAN. R-project. org/package= fitdistrplus
791 Dubinsky, E.A., Esmaili, L., Hulls, J.R., Cao, Y., Griffith, J.F., Andersen, G.L.,
792 2012. Application of phylogenetic microarray analysis to discriminate sources of fecal
793 pollution. Environ. Sci. Technol. 46, 4340-4347.

794 Durán, A.E., Muniesa, M., Méndez, X., Valero, F., Lucena, F., Jofre, J., 2002.
795 Removal and inactivation of indicator bacteriophages in fresh waters. J. Appl.
796 Microbiol. 92, 338–347. doi:10.1046/j.1365-2672.2002.01536.x'
797 Ebdon, J., Muniesa, M., Taylor, H., 2007. The application of a recently isolated
798 strain of Bacteroides (GB-124) to identify human sources of fecal pollution in a
799 temperate river catchment. Water Res. 41, 3683-3690.
800 Elzhov, T. V., Mullen, K. M., Spiess, A.-N. and Bolker, B. (2016). minpack.lm:
801 R Interface to the Levenberg-Marquardt Nonlinear Least-Squares Algorithm Found in
802 MINPACK, Plus Support for Bounds. R package version 1.2-1. https://CRAN.R-
803 project.org/package=minpack.lm
804 García-Aljaro, C., Martín-Díaz, J., Viñas-Balada, E., Calero-Cáceres, W.,
805 Lucena, F., Blanch, A.R., 2017. Mobilisation of microbial indicators, microbial source
806 tracking markers and pathogens after rainfall events. Water Res.
807 doi:10.1016/j.watres.2017.02.003
808 Gomez-Donate, M., Balleste, E., Muniesa, M., Blanch, A.R., 2012. New
809 molecular quantitative PCR assay for detection of host-specific Bifidobacteriaceae
810 suitable for microbial source tracking. Appl. Environ. Microbiol. 78, 5788-5795.
811 Gomez-Doñate, M., Payan, A., Cortes, I., Blanch, A.R., Lucena, F., Jofre, J.,
812 Muniesa, M., 2011. Isolation of bacteriophage host strains of Bacteroides species
813 suitable for tracking sources of animal fecal pollution in water. Environ. Microbiol. 13,
814 1622-1631.
815 Gomez-Doñate, M., Casanovas-Massana, A., Muniesa, M., Blanch, A.R., 2016.
816 Development of new host-specific Bacteroides qPCRs for the identification of fecal
817 contamination sources in water. Microbiology Open. Jan 14. doi: 10.1002/mbo3.313.

818 Grothendieck, G., 2013. nls2: Non-linear regression with brute force. R package
819 version, 2.
820 Grubbs, F.E., 1950. Sample criteria for testing outlying observations. Ann. Math. Stat.
821 21
822 Harwood, Valerie J. and Hodon, Ryu and Santo Domingo, Jorge. 2011.
823 Microbial Source Tracking, p 189-216. In Sadowsky, Michael J. and Whitman, Richard
824 L.(ed), The Fecal Bacteria. doi:10.1128/9781555816865.ch9
825 Harwood, V., Shanks, O., Koraijkic, A., Verbyla, M., Ahmed, W. and Iriate, M.
826 2017.General and host-associated bacterial indicators of fecal pollution. In: J.B. Rose
827 and B. Jiménez-Cisneros, (eds) Global Water Pathogen Project.
828 http://www.waterpathogens.org (A.Farnleitner, and A. Blanch (eds) Part 2 Indicators
829 and Microbial Source Tracking Markers)
830 http://www.waterpathogens.org/book/bacterial-indicators Michigan State University, E.
831 Lansing, MI, UNESCO. https://doi.org/10.14321/waterpathogens.6
832 Hervé, M. (2019). RVAideMemoire: Testing and Plotting Procedures for
833 Biostatistics. R package version 0.9-73. https://CRAN.R-
834 project.org/package=RVAideMemoire
835 ISO, 2000. ISO 10705-2: Water quality. Detection and enumeration of
836 bacteriophages -part 2: Enumeration of somatic coliphages. Geneva, Switzerland:
837 International Organisation for Standardisation.
838 ISO, 2001. ISO 10705-4: Water quality. Detection and enumeration of
839 bacteriophages - Part 4: Enumeration of bacteriophages infecting Bacteroides fragilis.
840 Geneva, Switzerland: International Organisation for Standardisation.
841 Jebri, S., Muniesa, M. and Jofre, J. 2017. General and host-associated
842 bacteriophage indicators of fecal pollution. In: J.B. Rose and B. Jiménez-Cisneros, (eds)

843 Global Water Pathogen Project. http://www.waterpathogens.org (A.Farnleitner, and A.
844 Blanch (eds) Part 2 Indicators and Microbial Source Tracking Markers)
845 http://www.waterpathogens.org/book/coliphage Michigan State University, E. Lansing,
846 MI, UNESCO. https://doi.org/10.14321/waterpathogens.7
847 Jofre, J., Blanch, A.R., Lucena, F., Muniesa, M., 2014. Bacteriophages infecting
848 Bacteroides as a marker for microbial source tracking. Water Res. 55, 1-11.
849 Köck-Schulmeyer, M., Ginebreda, A., Postigo, C., López-Serna, R., Pérez, S.,
850 Brix, R., Llorca, M., de Alda, M.L., Petrović, M., Munné, A., Tirapu, L., Barceló, D.,
851 2011. Wastewater reuse in Mediterranean semi-arid areas: The impact of discharges of
852 tertiary treated sewage on the load of polar micro pollutants in the Llobregat River (NE
853 Spain). Chemosphere 82, 670–8
854 Kuhn, M.; Contributions from Wing, J., Weston, S., Williams, A., Keefer, C.,
855 Engelhardt, A., Cooper, T., Mayer, Z., Kenkel, B., the R Core Team, Benesty, M.,
856 Lescarbeau, R., Ziem, A., Scrucca, L., Tang, Y., Candan C., and Hunt, T. (2019). caret:
857 Classification and Regression Training. R package version 6.0-82. https://CRAN.R-
858 project.org/package=caret
859 Lee, J.E., Lee, S., Sung, J., Ko, G., 2011. Analysis of human and animal fecal
860 microbiota for microbial source tracking. ISME J. 5, 362-365.
861 Lee, J.E., Lim, M.Y., Kim, S.Y., Lee, S., Lee, H., Oh, H.M., Hur, H.G., Ko, G.,
862 2009. Molecular characterization of bacteriophages for microbial source tracking in
863 Korea. Appl. Environ. Microbiol. 75, 7107-7114.
864 Mauffret, A., Mieszkin, S., Morizur, M., Alfiansah, Y., Lozach, S., Gourmelon,
865 M., 2013. Recent innovation in microbial source tracking using bacterial real-time PCR
866 markers in shellfish. Mar. Pollut. Bull. 68, 21-29.

McMinn, B.R., Korajkic, A., Ashbolt, N.J., 2014. Evaluation of Bacteroides fragilis GB-124 bacteriophages as novel human-associated fecal indicators in the United States. Lett Appl. Microbiol. 59, 115-121.

Moce-Llivina, L., Lucena, F., Jofre, J., 2005. Enteroviruses and bacteriophages in bathing waters. Appl. Environ. Microbiol. 71, 6838-6844.

Muniesa, M., Lucena, F., Blanch, A.R., Payan, A., Jofre, J., 2012. Use of abundance ratios of somatic coliphages and bacteriophages of Bacteroides thetaiotaomicron GA17 for microbial source identification. Water Res. 46, 6410-6418.

Noble, R.T., Allen, S.M., Blackwood, A.D., Chu, W., Jiang, S.C., Lovelace, G.L., Sobsey, M.D., Stewart, J.R., Wait, D.A., 2003. Use of viral pathogens and indicators to differentiate between human and non-human fecal pollution in a microbial source tracking comparison study. J. Water Health. 1, 195-207.

Payan, A., Ebdon, J., Taylor, H., Gantzer, C., Ottoson, J., Papageorgiou, G.T., Blanch, A.R., Lucena, F., Jofre, J., Muniesa, M., 2005. Method for isolation of Bacteroides bacteriophage host strains suitable for tracking sources of fecal pollution in water. Appl. Environ. Microbiol. 71, 5659-5662.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D. (2019). R Core Team. nlme: Linear and Nonlinear Mixed Effects Models. R package version 3.1-139, https://CRAN.R-project.org/package=nlme.

Pouillot, R., Delignette-Muller, M.L., 2010. Evaluating variability and uncertainty separately in microbial quantitative risk assessment using two R packages. Int. J. Food Microbiol. 142, 330-340.

Puig, A., Queralt, N., Jofre, J., Araujo, R., 1999. Diversity of Bacteroides fragilis strains in their capacity to recover phages from human and animal wastes and from fecally polluted wastewater. Appl. Environ. Microbiol. 65, 1772-1776.

R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Sekhon, J.S., 2011. Multivariate and propensity score matching software with automated balance optimization: the matching package for R. Journal of Statistical Software, Forthcoming.

Shanks O, Green H, Korajkic A, Field K. 2016. Overview of Microbial Source Tracking Methods Targeting Human Fecal Pollution Sources, p 3.4.3-1-3.4.3-8. In Yates M, Nakatsu C, Miller R, Pillai S (ed), Manual of Environmental Microbiology, Fourth Edition. ASM Press, Washington, DC. doi: 10.1128/9781555818821.ch3.4.3

Stapleton, C.M., Wyer, M.D., Kay, D., Crowther, J., McDonald, A.T., Walters, M., Gawler, A., Hindle, T., 2007. Microbial source tracking: a forensic technique for microbial source identification? J. Environ. Monit. 9, 427-439.

Tartera, C., Lucena, F., Jofre, J., 1989. Human origin of Bacteroides fragilis bacteriophages present in the environment. Appl. Environ. Microbiol. 55, 2696-2701.

Venegas, C., Diez, H., Blanch, A.R., Jofre J., Campos, C., 2015. Microbial source markers assessment in the Bogotá River basin (Colombia). J Water and Health 13, 801-809.

Yahya, M., Hmaied, F., Jebri, S., Jofre, J., Hamdi, M., 2015. Bacteriophages as indicators of human and animal fecal contamination in raw and treated wastewaters from Tunissia. J Appl Microbiol 118, 1271-1225.

Wu., H., A. Godfrey, J. R., Govindaraju, K. and Pirikahu S., (2015). ExtDist: Extending the Range of Functions for Probability Distributions. R package version 0.6-3. https://CRAN.R-project.org/package=ExtDist

Table 1. Descriptive statistics of the samples used to obtain probability function distributions applied in Monte Carlo modeling and to determine bacteriophage inactivation by wastewater treatments. Values expressed as PFU / 100 mL.

	Urban wastewaters		Abattoir sewages		Secondary effluents		Tertiary effluents	
	SOMCPH	**GA17PH**	**SOMCPH**	**GA17PH**	**SOMCPH**	**GA17PH**	**SOMCPH**	**GA17PH**
n	41	41	31	31	108	108	65	65
Mean	$4.01 \cdot 10^{6}$	$8.33 \cdot 10^{4}$	$8.68 \cdot 10^{6}$	0	$1.31 \cdot 10^{6}$	$8.76 \cdot 10^{3}$	$2.68 \cdot 10^{4}$	$5.96 \cdot 10^{2}$
Median	$4.25 \cdot 10^{6}$	$8.95 \cdot 10^{4}$	$5.50 \cdot 10^{6}$	0	$5.00 \cdot 10^{5}$	$2.16 \cdot 10^{3}$	$3.80 \cdot 10^{2}$	$1.70 \cdot 10^{1}$
Sd	$2.42 \cdot 10^{6}$	$4.88 \cdot 10^{4}$	$9.01 \cdot 10^{6}$	-	$3.01 \cdot 10^{6}$	$1.39 \cdot 10^{4}$	$9.20 \cdot 10^{4}$	$1.83 \cdot 10^{3}$
2.5th perc.	$1.00 \cdot 10^{6}$	$8.00 \cdot 10^{3}$	$6.70 \cdot 10^{4}$	0	$2.35 \cdot 10^{3}$	$5.84 \cdot 10^{1}$	$2.45 \cdot 10^{0}$	$2.11 \cdot 10^{0}$
97.5th perc.	$9.00 \cdot 10^{6}$	$1.59 \cdot 10^{5}$	$2.77 \cdot 10^{7}$	0	$4.44 \cdot 10^{6}$	$4.86 \cdot 10^{4}$	$1.61 \cdot 10^{5}$	$5.08 \cdot 10^{3}$
rho	0.51		-		0.75		0.82	

922 Table 2. Descriptive statistics of the samples of known origin used to evaluate the

923 predictive models. Values expressed as PFU / 100 mL.

924

	Secondary effluents		Tertiary effluents		Llobregat River		Riudaura Stream	
	SOMCPH	**GA17PH**	**SOMCPH**	**GA17PH**	**SOMCPH**	**GA17PH**	**SOMCPH**	**GA17PH**
n	54	54	36	36	14	14	81	81
Mean	$2.29 \cdot 10^{6}$	$1.67 \cdot 10^{4}$	$2.79 \cdot 10^{3}$	$1.63 \cdot 10^{2}$	$7.88 \cdot 10^{3}$	$6.65 \cdot 10^{2}$	$9.75 \cdot 10^{3}$	$5.76 \cdot 10^{1}$
Median	$1.30 \cdot 10^{6}$	$1.19 \cdot 10^{4}$	$6.75 \cdot 10^{2}$	$4.35 \cdot 10^{1}$	$6.35 \cdot 10^{3}$	$5.90 \cdot 10^{2}$	$9.45 \cdot 10^{3}$	$3.60 \cdot 10^{1}$
Sd	$4.01 \cdot 10^{6}$	$1.76 \cdot 10^{4}$	$3.94 \cdot 10^{3}$	$3.51 \cdot 10^{2}$	$4.60 \cdot 10^{3}$	$5.03 \cdot 10^{2}$	$7.07 \cdot 10^{3}$	$7.46 \cdot 10^{1}$
2.5th perc.	$2.93 \cdot 10^{4}$	$1.83 \cdot 10^{2}$	$5.40 \cdot 10^{1}$	$1.00 \cdot 10^{1}$	$3.53 \cdot 10^{3}$	$1.78 \cdot 10^{2}$	$2.10 \cdot 10^{3}$	$4.00 \cdot 10^{0}$
97.5th perc.	$1.59 \cdot 10^{7}$	$6.15 \cdot 10^{4}$	$1.36 \cdot 10^{4}$	$1.42 \cdot 10^{3}$	$1.79 \cdot 10^{4}$	$1.74 \cdot 10^{3}$	$3.10 \cdot 10^{4}$	$3.20 \cdot 10^{2}$
rho	0.75		0.66		0.34		0.15	

925

926

927 Table 3. Best parameters for the probability distribution function fitting. SOMCPH:

928 somatic coliphages. GA17PH: human-host specific bacteriophages that infect

929 *Bacteroides thetaiotaomicron* strain GA17

		Function	Parameters	AIC
Human	SOMCPH	Gamma	shape = $2.816 \cdot 10^{0}$; rate = $7.021 \cdot 10^{-7}$	1322.11
	GA17PH	Normal	mean = $8.334 \cdot 10^{4}$; sd = $4.848 \cdot 10^{4}$	1005.00
Non-human	SOMCPH	Gamma	shape = $9.587 \cdot 10^{-1}$; rate = $1.105 \cdot 10^{-7}$	1055.74
	GA17PH	Constant	value = 10	-

930

Table 4. Classification accuracy after testing models with secondary and tertiary effluents. Regression states for the mixed effect regression model. KNN (ratio) denotes classification with the KNN algorithm using the ratio. KNN (ratio + cor.) states for classification with the KNN algorithm using the ratio and correlation as predictor variables. Samples represent the number of samples used in the analysis. Raw and Inac. respectively represent the classification of samples not including and including the effect of bacteriophage inactivation.

	Secondary						Tertiary					
	KNN (ratio)		KNN (ratio + cor.)		Regression		KNN (ratio)		KNN (ratio + cor.)		Regression	
Samples	**Raw**	**Inac.**	**Raw**	**Inac.**	**Raw**	**Inac.**	**Raw**	**Inac.**	**Raw**	**Inac.**	**Raw**	**Inac.**
3	0.24	0.29	0.90	0.94	0.90	0.96	0.08	0.27	0.77	0.77	0.07	0.16
4	0.12	0.28	0.88	0.90	0.89	1.00	0.08	0.19	0.76	0.77	0.01	0.13
5	0.20	0.21	0.86	0.95	0.98	0.99	0.06	0.12	0.82	0.79	0.01	0.07
6	0.07	0.28	0.93	0.97	0.95	0.99	0.00	0.09	0.80	0.91	0.00	0.02
7	0.07	0.16	0.94	0.97	0.96	1.00	0.05	0.06	0.81	0.85	0.00	0.01
8	0.03	0.16	0.93	0.97	0.96	1.00	0.03	0.03	0.81	0.90	0.00	0.03
9	0.06	0.20	0.96	1.00	1.00	1.00	0.00	0.08	0.88	0.91	0.00	0.04
10	0.01	0.19	0.98	1.00	1.00	1.00	0.03	0.05	0.86	0.95	0.00	0.00
11	0.01	0.12	0.99	0.98	1.00	1.00	0.01	0.03	0.94	0.93	0.00	0.00
12	0.01	0.15	0.99	1.00	1.00	1.00	0.00	0.01	0.92	0.95	0.00	0.00
13	0.03	0.10	1.00	1.00	1.00	1.00	0.00	0.01	0.95	0.95	0.00	0.00
14	0.02	0.09	1.00	1.00	1.00	1.00	0.00	0.01	0.96	0.96	0.00	0.00
15	0.02	0.08	0.99	1.00	1.00	1.00	0.00	0.02	0.93	1.00	0.00	0.00
16	0.00	0.04	1.00	1.00	1.00	1.00	0.00	0.00	0.99	1.00	0.00	0.00
***p*-value**	≤ 0.05		≤ 0.05		≤ 0.05		≤ 0.05		≤ 0.05		≤ 0.05	

940 Supplementary-Table 1. Functions, R^2 and AIC of the tested regression models between
941 somatic coliphages and human-host specific bacteriophages infecting *B.*
942 *thetaiotaomicron* strain GA17. The first row shows a linear regression, *exp* denotes the
943 exponential function and *log* denotes the natural logarithm function.

Function	R^2	AIC
$a \cdot x + b$	0.675	84404.71
$(a + b \cdot x)/(1 + c \cdot x + d \cdot x^2)$	0.858	76099.87
$(a \cdot b + c \cdot x^d)/(b + x^d)$	0.857	76188.44
$1/(a + b \cdot x^c)$	0.856	76242.76
$\exp(a + b/x + c \cdot \log(x))$	0.856	76260.06
$a+(b - a) \cdot (1 - \exp(-\exp(c \cdot (\log(x)-\log(d)))))$	0.855	76332.41
$a \cdot \exp(-(x - b)^2/(2 \cdot c^2))$	0.854	76383.94
$(a \cdot x^3 + b \cdot x^2 + c \cdot x + d)$	0.851	76579.72
$a + b \cdot x + c/x^2$	0.849	76743.56
$a \cdot x^{(b \cdot x)}$	0.848	76827.12
$a + (b - c) \cdot (1 - \exp(-x/d))$	0.846	76912.27
$a \cdot \exp(b \cdot x)$	0.843	77106.57
$a \cdot \exp(b/x)$	0.843	77106.57
$(a + b \cdot x)^{(-1/c)}$	0.843	77108.65
$a \cdot b^x \cdot x^c$	0.839	77369.88
$a \cdot x^2 + b \cdot x + c$	0.837	77487.28
$a \cdot (x - b)^c$	0.837	77499.57
$a \cdot x^b$	0.835	77641.62

944

945

Supplementary-Table 2. Classification accuracy after testing models with river matrices with an estimated aging from 0 to 360 hours. Only the classification achieved with the KNN algorithm using the ratio and correlation as predictor variables is shown. (A) (top) shows the results for the Llobregat River and (B) (bottom) the results for the Riudaura Stream. Samples represent the number of samples used in the analysis.

A	Winter					Summer				
Samples	**0**	**72**	**120**	**168**	**360**	**0**	**72**	**120**	**168**	**360**
3	0.72	0.64	0.70	0.72	0.73	0.70	0.67	0.69	0.64	0.65
4	0.65	0.66	0.64	0.65	0.60	0.64	0.61	0.58	0.65	0.62
5	0.62	0.58	0.56	0.69	0.66	0.60	0.61	0.57	0.63	0.51
6	0.61	0.64	0.65	0.71	0.59	0.60	0.52	0.57	0.62	0.63
7	0.56	0.65	0.62	0.56	0.59	0.53	0.66	0.57	0.68	0.58
8	0.59	0.57	0.57	0.66	0.65	0.72	0.62	0.63	0.58	0.55
9	0.62	0.73	0.54	0.69	0.63	0.59	0.64	0.67	0.61	0.62
10	0.57	0.64	0.69	0.60	0.58	0.60	0.68	0.63	0.62	0.69
11	0.62	0.59	0.61	0.64	0.56	0.68	0.56	0.68	0.62	0.62
12	0.64	0.67	0.59	0.65	0.59	0.69	0.67	0.64	0.63	0.66
13	0.64	0.62	0.62	0.57	0.67	0.64	0.60	0.67	0.60	0.70
14	0.70	0.70	0.68	0.54	0.71	0.71	0.66	0.64	0.72	0.67
15	0.74	0.71	0.68	0.62	0.63	0.72	0.74	0.61	0.72	0.70
16	0.76	0.67	0.69	0.73	0.66	0.63	0.67	0.69	0.66	0.64

B	Winter					Summer				
Samples	**0**	**72**	**120**	**168**	**360**	**0**	**72**	**120**	**168**	**360**
3	0.79	0.87	0.81	0.93	0.91	0.81	0.84	0.79	0.80	0.84
4	0.94	0.87	0.84	0.89	0.91	0.92	0.88	0.89	0.87	0.87
5	0.90	0.94	0.92	0.96	0.92	0.90	0.91	0.92	0.92	0.95
6	0.95	0.94	0.95	0.90	0.92	0.97	0.97	0.94	0.95	0.97
7	0.99	0.97	0.96	0.96	0.96	0.96	0.97	0.99	0.97	0.98
8	0.96	0.99	0.96	0.97	0.98	0.98	0.99	0.94	0.95	1.00
9	0.96	1.00	0.99	1.00	1.00	1.00	1.00	0.98	0.98	0.99
10	1.00	0.96	0.98	1.00	0.98	0.98	0.98	1.00	0.99	0.98
11	0.99	0.98	0.99	0.97	1.00	1.00	0.97	1.00	0.98	0.98
12	1.00	0.99	0.99	0.99	1.00	1.00	1.00	0.99	1.00	1.00
13	0.99	0.99	1.00	1.00	1.00	0.98	0.99	0.99	1.00	0.99
14	0.99	1.00	0.98	1.00	1.00	1.00	1.00	1.00	0.98	1.00
15	0.98	1.00	0.99	0.99	0.99	1.00	1.00	1.00	1.00	1.00
16	1.00	1.00	1.00	0.99	1.00	1.00	1.00	1.00	1.00	0.99