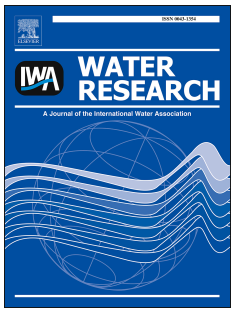
Journal Pre-proof



Improving the identification of the source of faecal pollution in water using a modelling approach: From multi-source to aged and diluted samples

Elisenda Ballesté, Luis Antonio Belanche, Andreas H. Farnleitner, Rita Linke, Regina Sommer, Ricardo Santos, Silvia Monteiro, Leena Maunula, Satu Oristo, Andreas Tiehm A, Claudia Stange, Anicet R. Blanch

PII: S0043-1354(19)31166-2

DOI: https://doi.org/10.1016/j.watres.2019.115392

Reference: WR 115392

To appear in: *Water Research*

Received Date: 1 August 2019

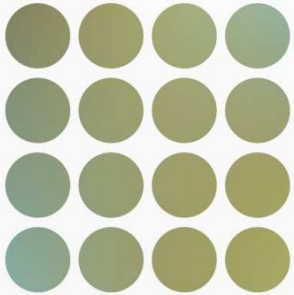
Revised Date: 9 December 2019

Accepted Date: 11 December 2019

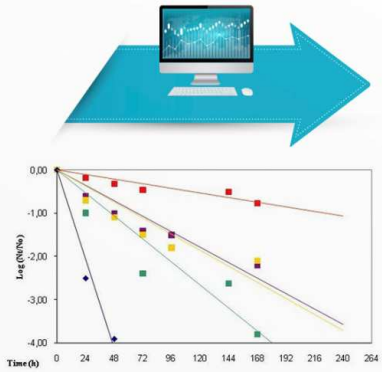
Please cite this article as: Ballesté, E., Belanche, L.A., Farnleitner, A.H., Linke, R., Sommer, R., Santos, R., Monteiro, S., Maunula, L., Oristo, S., Tiehm A, A., Stange, C., Blanch, A.R., Improving the identification of the source of faecal pollution in water using a modelling approach: From multi-source to aged and diluted samples, *Water Research* (2020), doi: https://doi.org/10.1016/j.watres.2019.115392.

Faecal Indicators
Source Tracking Markers

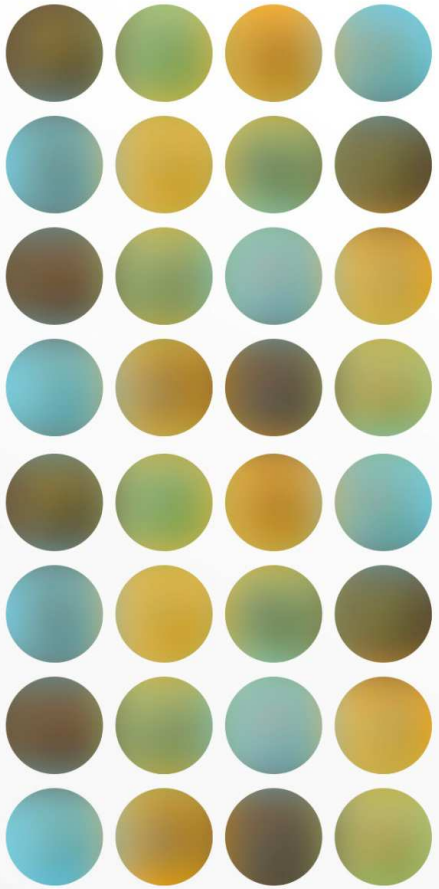
Point Source
Data Matrix



In silico aging and diluting



Aged - diluted Data Matrix



Source Tracking
Prediction Models

Who polluted
the water?

Improving the identification of the source of faecal pollution in water using a modelling approach: from multi-source to aged and diluted samples

Elisenda Ballesté[1*], Luis Antonio Belanche[2], Andreas H. Farnleitner[3,4], Rita Linke[3], Regina Sommer[5], Ricardo Santos[6], Silvia Monteiro[6], Leena Maunula[7], Satu Oristo[7], Andreas Tiehm A[8], Claudia Stange[8], Anicet R. Blanch [1]

[1]Dept. Genetics, Microbiology and Statistics, University of Barcelona, Catalonia. Spain

[2]Dept. of Computer Science, Technical University of Catalonia, Spain

[3]Institute of Chemical, Environmental and Bioscience Engineering, Research Group Environmental Microbiology and Molecular Diagnostics 166/5/3, TU Wien, Getreidemarkt 9/166, 1060 Vienna, Austria.

[4]Karl Landsteiner University of Health Sciences, Research Division Water Quality and Health, Dr.-Karl-Dorrek-Straße 30, 3500 Krems an der Donau, Austria

[5]Unit of Water Hygiene, Institute for Hygiene and Applied Immunology, Medical University of Vienna, Kinderspitalgasse 15, 1090 Vienna, Austria.

[6] Laboratório Analises, Instituto Superior Tecnico. Universidade Lisboa. Lisbon. Portugal[7]Dept. Food Hygiene and Environmental Health, Faculty of Veterinary Medicine, University of Helsinki. Finland

[8]Dept. Microbiology and Molecular Biology, DVGW-Technologiezentrum Wasser, Germany

Running title: Determining faecal pollution sources in water

Keywords: microbial source tracking; faecal pollution; machine learning methods; modelling; water management;

*To whom correspondence should be addressed:

Phone: +34 934039040

Fax: +34 934039047

E-mail:eballeste@ub.edu

Abstract

The last decades have seen the development of several source tracking (ST) markers to determine the source of pollution in water, but none of them show 100% specificity and sensitivity. Thus, a combination of several markers might provide a more accurate classification. In this study Ichnaea® software was improved to generate predictive models, taking into account ST marker decay rates and dilution factors to reflect the complexity of ecosystems. A total of 106 samples from 4 sources were collected in 5 European regions and 30 faecal indicators and ST markers were evaluated, including *E. coli*, enterococci, clostridia, bifidobacteria, somatic coliphages, host-specific bacteria, human viruses, host mitochondrial DNA, host-specific bacteriophages and artificial sweeteners. Models based on linear discriminant analysis (LDA) able to distinguish between human and non-human faecal pollution and identify faecal pollution of several origins were developed and tested with 36 additional laboratory-made samples. Almost all the ST markers showed the potential to correctly target their host in the 5 areas, although some were equivalent and redundant. The LDA-based models developed with fresh faecal samples were able to differentiate between human and non-human pollution with 98.1% accuracy in leave-one-out cross-validation (LOOCV) when using 2 molecular human ST markers (HF183 and HMBif), whereas 3 variables resulted in 100% correct classification. With 5 variables the model correctly classified all the fresh faecal samples from 4 different sources. Ichnaea® is a machine-learning software developed to improve the classification of the faecal pollution source in water, including in complex samples. In this project the models were developed using samples from a broad geographical area, but they can be tailored to determine the source of faecal pollution for any user.

Introduction

Since the beginning of the millennium a big research effort has led to the development of new methodologies and indicators for determining the origin of faecal pollution in water, known as

source tracking (ST) markers. These tools complement the traditional faecal indicator bacteria such as *Escherichia coli* and enterococci, and their capacity to identify the source of faecal pollution has improved the management and assessment of water bodies (Bradshaw et al., 2016; Heaney et al., 2015). Research in this area has been focused mainly on the development of new molecular markers targeting closely related host-specific microorganisms (Hagedorn et al., 2011), establishing protocols, and determining levels of specificity and sensitivity (Bernhard and Field, 2000; Bonjoch et al., 2004; Dick et al., 2005; García-Aljaro et al., 2017; H C Green et al., 2014; Mieszkin et al., 2009; Reischer et al., 2006). Other methods rely on phage detection by culture (Ebdon et al., 2007; Gómez-Doñate et al., 2011).

However, ST methods have some limitations. i) As geographical areas differ in host genetics, immunological factors, antibiotic usage, and diet, all of which affect microbiota, ST markers should be monitored and validated in the target region prior to their application (Gawler et al., 2007; Mayer et al., 2018; Reischer et al., 2013; Yahya et al., 2017). ii) No available marker shows 100% sensitivity and specificity. Accuracy may nevertheless be enhanced by using a combination of several ST markers and ratios, which can be evaluated using predictive models to improve decision-making strategies (Ahmed et al., 2007; Ballesté et al., 2010; Blanch et al., 2006; Gourmelon et al., 2010). iii) There is a lack of standardized methods. Few studies have focused on the standardization and validation of protocols developed by independent laboratories, although this is a crucial step for the consolidation of feasible and reliable approaches (Blanch et al., 2004; Stewart et al., 2013). Furthermore, iv) environment factors need to be considered when monitoring a faecal pollution event, including dilution in the water body, inactivation of the tested parameters, and mixing with other potential pollution sources (Casanovas-Massana et al., 2015). Accordingly, several authors have evaluated the environmental persistence and water treatment resistance of ST markers as factors in management strategies (Ahmed et al., 2007; Bae and Wuertz, 2009; Balleste and Blanch, 2010a; Brooks and Field, 2017; Green et al., 2011; He et al., 2015; Jeanneau et al., 2012; Walters and

Field, 2009). The incorporation of inactivation parameters, together with the dilution effect in the water body, is essential for developing ST predictive models adjusted for the complexity of ecosystems and water flows.

Computational techniques have a wide scope of application in microbiology, ranging from predicting human health and ethnicity through the microbiome to defining the microbial load of a sea sponge (Mason et al., 2013; Walters et al., 2014). Two software systems designed to assess the source of faecal pollution in water are Ichnaea® (Sánchez et al. 2011), which analyses different markers and indicators commonly monitored in water samples, and SourceTracker (Knights et al., 2011), which relies on the results obtained by high-throughput sequencing. Ichnaea® supports the definition and building of models that can predict multiple sources of faecal pollution. It combines different ST markers, thereby obtaining better sensitivity and specificity than a single marker, and takes into account the effects of dilution of the pollution event and the aging of selected ST markers once they reach the environment. The software incorporates models of phenomena based on empirical data (Sánchez et al., 2011), which allows pattern recognition, classification and prediction (Tarca et al., 2007).

In this international and interlaboratory study, the combined use of culture-dependent and -independent methods to identify pollution was tested and a standardised approach was developed. The ultimate aim was to provide a new practical, feasible and integrated approach to pollution analysis. Environmental samples from diverse geographic, climatic and dietary sources were used to address the issues of geographical variability and to carry out testing over a broader area. Several ST markers were selected, including host-specific bacteria targeted by molecular methods (Gomez-Donate et al., 2012; Hyatt C Green et al., 2014; Layton et al., 2006; Mieszkin et al., 2009; Reischer et al., 2006), human viruses (Fong et al., 2005; Maunula et al., 2012; McQuaig et al., 2012; Pina et al., 1998; Rusiñol et al., 2014; Wong et al., 2012; Wyn-Jones et al., 2011), host mitochondrial DNA (mtDNA) (Schill and Mathes, 2008), host-specific bacteriophages detected by culture methods (Gómez-Doñate et al., 2011) and artificial

sweeteners (Scheurer et al., 2009). Standard microbial indicators were measured to assess the total load of faecal pollution (*E. coli*, enterococci, clostridia, total bifidobacteria and somatic coliphages) together with ST markers. The previously developed machine learning-based software Ichnaea® was adapted, trained and tested. Models based on linear discriminant analysis were obtained and the best subsets of indicators and/or ST markers (low number and/or cost, and high predictive ability) to discern the source of faecal pollution were determined.

MATERIALS AND METHODS

Selection of indicators and ST markers

Indiscriminate testing of a large number of protocols and ST markers was not practical, given a tight timeframe and the increasing cost of performing international and integrative ST assays. Consequently, a careful selection of markers (culture-dependent, molecular and chemical) used in several countries of Europe was made according to the following criteria: i) representation of the diversity of currently available methods; ii) library-independent methods; iii) availability of quantification methods; v) and of standard operating procedures (SOP); and vi) ample evidence supporting applicability in an aquatic environment. The selection was also based on the resources and expertise of the participant laboratories and a previous review of the literature. Emphasis was placed on the pre-selection of molecular faecal markers as potential targets in any further technological platforms or automated approaches. The selected ST markers used as variables for modelling are given in Table 1.

Establishing operating principles and quality assurance

Participant laboratories agreed on the use of international standard protocols (ISO, CEN) when available. Other protocols of new indicators were written up and added to those from the literature, together with internal protocols used by some of the laboratories, in a booklet of

standard operating procedures for the use of all participants (http://aquavalens.org/project/latest-results-cluster-1). The results obtained from each laboratory underwent quality control through an initial verification test with blind water samples. The verification test took into account traditional microbial parameters and some culture-dependent ST markers following the agreed SOP: *E. coli* (EC), enterococci, *Clostridium perfringens* (CP), somatic coliphages (SOMCPH) and total and fermenting-sorbitol bifidobacteria (BifTot and BifSorb). Two raw urban sewage samples with high and low faecal concentration were sent blind to all partners. Samples were sent at 4ºC, were delivered in 24 h, and were analysed by all the participants on the same day. Results (enumerations) were sent to the organizer laboratory for statistical analysis.

Samples and sampling campaigns

The five research institutions participating in this study formed an axis across continental Europe (Portugal, Spain, Austria, Germany, and Finland). This consortium allowed the sampling to cover a wide diversity of geographical and climate situations as well as human diets, thus addressing limitations of previous ST studies. Each participant was responsible for collecting samples from their own region, and determining the main culture-based indicators (EC, enterococci, SOMCPH, CP, total BifTot and BifSorb) and their own selected markers. The samples were shipped in cold conditions to the other partner laboratories for the analysis of the other ST markers.

The sampling approach was similar to the procedure followed by a previous integrative and international ST project (Blanch et al., 2006), although the latter was focused on providing predictive models at the faecal point source and distinguishing between human and non-human faecal sources. In the current study, two sampling campaigns were performed to obtain a) point source fresh (PSF) and b) laboratory-made environmental (LME) samples.

The aim of the first sampling campaign was to obtain data from PSF samples to be used as a training matrix in the mathematical modelling. This data matrix was used to classify and select subsets of the best indicators and develop different predictive models (Fig 1). Models were

defined to resolve different scenarios: to distinguish between human and non-human sources or between four sources (human, bovine, porcine and poultry) in fresh samples or those affected by dilution and aging.

In the second sampling campaign, each partner sent blind faecal polluted water samples to the other participants to be analysed and tested by the developed predictive models. These samples could be from faecal point sources or have been diluted and/or aged in the laboratory. The final distribution of samples by sampling campaign was as follows.

Point source fresh samples: A total of 106 faecal and wastewater samples were collected between November 2013 and September 2014 from wastewater treatment plants (WWTP), abattoirs and farms in five different countries: Austria, Finland, Germany, Portugal and Spain. Samples were almost exclusively composed of a unique faecal source: human (35), porcine (24), bovine (23) and poultry (24). Sewage samples came from communities with 2,100 to 4.0 million inhabitants. Wastewater was taken from different abattoirs processing between 400 and 8,000 porcine and ruminant animals per day, and around 100,000 poultry specimens. Other samples were of animal faecal slurry composed of a mix proceeding from at least 10 different individuals. Details of each sample are provided in Supplementary Materials. They were collected in sterile containers and kept at 4°C while in transit to the laboratory. One hundred ml of each sample was sent to the other partner institutions in cold conditions for the assigned analysis.

Laboratory-made environmental samples: A total of 37 samples were laboratory-made by diluting and aging faecal and wastewater samples of different sources to simulate potential environmental samples. The original samples were collected from March to May 2015 from the same WWTP, abattoirs, farms and countries as the PSF samples. Dilutions of faeces/wastewater were made from 1:3 to 1:100,000 using bottled water without faecal pollution and were kept from 0 to 168 h at room temperature for aging. Details of each sample are provided in

Supplementary Materials. Five-hundred ml of each sample was sent blind to each partner institution to be analysed for the selected markers as described above.

Detection and enumeration of general faecal indicators

Five general faecal indicators were measured in each partner laboratory: EC, enterococci, CP spores measured by membrane filtration on 0.45-µm-pore-size membranes, BifTot and BifSorb by spread-plating, and somatic coliphages by a double-agar-layer technique. Enumeration of EC was based on the ISO standard method 16649-1:2001 with an initial resuscitation stage on MMGA (4 h at 37ºC) followed by incubation in chromogenic TBX agar at 44ºC (ISO, 2001a). Enterococci were enumerated following the ISO standard method 7899-2:2000 using Slanetz-Bartley medium at 37ºC for 48 h and confirmed by Bilis Esculine Azide agar at 44ºC for 4 h (ISO, 2000a). CP was analysed according to the ISO standard method 14189 using TSC agar (ISO, 2013a). BifTot and BifSorb enumeration was performed using human bifidobacteria sorbitol-fermenting agar (HBSA) at 37ºC for 48 h in anaerobic conditions as previously described (Bonjoch et al., 2005)**.** Somatic coliphages were enumerated by the double-agar-layer technique using *E. coli* strain WG5 at 37ºC for 24 h, as described in the ISO standard method 10705-2 (ISO, 2000b).

Detection of source tracking markers

Based on the available facilities and experience of the different laboratories, each partner analysed different ST markers in all the samples collected in the 5 regions.

Detection of chemical markers

Four artificial sweeteners, acesulfame, cyclamate, saccharin and sucralose, were measured by high-performance liquid chromatography - electrospray tandem mass spectrometry (HPLC-ESI-MS/MS) as previously described (Scheurer et al., 2009).

Detection of host-specific Bacteroides *phages*

Phages infecting host-specific *Bacteroides* species were enumerated as described in the ISO standard method 10705-4 (ISO, 2001b). PFU of host-specific *Bacteroides* phages were enumerated by the double-agar-layer technique using the strains GA17, PG76, CW18 and PL122 to detect human, porcine, bovine and poultry pollution, respectively (Gómez-Doñate et al., 2011; Payan et al., 2005). One-ml of PSF samples was analysed directly. However, for the highly diluted LME samples, 250 ml was concentrated by membrane filtration using 0.22 µm-pore-size mixed cellulose ester membrane (Merck Millipore, Cork, Ireland) after adding 0.05 mM of $MgCl_2$. The filters were eluted in 12 ml Elution Buffer (1% Beef Extract, 0.5 M NaCl and 3% Tween 80) using an ultrasound bath for 4 min (Méndez et al., 2004). The elution solution pH was brought to 7 and filtered through a low protein-binding 0.2-µm-pore-size PES syringe filter (Merck Millipore) to remove any remaining bacterial cells. One ml of the solution was titred in triplicate with the corresponding host strain.

Detection of molecular ST markers

The genetic material of the shipped samples was extracted in each laboratory where the corresponding markers would be analysed according to routine protocol specifications.

Bifidobacterium host-specific markers

DNA from PSF samples was extracted directly from 1 ml using the QIAamp DNA Blood Mini Kit (Qiagen). In LME samples, 250 ml was concentrated by filtration through a 0.22-µm-pore-size filter (SO-PAK, Millipore, Germany) and DNA was extracted following a previously described protocol (Gourmelon et al., 2007). Filtration and DNA extraction controls were run together with the samples. Total and host-specific *Bifidobacterium* species (HMBif, CWBif, PLBif and PGBif) targeting the 16S rRNA gene were analysed with TaqMan Environmental Master Mix 2.0 (Applied Biosystems) using ABI StepOne Real-Time qPCR as described in the literature (Gomez-Donate et al., 2012) (Table S1).

Host-specific *Bacteroidales* markers

Ten ml of PSF samples and 500 ml of LME samples were concentrated by membrane filtration through Isopore 0.2 µm polycarbonate membrane filters (Millipore, Bedford, MA). DNA was extracted using phenol-chloroform-isoamyl alcohol as described in the literature (Reischer et al., 2008). The respective human, ruminant and swine host-specific *Bacteroidales* markers HF183 (Hyatt C Green et al., 2014), BacR (Reischer et al., 2006) and Pig2Bac (Mieszkin et al., 2009) were analysed together with general *Bacteroidales* marker AllBac (Layton et al., 2006) (Table S1). The QIAGEN Rotor-Gene Multiplex PCR Kit (Qiagen, Hilden, Germany) was used for the qPCR reactions with a Rotor-gene cycler (Qiagen). An internal amplification control (Applied Biosystems, Vienna, Austria) was included for each reaction and samples were always analysed using 1:4 or 1:16 dilution extracts to avoid any potential reaction inhibitors. Filtration and DNA extraction controls were run together with the samples.

Mitochondrial DNA

The analysis of mtDNA to detect faecal contamination of human, bovine, porcine and poultry source was performed targeting the mitochondrial cytochrome *b* by qPCR (Schill and Mathes, 2008). 200 µl of PSF samples was extracted directly using the QIAamp DNA Blood Mini Kit (Qiagen), and in LME samples DNA was extracted following Martellini et al (Martellini et al., 2005). Mitochondrial DNA amplification was performed with TaqMan Environmental Master Mix 2.0 (Applied Biosystems) and using ABI 7300 Real-Time PCR (Applied Biosystems) (Table S1). Several quality control processes were added for the determination of mtDNA. A blank control (filtered, sterile distilled water) was processed in parallel with the LME samples from the concentration stage to the qPCR. Similarly, a blank extraction control was added for both sampling periods. In each run, 10- and 100-fold dilutions of every sample were also tested to account for inhibition. Every qPCR run also had a standard curve and a positive and negative control.

Viral source tracking markers: Adenovirus and Norovirus

Human adenoviruses (HAdV) were amplified following a previously described protocol (Hernroth et al., 2002) using the same DNA extracted from PSF samples for the analysis of mtDNA. As for mtDNA, in addition to the original samples, each HAdV run was comprised of 10- and 100-fold dilutions of every sample, a standard curve and positive and negative controls.

Norovirus GI and GII were amplified following the ISO/TS 15216-1 (ISO, 2013b; Oristo et al., 2018) with some modifications. A sample volume of 250 µl of PSF (or 500µl for diluted samples) was used for RNA extraction. For LME samples, 500 ml was first concentrated by filtration through a positively charged Sartolon membrane (0.45µm-pore-size disc, Sartorius). Viruses from the membrane and the empty bottle were eluted with 100 mM Tris - 50 mM glycine - 1 % beef extract (TGBE) buffer, pH 9.5, after which the pH was adjusted to neutral. RNA from both PSF and LME samples was extracted using the NucliSens® Magnetic Extraction Kit and NucliSens® MiniMag® instrument (Biomerieux, Boxtel, The Netherlands) according to the manufacturer's instructions. The initial sample was spiked with mengovirus to be used as a process positive control (Table S1). Samples were amplified using the QuantiTect Probe RT-PCR Kit (Qiagen, Hilden, Germany) and Rotor-gene PCR cycler (Corbett) (Table S1). For every set of samples, a negative extraction control, positive external RNA controls, and dilutions of purified plasmid dsDNA for the construction of a standard curve were added.

Faecal Enterococci quantification by qPCR

Faecal enterococci were also quantified by qPCR using the DNA extractions for host-specific *Bacteroidales* and following the protocol described elsewhere (Haugland et al., 2005) (Table S1).

Data treatment

PSF sample data were harmonized and standardized to create the *point source training matrix* containing 106 observations (samples) of four animal sources from which 42 variables were analysed: 30 single variables derived from the results of each parameter (8 general faecal

289 indicators, 22 ST markers) and 12 derived variables constituted of ratios of 2 independent
290 variables (Fig 1, Table 1). The results were expressed per 10 ml and data were transformed to
291 $\log_{10}$ units. The *point source training matrix* was instrumental for developing the *age-diluted*
292 *training matrix* by *in silico* dilutions and aging. This matrix was generated creating a realistic
293 scenario of dilution/aging that included 10,000 observations created by randomly sampling the
294 *point source training matrix*. The dilution degree was lognormal up to 4 log units of dilution
295 (alphas) and aging time in water was exponential up to 300 h of aging (times) (Fig S1)
296 considering the decay rate (K_s) of each marker as follows:

$$log10(\text{PSF random value}) - \text{alphas} + K_s * times$$

297 Values above the limit of quantification were assumed to be 10% of the limit of quantification.
298 The predictive models for the four sources using this extended data matrix (dilution and aging
299 included) are the models covering most real expected cases.

300 Similarly, the *testing matrix* was obtained from the harmonization and standardization of the
301 results from the LME samples following the criteria used to develop the *point source training*
302 *matrix*. Results were also expressed per 10 ml and values below the limit of quantification were
303 assumed to be 10% of the limit of quantification. After developing the models using both
304 training matrices and before their validation, the variables not showing significance in the
305 models were disregarded. Therefore, the 38 LME samples were analysed for just 21 of the
306 initial variables.

307 **Inactivation data**

308 The die-off regression in the environment for each measured ST marker and indicator was
309 provided by the partner responsible, based on experimental assays or obtained from the
310 literature (W Ahmed et al., 2014; Balleste and Blanch, 2010b; Dick et al., 2010; Fallahi and
311 Mattison, 2011; Green et al., 2011; Hirneisen and Kniel, 2013; Jeanneau et al., 2012; Korajkic
312 et al., 2014; Liang et al., 2012; Sokolova et al., 2012; Solecki et al., 2011; Tambalo et al., 2012;

313 Walters and Field, 2009). A first order decay model was assumed for all the parameters.
314 Inactivation values included T_{90} (time required to achieve 90% reduction in the initial
315 population), T_{99} (time required to achieve 99% reduction in the initial population), K_s and % of
316 degradation and they were all converted to K_s (Table S2). The effects of seasonality on the
317 environmental persistence of markers were also considered by using different die-off regression
318 models for different seasons. The die-off values were used to consider the decay of each
319 parameter when aging the faecal pollution in the development of predictive models.

320 **Statistical analysis and model evaluation**

321 Descriptive statistics were performed for each of the single variables using the software R (R
322 Core Team, 2016). For descriptive statistics, values above the limit of detection were not
323 considered. The Welch one-way test was applied to detect differences between targeted and
324 non-targeted hosts, and in this case values above the limit of detection were considered as zeros.
325 A Kruskal-Wallis ANOVA by ranks test for non-parametric data was used to evaluate
326 interlaboratory differences.

327 Different models were developed using data from PSF samples represented in the *point source*
328 *training matrix* and from the *age-diluted training matrix* with R software including the packets
329 "MASS", "FSelector", "rgl" "randomForest", and "varSelRF". For both matrices, 2 different
330 scenarios were established: discrimination between human and animal pollution or between
331 human, bovine, porcine and poultry pollution.

332 Numerical analyses were performed using linear discrimination analysis (LDA). This method is
333 a generalization of Fisher's linear discriminant, and is usually applied in statistics, pattern
334 recognition and machine learning to find a linear combination of features that characterizes or
335 separates two or more classes (in our study sources). Obtained results were validated with
336 Leave-one-out cross-validation (LOOCV), a model validation technique for assessing how the
337 results of a statistical analysis will generalize to an independent data set. LOOCV is usually

338 applied in analyses where the goal is prediction and it is necessary to estimate how accurately a
339 predictive model will perform in practice.

340 **RESULTS**

341 Before starting the sampling campaigns, standard operation procedures were established and
342 interlaboratory verification tests were performed involving all the partners. Using a Kruskal-
343 Wallis ANOVA by ranks test for non-parametric data, no statistically significant differences (P-
344 value > 0.05) were observed between laboratories when testing EC, enterococci, SOMCPH and
345 CP, although higher variance was observed for HBSA total and sorbitol-fermenting
346 bifidobacteria (Table S3).

347 **Indicator and marker description**

348 The significance of variables (faecal indicators) and their correlations were previously
349 calculated to support the selection of parameter subsets. Using Fisher's test, differences in the
350 ST markers between target and non-target sources were analysed. Ten of the human markers
351 tested (all except saccharin, for which only 3 human samples were positive) showed significant
352 differences between human and non-human samples (Table S4). The 4 pig ST markers (PGPH,
353 PigNeo, Pig2Bac and PGMit) showed differences between pig and non-pig samples. For the
354 ruminant (CWBif, BacR and CWMit) and poultry markers (PLBif and PLMit), significant
355 differences between target and non-target samples were also observed. However, no significant
356 differences were detected for the ruminant (CWPH) and poultry *Bacteroides* phages (PLPH)
357 analysed, probably due to their geographical specificity, as most of the positive samples were
358 from Spain, where the markers were developed.

359 The correlation between markers was analysed using Pearson's test (Pearson's correlation
360 coefficient r) to evaluate equivalence and redundancy. A strong correlation was observed
361 between the chemical human markers: acesulfame with cyclamate and with sucralose (r = 0.885
362 and 0.681, respectively). GA17PH strongly correlated with acesulfame, cyclamate, HAdV and

HF183 (r = 0.714, 0.669, 0.665 and 0.658), whereas HMBif and HMMit showed a low correlation (r < 0.650) with the remaining human marker. A strong correlation was detected between the animal mitochondrial markers and other ST markers: BacR and CWMit (r = 0.939), PLBif and PLMit (r = 0.816), and Pig2Bac and PGMit (r = 0.805). Phages infecting *Bacteroides* PG76 targeting pig contamination showed a low correlation with norovirus, PGMit and Pig2Bac (r < 0.650), whereas no significant correlation was observed for the ruminant and poultry host-specific *Bacteroides* phages.

Marker selection

Based on preliminary models evaluating the correlation between markers and the experience of the research laboratories, a pool of the analysed variables was disregarded for further analysis. The decision was taken after an agreement with all the project partners to reduce laboratory costs and efforts. Total and sorbitol-fermenting Bifidobacteria detected using HBSA media were discarded because of the subjectivity of colour analysis. Other markers were discarded for their low sensitivity (saccharin) or low specificity (HMMit). Chemical markers, PLPH and HAdV were considered redundant for their high correlation with molecular markers and absence in the preliminary models, and were thus also discarded for further sampling and analysis. The number of evaluated indicators was thereby reduced from 30 to 21. Additionally, ratios were no longer considered in the models as they did not give additional value.

Model Development

We obtained a list of prediction models to distinguish between human and non-human faecal pollution sources, and also between the main faecal pollution sources. Both scenarios were tested using the *point source training matrix* obtained experimentally and the *age-diluted training matrix* developed *in silico* considering the effect of dilution and aging. The following scenarios were evaluated:

Scenario 1: Human vs non-human faecal pollution using the point source training matrix

When using all the 21 variables, 100 % LOOCV accuracy was achieved with LDA, and all the samples were correctly classified (Figure 2A, 2B). After reducing the number of variables, several combinations gave a high percentage of detection (Table 2). LOOCV accuracy was a) 92.45 % when using only one variable (GA17PH); b) 94.34 % when combining GA17PH with HF183; and c) 98.11 % when using HMBif and HF183 (whose individual values were 80.19 % and 84.91 %, respectively). To achieve 100 % correct classification, a combination of 3 variables (HMBif, HF183 and EC) was needed, whereas 14 different options each using 4 variables achieved 100% LOOCV accuracy: all but one included EC, HF183 and HMBif and 1 other variable (SOMCPH, CP, CWPH, PGPH, Pig2Bac, CWMit, PLMit, PGMit, BacR, CWBif, PLBif, AllBac, NoV, FEqPCR). When using only molecular markers, HMBif, HF183 and PLMit should be measured together with BacR or NoV (Table 2).

Scenario 2: Assessment of four sources using the point source training matrix

When using all the 21 variables, 100% LOOCV accuracy was achieved with LDA (Figure 3A). However, 2 combinations of 3 markers, CWMit, PLMit and Pig2Bac or BacR, PLMit and Pig2Bac, gave a LOOCV accuracy of 97.17% and 96.23%, respectively, in distinguishing between samples from human and farm animal sources (bovine, porcine and poultry) (Table 2). Increasing the number of variables to 4 (GA17PH, PLBif, Pig2Bac and CWMit) increased the correct classification to 99.06 %, whereas 11 other combinations gave 98.11% correct classification. Five variables (GA17PH, PLBif, Pig2Bac, CWMit and BacR or HF183) were needed to correctly classify 100 % of the samples.

Scenario 3: Human vs non-human faecal pollution using the aged-diluted training matrix

When testing a more realistic scenario with the aged-diluted training matrix containing 10,000 *in silico*-made samples, a LOOCV accuracy of 99.78 % was achieved when using the 21 variables with a linear discriminant analysis (Fig 2C, Fig 2D). From a total of 3,342 human samples, 20 were misclassified as non-human, and 2 non-human samples from 6,658 were

misclassified as human. LOOCV accuracy was 95.71 % when the number of variables was reduced to 2 (GA17PH and HF183), and 99.59 % when using 3 (HMBif, HF183 and GA17PH). Seven more combinations with 3 variables gave similar values of 96.81 – 98.04 % (Table 3).

Scenario 4: Four sources assessed with the aged-diluted training matrix

An LDA-based model using all the variables showed 99.08 % LOOCV accuracy (Figure 3B). All the human samples were correctly classified, whereas 2.5 % of cow, 1.2 % of pig and 0.5 % of poultry samples were misclassified. A model using PLBif and Pig2Bac showed an LOOCV accuracy of 69.83 % (Table 3). When using 3 variables (BacR, GA17PH and PLBif), the LOOCV accuracy was 87.04 %, which increased to 93.88 % with the addition of a fourth variable (Pig2Bac) and 96.04 % after adding a fifth (HF183). Nine variables were needed to reach 98 % LOOCV accuracy (Table 3).

Model testing with laboratory-made environmental samples

The selected models were tested using *laboratory-made environmental samples*, which were sent blind to the different participant laboratories. The previously selected 21 markers were evaluated using the models developed for the different scenarios. The resulting data were incorporated into the models developed with the diluted and aged sample matrix, as LME samples were diluted and aged.

The different LDA-based models discriminating between human and non-human pollution (using 2 to 21 variables) correctly classified 84.2 % of the laboratory-made samples. The prediction model with only 2 variables (GA17PH and HF183) achieved 95.71 % LOOCV accuracy (Table 3). All the 6 misclassified samples were of human source identified as non-human.

Models using different combinations of markers to distinguish between the 4 sources were also tested with the *laboratory-made environmental samples*. In this case, 86.8% of the samples were correctly classified. However, the model able to classify the highest number of samples

was the one using 4 variables: BacR, GA17PH, Pig2Bac and PLBif. This model correctly classified 89.5 % of the samples with 93.88% LOOCV accuracy (Table 3). Three of the 4 misclassified samples were poultry samples misclassified as human. Models with a higher number of variables showed a lower % of correct validation.

DISCUSSION

The immense efforts invested in designing new reliable microbial ST markers to determine sources of water pollution have resulted in a toolbox full of markers (Ahmed et al., 2016; Harwood et al., 2014; Roslev and Bukh, 2011). However, as mentioned, a single marker may not be sensitive or specific enough to effectively identify the source of faecal pollution, but a combination of markers can improve the accuracy of classification (Ballesté et al., 2010; Jenkins et al., 2009; Mayer et al., 2018; Raith et al., 2013), sometimes in the form of ratios (Muniesa et al., 2012). Computational techniques based on machine-learning algorithms, like Ichnaea® or SourceTracker, are available for ST prediction (Knights et al., 2011; Sánchez et al., 2011). These algorithms may be based on artificial neural networks or random forests and can be trained with known samples to classify environmental samples of unknown origin (McLellan and Eren, 2014; Smith et al., 2010). SourceTracker relies on 16S rRNA gene amplicons obtained by high throughput sequencing and compares them to a database to calculate the probability that an operational taxonomic unit present in the bacterial community in environmental water samples comes from a given pollution source. This is therefore a library-dependent method (Henry et al., 2016; Knights et al., 2011). Computational methods based on the antibiotic resistance profile of *E. coli* strains have also been tested, resulting in 74.6 % correct classification when using LDA and 82.3 % with random forests (Smith et al., 2010). This approach is also library-dependent and requires the culture of *E. coli* strains. In contrast, Ichnaea® relies on library-independent markers and standardized methods selected for each laboratory reporting the abundance of faecal indicators or host-specific markers. It is a prototype

computer-based integrated system that can be trained by users with their own data matrix developed with the numerous ST indicators available. To improve classification accuracy, the software develops models combining different markers, which can also be chosen by the user (Casanovas-Massana et al., 2015; Sánchez et al., 2011).

The main aim of this study was to adapt Ichnaea® software to select a combination of ST markers from the general ST toolbox and build optimal models to determine the source of faecal pollution in a water body in a given area taking into account aging and dilution. We tested a total of 30 markers and 12 ratios in 106 fresh faecal samples from 5 European regions with different climates and cultural habits. When a new set of ST markers are developed and presented to the scientific community, in the first instance they are normally tested with fresh faecal samples and sometimes with environmental samples. Although this is a good starting point, assessing marker performance in the real environment is more challenging, because of the impact of other factors (Cho et al., 2016): dilution in the water body and the effect of rainfall (Sercu et al., 2011), aging of the pollution between discharge and sampling (Ballesté et al., 2018; Blaustein et al., 2013; Van Kessel et al., 2007), or mixing with other potential faecal sources. To approximate real conditions, an *in silico* matrix of 10,000 samples was generated using faecal samples, taking into account their potential dilution and aging in the environment. This proved to be an appropriate strategy for modelling and to our knowledge, it is the only computational approach reported to date that takes these factors into consideration.

Although the matrix was virtual, it allowed us to achieve our objective, as it included a large number of samples covering different situations found in the environment. A lognormal distribution was used to dilute the samples and an exponential distribution to age them. These distribution approaches to generate *in silico* matrices can be modified according to the application context. In this case, the *age-diluted training matrix* showed 0.13 % negative values (zeros), whereas the *laboratory-made testing matrix* sent to the partners as blind samples

494 showed 2.38 % negative values. A similar number of negative samples should be observed
495 between the *in silico*-created samples and real samples to indicate a reliable prediction potential.

496 When a large-scale study involving multiple laboratories is performed, it is crucial to address
497 the repeatability and reproducibility of the analysis. In this study, standard operating procedures
498 were established, and interlaboratory variability was assessed for parameters measured in each
499 laboratory. On the other hand, each laboratory was responsible for some of the selected
500 microbial ST markers, which they tested in all the collected samples, thereby avoiding inter-
501 laboratory differences due to protocols, equipment and consumables (Ebentier et al., 2013;
502 Stewart et al., 2013). No significant differences were found in the interlaboratory analysis,
503 although the parameters BifSorb and BifTot and clostridia spores showed some variance. These
504 discrepancies were checked (confirming the heat-treatment protocol and its performance) and
505 resolved for one parameter (clostridia spores), whereas the others were discarded from further
506 analysis (BifSorb and BifTot).

507 Models to distinguish between human and non-human faecal pollution sources and also to
508 identify faecal pollution of several origins (human, bovine, porcine and poultry) were defined
509 and built using linear discriminant analysis. When fresh faecal samples were used to develop the
510 models, 2 molecular human ST markers (HF183 and HMBif) were able to distinguish between
511 human and non-human pollution with 98.1 % LOOCV accuracy. The additional economic cost
512 of adding 1 complementary variable to achieve 100 % correct classification should be
513 considered by the end user. On the other hand, when using 5 variables the model correctly
514 classified all the fresh faecal samples of four potential sources: human, porcine, ruminant and
515 poultry. It should be noted that these models do not cover any other potential faecal source such
516 as seagulls or pets.

517 The models built using the aged-diluted matrix were more complex. After a certain degree of
518 dilution and aging, the predicted sources of samples may converge, making it difficult to obtain

519 a correct classification. Linear discriminant analysis and random forest (data not shown) using
520 the aged-diluted matrix gave models with similar indicators to those obtained with fresh
521 samples, although with lower LOOCV accuracy. However, when 4 variables were used, the
522 result was very promising, as the LOOCV accuracy was higher than 99%. Random forest has
523 been reported to improve the accuracy of identification (McLellan and Eren, 2014; Smith et al.,
524 2010), but in the current study its performance was below that of LDA (data not shown).

525 A high percentage of the laboratory-made environmental samples sent for blind testing by the
526 selected models were correctly classified, the failures being mainly in human samples. The
527 results highlight the dependence of the method on the specific set of selected markers. For
528 example, the 4-variable model {GA17 + PLBif + BacR + Pig2Bac} gave the best predictive
529 performance, correctly classifying 89.50% of the samples with 93.88% LOOCV accuracy.
530 Some models with more variables achieved a better performance, for example, LOOCV
531 accuracy was 96.04% for the 5-variable model {GA17PH + PLBif + BacR + Pig2Bac + HF183}
532 and 98.03% for the 9-variable model {GA17PH + PLBif + BacR + Pig2Bac + AllBac + HF183
533 + FEqPCR + PGMit + NoV}. However, when these models were tested, their performance level
534 dropped to 76.32% correctly predicted samples. This phenomenon is well-known in pattern
535 recognition and is explained by the fact that the chances of (linear) separability increase with
536 dimension (number of markers). Redundancy (information shared or conveyed by different
537 markers) also contributes to the phenomenon. Altogether, our results suggest that some of the
538 selected markers may not be sufficiently independent (in the sense of conveying new
539 separability information) and therefore could be removed. From an operational point of view,
540 the results highlight the importance of adding a parsimony principle (in the number of selected
541 markers) when choosing the best model.

542 The development of different models allows the user of Ichnaea® to decide the number of
543 variables to be included and the desired rate of correct classification, considering that a high

number of variables increases not only the correct classification but also the time and cost of laboratory analysis.

As ST markers from different geographical areas can vary in sensitivity and specificity (Haramoto and Osada, 2018; Mayer et al., 2018; Yahya et al., 2017), a more local study using regionally tailored ST markers with samples from a smaller geographical range could reduce the number of markers while increasing the power of the models. It should be born in mind that the indicators selected here were the best in a given framework, but they may differ when using another input matrix (different markers, indicators and source samples) or altering the given inactivation, which can vary according to the season and environmental conditions (W. Ahmed et al., 2014; Ballesté et al., 2018; Blaustein et al., 2013; Solecki et al., 2011). Thus, the decay rate and dilution will vary according to the target scenario. The aged-diluted matrix in this study covered up to 300 h (12.5 days), but other types of aging and dilution may occur, depending on the environment. For example, in rainy seasons or after snowmelt the dilution factor becomes more significant, and may also influence flow velocity and transport distances, thereby affecting the age of the pollution (Jonsson and Agerberg, 2015; Reischer et al., 2008). Hence, the distributions used in this Ichnaea® approach to develop the aged-dilution matrix can be modified to match regional conditions and draw scenarios with a better fit. Tailoring models to the area of study by using local fresh faecal samples, as well as more accurate factors of regional and seasonal inactivation, dilution and aging would improve accuracy while reducing the number of variables to be tested. Further approaches should include mixing of different potential sources.

CONCLUSIONS

- Almost all the ST markers tested showed the potential to correctly target their host in the 5 geographical areas. Redundancy among some of the markers showed they can be used indistinctly.

- Ichnaea®, a machine-learning software based on the R script, provides useful and easy-to-use tools to improve the classification of faecal pollution in water, including complex samples potentially aged and diluted. The software can generate tailored models to determine the source of faecal pollution.

- The creation of an *in silico* matrix of aged and diluted samples using point source fresh faecal samples is an effective approach to obtain a high amount of data covering different scenarios and reproducing environmental conditions.

- When a water sample is aged and diluted, the levels of the markers decrease, and becomes difficult to distinguish between samples with different degrees of dilution and aging. In this scenario, no model gives 100% LOOCV accuracy, although 99% was achieved.

- Models based on linear discriminant analysis using a low number of ST markers (between 2 and 5) can achieve LOOCV accuracies of over 95%. Different models can be generated to discriminate between human and non-human pollution or identify 4 potential sources: human, porcine, bovine and poultry.

- Testing with real samples is a crucial step in generating models with better performance.

Acknowledgements

This study was supported by the European Project FP7 KBBE AQUAVALENS, Grant agreement no: 311846, Spanish Government research project CGL2011-25401 and the 2017SGR170 project by the Catalan Government. We thank Kirsi Söderberg for technical support at UH, Nathalie Schuster for her laboratory expertise and Gerhard Lindner for providing

lab support at Institute of Chemical, Environmental and Bioscience Engineering (Vienna); Laura Sala and Marta Gómez for her lab support in the UB.

References

Ahmed, W, Gyawali, P., S Sidhu, J.P., Toze, S., 2014. Relative Inactivation of Faecal Indicator Bacteria and Sewage Markers in Freshwater and Seawater Microcosms. Lett. Appl. Microbiol. n/a-n/a. doi:10.1111/lam.12285

Ahmed, W., Gyawali, P., Sidhu, J.P.S., Toze, S., 2014. Relative inactivation of faecal indicator bacteria and sewage markers in freshwater and seawater microcosms. Lett. Appl. Microbiol. 59, 348–354. doi:10.1111/lam.12285

Ahmed, W., Hughes, B., Harwood, V.J., 2016. Current status of marker genes of bacteroides and related taxa for identifying sewage pollution in environmental waters. Water (Switzerland) 8, 231. doi:10.3390/w8060231

Ahmed, W., Stewart, J., Gardner, T., Powell, D., Brooks, P., Sullivan, D., Tindale, N., 2007. Sourcing faecal pollution: a combination of library-dependent and library-independent methods to identify human faecal pollution in non-sewered catchments. Water Res. 41, 3771–3779. doi:10.1016/j.watres.2007.02.051

Bae, S., Wuertz, S., 2009. Rapid decay of host-specific fecal Bacteroidales cells in seawater as measured by quantitative PCR with propidium monoazide. Water Res. 43, 4850–4859.

Balleste, E., Blanch, A.R., 2010a. Persistence of Bacteroides species populations in a river as measured by molecular and culture techniques. Appl Env. Microbiol 76, 7608–7616. doi:10.1128/AEM.00883-10

Balleste, E., Blanch, A.R., 2010b. Persistence of Bacteroides species populations in a river as measured by molecular and culture techniques. Appl Env. Microbiol 76, 7608–7616. doi:10.1128/AEM.00883-10

Ballesté, E., Bonjoch, X., Belanche, L.A.L.A., Blanch, A.R., 2010. Molecular indicators used in the development of predictive models for microbial source tracking. Appl. Environ. Microbiol. 76, 1789–1795. doi:10.1128/AEM.02350-09

Ballesté, E., García-Aljaro, C., Blanch, A.R., 2018. Assessment of the decay rates of microbial source tracking molecular markers and faecal indicator bacteria from different sources. J. Appl. Microbiol. 125, 1938–1949. doi:10.1111/jam.14058

Bernhard, A.E., Field, K.G., 2000. A PCR assay to discriminate human and ruminant feces on the basis of host differences in Bacteroides-Prevotella genes encoding 16S rRNA. Appl.Environ.Microbiol. 66, 4571–4574.

Blanch, A.R., Belanche-Muñoz, L., Bonjoch, X., Ebdon, J., Gantzer, C., Lucena, F., Ottoson, J., Kourtis, C., Iversen, A., Kühn, I., Moce, L., Muniesa, M., Schwartzbrod, J., Skraber, S., Papageorgiou, G., Taylor, H.D., Wallis, J., Jofre, J., 2004. Tracking the origin of faecal pollution in surface water: an ongoing project within the European Union research programme. J. Water Health 2, 249–60.

631 Blanch, A.R., Belanche-Munoz, L., Bonjoch, X., Ebdon, J., Gantzer, C., Lucena, F., Ottoson, J.,
632 Kourtis, C., Iversen, A., Kuhn, I., Moce, L., Muniesa, M., Schwartzbrod, J., Skraber, S.,
633 Papageorgiou, G.T., Taylor, H., Wallis, J., Jofre, J., 2006. Integrated analysis of
634 established and novel microbial and chemical methods for microbial source tracking.
635 Appl.Environ.Microbiol. 72, 5915–5926.

636 Blaustein, R. a., Pachepsky, Y., Hill, R.L., Shelton, D.R., Whelan, G., 2013. Escherichia coli
637 survival in waters: Temperature dependence. Water Res. 47, 569–578.
638 doi:10.1016/j.watres.2012.10.027

639 Bonjoch, X., Ballesté, E., Blanch, A.R., 2005. Enumeration of bifidobacterial populations with
640 selective media to determine the source of waterborne fecal pollution. Water Res. 39,
641 1621–1627.

642 Bonjoch, X., Ballesté, E., Blanch, A.R.R., Balleste, E., Blanch, A.R.R., Ballesté, E., Blanch,
643 A.R.R., 2004. Multiplex PCR with 16S rRNA gene-targeted primers of bifidobacterium
644 spp. to identify sources of fecal pollution. Appl. Environ. Microbiol. 70, 3171–3175.
645 doi:10.1128/AEM.70.5.3171-3175.2004

646 Bradshaw, J.K., Snyder, B.J., Oladeinde, A., Spidle, D., Berrang, M.E., Meinersmann, R.J.,
647 Oakley, B., Sidle, R.C., Sullivan, K., Molina, M., 2016. Characterizing relationships
648 among fecal indicator bacteria, microbial source tracking markers, and associated
649 waterborne pathogen occurrence in stream water and sediments in a mixed land use
650 watershed. Water Res. 101, 498–509. doi:10.1016/j.watres.2016.05.014

651 Brooks, L.E., Field, K.G., 2017. Global model fitting to compare survival curves for faecal
652 indicator bacteria and ruminant-associated genetic markers. J. Appl. Microbiol. 122, 1704–
653 1713. doi:10.1111/jam.13454

654 Casanovas-Massana, A., Gomez-Donate, M., Sanchez, D., Belanche-Munoz, L.A., Muniesa,
655 M., Blanch, A.R., 2015. Predicting fecal sources in waters with diverse pollution loads
656 using general and molecular host-specific indicators and applying machine learning
657 methods. J Env. Manag. 151, 317–325. doi:10.1016/j.jenvman.2015.01.002

658 Cho, K.H., Pachepsky, Y.A., Oliver, D.M., Muirhead, R.W., Park, Y., Quilliam, R.S., Shelton,
659 D.R., 2016. Modeling fate and transport of fecally-derived microorganisms at the
660 watershed scale: State of the science and future opportunities. Water Res. 100, 38–56.
661 doi:10.1016/j.watres.2016.04.064

662 Dick, L.K., Bernhard, A.E., Brodeur, T.J., Santo Domingo, J.W., Simpson, J.M., Walters, S.P.,
663 Field, K.G., 2005. Host distributions of uncultivated fecal Bacteroidales bacteria reveal
664 genetic markers for fecal source identification. Appl.Environ.Microbiol. 71, 3184–3191.

665 Dick, L.K., Stelzer, E.A., Bertke, E.E., Fong, D.L., Stoeckel, D.M., 2010. Relative decay of
666 Bacteroidales microbial source tracking markers and cultivated Escherichia coli in
667 freshwater microcosms. Appl Env. Microbiol 76, 3255–3262. doi:10.1128/AEM.02636-09

668 Ebdon, J., Muniesa, M., Taylor, H., 2007. The application of a recently isolated strain of
669 Bacteroides (GB-124) to identify human sources of faecal pollution in a temperate river
670 catchment. Water Res. 41, 3683–3690. doi:10.1016/j.watres.2006.12.020

671 Ebentier, D.L., Hanley, K.T., Cao, Y., Badgley, B.D., Boehm, A.B., Ervin, J.S., Goodwin, K.D.,
672 Gourmelon, M., Griffith, J.F., Holden, P. a., Kelty, C. a., Lozach, S., McGee, C., Peed, L.
673 a., Raith, M., Ryu, H., Sadowsky, M.J., Scott, E. a., Domingo, J.S., Schriewer, A.,
674 Sinigalliano, C.D., Shanks, O.C., Van De Werfhorst, L.C., Wang, D., Wuertz, S., Jay, J. a.,

675 2013. Evaluation of the repeatability and reproducibility of a suite of qPCR-based
676 microbial source tracking methods. Water Res. 47, 6839–6848.
677 doi:10.1016/j.watres.2013.01.060

678 Fallahi, S., Mattison, K., 2011. Evaluation of Murine Norovirus Persistence in Environments
679 Relevant to Food Production and Processing. J. Food Prot. 74, 1847–1851.
680 doi:10.4315/0362-028X.JFP-11-081

681 Fong, T.T., Griffin, D.W., Lipp, E.K., 2005. Molecular assays for targeting human and bovine
682 enteric viruses in coastal waters and their application for library-independent source
683 tracking. Appl. Environ. Microbiol. doi:10.1128/AEM.71.4.2070-2078.2005

684 García-Aljaro, C., Ballesté, E., Muniesa, M., Jofre, J., 2017. Determination of crAssphage in
685 water samples and applicability for tracking human faecal pollution. Microb. Biotechnol.
686 10, 1775–1780. doi:10.1111/1751-7915.12841

687 Gawler, A.H., Beecher, J.E., Brandao, J., Carroll, N.M., Falcao, L., Gourmelon, M., Masterson,
688 B., Nunes, B., Porter, J., Rince, A., Rodrigues, R., Thorp, M., Walters, J.M., Meijer, W.G.,
689 2007. Validation of host-specific Bacteriodales 16S rRNA genes as markers to determine
690 the origin of faecal pollution in Atlantic Rim countries of the European Union. Water Res.
691 41, 3780–3784.

692 Gomez-Donate, M., Balleste, E., Muniesa, M., Blanch, A.R., 2012. New Molecular Quantitative
693 PCR Assay for Detection of Host-Specific Bifidobacteriaceae Suitable for Microbial
694 Source Tracking. Appl. Environ. Microbiol. 78, 5788–5795. doi:10.1128/AEM.00895-12

695 Gómez-Doñate, M., Payán, A., Cortés, I., Blanch, A.R., Lucena, F., Jofre, J., Muniesa, M.,
696 2011. Isolation of bacteriophage host strains of *Bacteroides* species suitable for tracking
697 sources of animal faecal pollution in water. Environ. Microbiol. 13, 1622–1631.
698 doi:10.1111/j.1462-2920.2011.02474.x

699 Gourmelon, M., Caprais, M.P., Mieszkin, S., Marti, R., Wery, N., Jarde, E., Derrien, M., Jadas-
700 Hecart, A., Communal, P.Y., Jaffrezic, A., Pourcher, A.M., 2010. Development of
701 microbial and chemical MST tools to identify the origin of the faecal pollution in bathing
702 and shellfish harvesting waters in France. Water Res. 44, 4812–4824.

703 Gourmelon, M., Caprais, M.P., Segura, R., Le, M.C., Lozach, S., Piriou, J.Y., Rince, A., 2007.
704 Evaluation of two library-independent microbial source tracking methods to identify
705 sources of fecal contamination in French estuaries. Appl.Environ.Microbiol. 73, 4857–
706 4866.

707 Green, H C, Haugland, R.A., Varma, M., Millen, H.T., Borchardt, M.A., Field, K.G., Walters,
708 W.A., Knight, R., Sivaganesan, M., Kelty, C.A., Shanks, O.C., 2014. Improved HF183
709 Quantitative Real-Time PCR Assay for Characterization of Human Fecal Pollution in
710 Ambient Surface Water Samples. Appl. Environ. Microbiol. 80, 3086–3094.
711 doi:10.1128/AEM.04137-13

712 Green, Hyatt C, Haugland, R.A., Varma, M., Millen, H.T., Borchardt, M.A., Field, K.G.,
713 Walters, W.A., Knight, R., Sivaganesan, M., Kelty, C.A., Shanks, O.C., 2014. Improved
714 HF183 quantitative real-time PCR assay for characterization of human fecal pollution in
715 ambient surface water samples. Appl. Environ. Microbiol. 80, 3086–94.
716 doi:10.1128/AEM.04137-13

717 Green, H.C., Shanks, O.C., Sivaganesan, M., Haugland, R.A., Field, K.G., 2011. Differential
718 decay of human faecal Bacteroides in marine and freshwater. Env. Microbiol 13, 3235–

719 3249. doi:10.1111/j.1462-2920.2011.02549.x

720 Hagedorn, C., Blanch, A.R., Harwood, V.J., Farnleitner, A.H., Reischer, G.H., Stadler, H.,
721 Kollanur, D., Sommer, R., Zerobin, W., Blöschl, G., Barrella, K.M., Truesdale, J.A.,
722 Casarez, E.A., Giovanni, G.D., 2011. Microbial Source Tracking: Methods, Applications,
723 and Case Studies, Source. Springer New York, New York, NY, NY. doi:10.1007/978-1-
724 4419-9386-1

725 Haramoto, E., Osada, R., 2018. Assessment and application of host-specific Bacteroidales
726 genetic markers for microbial source tracking of river water in Japan. PLoS One 13,
727 e0207727. doi:10.1371/journal.pone.0207727

728 Harwood, V.J., Staley, C., Badgley, B.D., Borges, K., Korajkic, A., 2014. Microbial source
729 tracking markers for detection of fecal contamination in environmental waters:
730 relationships between pathogens and human health outcomes. FEMS Microbiol Rev 38, 1–
731 40. doi:10.1111/1574-6976.12031

732 Haugland, R.A., Siefring, S.C., Wymer, L.J., Brenner, K.P., Dufour, A.P., 2005. Comparison of
733 Enterococcus measurements in freshwater at two recreational beaches by quantitative
734 polymerase chain reaction and membrane filter culture analysis. Water Res. 39, 559–568.
735 doi:10.1016/j.watres.2004.11.011

736 He, X., Chen, H., Shi, W., Cui, Y., Zhang, X.-X.X., 2015. Persistence of mitochondrial DNA
737 markers as fecal indicators in water environments. Sci Total Env. 533, 383–390.
738 doi:10.1016/j.scitotenv.2015.06.119

739 Heaney, C.D., Myers, K., Wing, S., Hall, D., Baron, D., Stewart, J.R., 2015. Source tracking
740 swine fecal waste in surface water proximal to swine concentrated animal feeding
741 operations. Sci. Total Environ. 511, 676–83. doi:10.1016/j.scitotenv.2014.12.062

742 Henry, R., Schang, C., Coutts, S., Kolotelo, P., Prosser, T., Crosbie, N., Grant, T., Cottam, D.,
743 O'Brien, P., Deletic, A., McCarthy, D., 2016. Into the deep: Evaluation of SourceTracker
744 for assessment of faecal contamination of coastal waters. Water Res. 93, 242–253.
745 doi:10.1016/j.watres.2016.02.029

746 Hernroth, B.E., Conden-Hansson, A.-C., Rehnstam-Holm, A.-S., Girones, R., Allard, A.K.,
747 2002. Environmental Factors Influencing Human Viral Pathogens and Their Potential
748 Indicator Organisms in the Blue Mussel, Mytilus edulis: the First Scandinavian Report.
749 Appl. Environ. Microbiol. 68, 4523–4533. doi:10.1128/AEM.68.9.4523-4533.2002

750 Hirneisen, K. a, Kniel, K.E., 2013. Comparing human norovirus surrogates: murine norovirus
751 and Tulane virus. J. Food Prot. 76, 139–43. doi:10.4315/0362-028X.JFP-12-216

752 ISO, 2013a. Water Quality – Detection and Enumeration of Clostridium perfringens – Part 2:
753 Method by Membrane filtration (ISO/CD 6461-2). International Organization of
754 Standardization, Geneve, Switzerland.

755 ISO, 2013b. Microbiology of food and animal feed -- Horizontal method for determination of
756 hepatitis A virus and norovirus in food using real-time RT-PCR -- Part 1: Method for
757 quantification. Geneva, Switzerland.

758 ISO, 2001a. Microbiology of food and animal feeding stuffs - Horizontal method for the
759 enumration of B-glucoronidase-positive Escherichia coli - Part 1: Colony-count technique
760 at 44^{o}C using membranes and 5-bromo-4-chloro-3-indolyl B-glucuronide. (ISO 16649-
761 1:2001 04. International Organization of Standardization, Geneva, Switzerland.

ISO, 2001b. ISO 10705-4: Water quality - Detection and enumeration of bacteriophages. Part 4: Enumeration of bacteriophages infecting Bacteroides fragilis.

ISO, 2000a. Water Quality – Detection and Enumeration of Intestinal Enterococci – Part 2: Membrane Filtration Method (ISO 7899-2: 2000). International Organization of Standardization, Geneve, Switzerland.

ISO, 2000b. Water quality - Detection and enumeration of bacteriophages - Part 2: Enumeration of somatic coliphages. (ISO 10705-2). International Organization of Standardization, Geneva, Switzerland.

Jeanneau, L., Solecki, O., Wery, N., Jarde, E., Gourmelon, M., Communal, P.Y., Jadas-Hecart, A., Caprais, M.P., Gruau, G., Pourcher, A.M., 2012. Relative decay of fecal indicator bacteria and human-associated markers: a microcosm study simulating wastewater input into seawater and freshwater. Env. Sci Technol 46, 2375–2382. doi:10.1021/es203019y

Jenkins, M.W., Tiwari, S., Lorente, M., Gichaba, C.M., Wuertz, S., 2009. Identifying human and livestock sources of fecal contamination in Kenya with host-specific Bacteroidales assays. Water Res 43, 4956–4966. doi:10.1016/j.watres.2009.07.028

Jonsson, A., Agerberg, S., 2015. Modelling of E. coli transport in an oligotrophic river in northern Scandinavia. Ecol. Modell. 306, 145–151. doi:10.1016/j.ecolmodel.2014.10.021

Knights, D., Kuczynski, J., Charlson, E.S., Zaneveld, J., Mozer, M.C., Collman, R.G., Bushman, F.D., Knight, R., Kelley, S.T., 2011. Bayesian community-wide culture-independent microbial source tracking. Nat Meth 8, 761–763.

Korajkic, A., McMinn, B.R., Shanks, O.C., Sivaganesan, M., Fout, G.S., Ashbolt, N.J., 2014. Biotic interactions and sunlight affect persistence of fecal indicator bacteria and microbial source tracking genetic markers in the upper mississippi river. Appl. Environ. Microbiol. 80, 3952–3961. doi:10.1128/AEM.00388-14

Layton, A., McKay, L., Williams, D., Garrett, V., Gentry, R., Sayler, G., 2006. Development of Bacteroides 16S rRNA gene TaqMan-based real-time PCR assays for estimation of total, human, and bovine fecal pollution in water. Appl.Environ.Microbiol. 72, 4214–4224.

Liang, Z., He, Z., Zhou, X., Powell, C.A., Yang, Y., Roberts, M.G., Stoffella, P.J., 2012. High diversity and differential persistence of fecal Bacteroidales population spiked into freshwater microcosm. Water Res 46, 247–257. doi:10.1016/j.watres.2011.11.004

Martellini, A., Payment, P., Villemur, R., 2005. Use of eukaryotic mitochondrial DNA to differentiate human, bovine, porcine and ovine sources in fecally contaminated surface water. Water Res. 39, 541–8. doi:10.1016/j.watres.2004.11.012

Mason, M.R., Nagaraja, H.N., Camerlengo, T., Joshi, V., Kumar, P.S., 2013. Deep Sequencing Identifies Ethnicity-Specific Bacterial Signatures in the Oral Microbiome. PLoS One 8. doi:10.1371/journal.pone.0077287

Maunula, L., Söderberg, K., Vahtera, H., Vuorilehto, V.P., Von Bonsdorff, C.H., Valtari, M., Laakso, T., Lahti, K., 2012. Presence of human noro- and adenoviruses in river and treated wastewater, a longitudinal study and method comparison. J. Water Health. doi:10.2166/wh.2011.095

Mayer, R.E., Reischer, G.H., Ixenmaier, S.K., Derx, J., Blaschke, A.P., Ebdon, J.E., Linke, R., Egle, L., Ahmed, W., Blanch, A.R., Byamukama, D., Savill, M., Mushi, D., Cristóbal,

804 H.A., Edge, T.A., Schade, M.A., Aslan, A., Brooks, Y.M., Sommer, R., Masago, Y., Sato,
805 M.I., Taylor, H.D., Rose, J.B., Wuertz, S., Shanks, O.C., Piringer, H., Mach, R.L., Savio,
806 D., Zessner, M., Farnleitner, A.H., 2018. Global Distribution of Human-Associated Fecal
807 Genetic Markers in Reference Samples from Six Continents. Environ. Sci. Technol. 52,
808 5076–5084. doi:10.1021/acs.est.7b04438

809 McLellan, S.L., Eren, A.M., 2014. Discovering new indicators of fecal pollution. Trends
810 Microbiol. 22, 697–706. doi:10.1016/j.tim.2014.08.002

811 McQuaig, S., Griffith, J., Harwood, V.J., 2012. Association of fecal indicator bacteria with
812 human viruses and microbial source tracking markers at coastal beaches impacted by
813 nonpoint source pollution. Appl. Environ. Microbiol. 78, 6423–6432.
814 doi:10.1128/AEM.00024-12

815 Méndez, J., Audicana, A., Isern, A., Llaneza, J., Moreno, B., Tarancón, M.L., Jofre, J., Lucena,
816 F., 2004. Standardised evaluation of the performance of a simple membrane filtration-
817 elution method to concentrate bacteriophages from drinking water. J. Virol. Methods 117,
818 19–25. doi:10.1016/j.jviromet.2003.11.013

819 Mieszkin, S., Furet, J.P., Corthier, G., Gourmelon, M., 2009. Estimation of pig fecal
820 contamination in a river catchment by real-time PCR using two pig-specific Bacteroidales
821 16S rRNA genetic markers. Appl.Environ.Microbiol. 75, 3045–3054.

822 Muniesa, M., Lucena, F., Blanch, A.R., Payán, A., Jofre, J., 2012. Use of abundance ratios of
823 somatic coliphages and bacteriophages of Bacteroides thetaiotaomicron GA17 for
824 microbial source identification. Water Res. 46, 6410–8. doi:10.1016/j.watres.2012.09.015

825 Oristo, S., Lee, H.-J., Maunula, L., 2018. Performance of pre-RT-qPCR treatments to
826 discriminate infectious human rotaviruses and noroviruses from heat-inactivated viruses:
827 applications of PMA/PMAxx, benzonase and RNase. J. Appl. Microbiol. 124, 1008–1016.
828 doi:10.1111/jam.13737

829 Payan, A., Ebdon, J., Taylor, H., Gantzer, C., Ottoson, J., Papageorgiou, G.T., Blanch, A.R.,
830 Lucena, F., Jofre, J., Muniesa, M., 2005. Method for isolation of Bacteroides
831 bacteriophage host strains suitable for tracking sources of fecal pollution in water. Appl.
832 Environ. Microbiol. 71, 5659–5662. doi:10.1128/AEM.71.9.5659-5662.2005

833 Pina, S., Puig, M., Lucena, F., Jofre, J., Girones, R., 1998. Viral pollution in the environment
834 and in shellfish: Human adenovirus detection by PCR as an index of human viruses. Appl.
835 Environ. Microbiol.

836 R Core Team, 2016. R: A Language and Environment for Statistical Computing, R Foundation
837 for Statistical Computing. doi:10.1007/978-3-540-74686-7

838 Raith, M.R., Kelty, C.A., Griffith, J.F., Schriewer, A., Wuertz, S., Mieszkin, S., Gourmelon, M.,
839 Reischer, G.H., Farnleitner, A.H., Ervin, J.S., Holden, P.A., Ebentier, D.L., Jay, J.A.,
840 Wang, D., Boehm, A.B., Aw, T.G., Rose, J.B., Balleste, E., Meijer, W.G., Sivaganesan,
841 M., Shanks, O.C., 2013. Comparison of PCR and quantitative real-time PCR methods for
842 the characterization of ruminant and cattle fecal pollution sources. Water Res. 47, 6921–8.

843 Reischer, G.H., Ebdon, J.E., Bauer, J.M., Schuster, N., Ahmed, W., Åström, J., Blanch, A.R.,
844 Blöschl, G., Byamukama, D., Coakley, T., Ferguson, C., Goshu, G., Ko, G., de Roda
845 Husman, A.M., Mushi, D., Poma, R., Pradhan, B., Rajal, V., Schade, M.A., Sommer, R.,
846 Taylor, H., Toth, E.M., Vrajmasu, V., Wuertz, S., Mach, R.L., Farnleitner, A.H., 2013.
847 Performance Characteristics of qPCR Assays Targeting Human- and Ruminant-Associated

Bacteroidetes for Microbial Source Tracking across Sixteen Countries on Six Continents. Environ. Sci. Technol. 47, 8548–8556. doi:10.1021/es304367t

Reischer, G.H., Haider, J.M., Sommer, R., Stadler, H., Keiblinger, K.M., Hornek, R., Zerobin, W., Mach, R.L., Farnleitner, A.H., 2008. Quantitative microbial faecal source tracking with sampling guided by hydrological catchment dynamics. Environ. Microbiol. 10, 2598–2608. doi:10.1111/j.1462-2920.2008.01682.x

Reischer, G.H., Kasper, D.C., Steinborn, R., Mach, R.L., Farnleitner, A.H., 2006. Quantitative PCR method for sensitive detection of ruminant fecal pollution in freshwater and evaluation of this method in alpine karstic regions. Appl.Environ.Microbiol. 72, 5610–5614.

Roslev, P., Bukh, A.S., 2011. State of the art molecular markers for fecal pollution source tracking in water. Appl.Microbiol.Biotechnol. 89, 1341–1355.

Rusiñol, M., Fernandez-Cassi, X., Hundesa, A., Vieira, C., Kern, A., Eriksson, I., Ziros, P., Kay, D., Miagostovich, M., Vargha, M., Allard, A., Vantarakis, A., Wyn-Jones, P., Bofill-Mas, S., Girones, R., 2014. Application of human and animal viral microbial source tracking tools in fresh and marine waters from five different geographical areas. Water Res. doi:10.1016/j.watres.2014.04.013

Sánchez, D., Belanche, L., Blanch, A.R., 2011. A Software System for the Microbial Source Tracking Problem. J. Mach. Learn. Res. 17, 7.

Scheurer, M., Brauch, H.-J., Lange, F.T., 2009. Analysis and occurrence of seven artificial sweeteners in German waste water and surface water and in soil aquifer treatment (SAT). Anal. Bioanal. Chem. 394, 1585–1594. doi:10.1007/s00216-009-2881-y

Schill, W.B., Mathes, M. V, 2008. Real-time PCR detection and quantification of nine potential sources of fecal contamination by analysis of mitochondrial cytochrome b targets. Environ. Sci. Technol. 42, 5229–34. doi:10.1021/es800051z

Sercu, B., Van De Werfhorst, L.C., Murray, J.L., Holden, P.A., 2011. Terrestrial sources homogenize bacterial water quality during rainfall in two urbanized watersheds in Santa Barbara, CA. Microb Ecol 62, 574–583. doi:10.1007/s00248-011-9874-z

Smith, A., Sterba-Boatwright, B., Mott, J., 2010. Novel application of a statistical technique, Random Forests, in a bacterial source tracking study. Water Res. 44, 4067–4076. doi:10.1016/j.watres.2010.05.019

Sokolova, E., Astrom, J., Pettersson, T.J., Bergstedt, O., Hermansson, M., 2012. Decay of Bacteroidales genetic markers in relation to traditional fecal indicators for water quality modeling of drinking water sources. Env. Sci Technol 46, 892–900. doi:10.1021/es2024498

Solecki, O., Jeanneau, L., Jarde, E., Gourmelon, M., Marin, C., Pourcher, A.M., 2011. Persistence of microbial and chemical pig manure markers as compared to faecal indicator bacteria survival in freshwater and seawater microcosms. Water Res 45, 4623–4633. doi:10.1016/j.watres.2011.06.012

Stewart, J.R., Boehm, A.B., Dubinsky, E. a., Fong, T.T., Goodwin, K.D., Griffith, J.F., Noble, R.T., Shanks, O.C., Vijayavel, K., Weisberg, S.B., 2013. Recommendations following a multi-laboratory comparison of microbial source tracking methods. Water Res. 47, 6829–6838. doi:10.1016/j.watres.2013.04.063

Tambalo, D.D., Fremaux, B., Boa, T., Yost, C.K., 2012. Persistence of host-associated Bacteroidales gene markers and their quantitative detection in an urban and agricultural mixed prairie watershed. Water Res 46, 2891–2904. doi:10.1016/j.watres.2012.02.048

Tarca, a L., Carey, V.J., Chen, X.W., Romero, R., Draghici, S., 2007. Machine learning and its applications to biology. PLoS Comput. Biol. 3, e116. doi:10.1371/journal.pcbi.0030116

Van Kessel, J.S., Pachepsky, Y.A., Shelton, D.R., Karns, J.S., 2007. Survival of Escherichia coli in cowpats in pasture and in laboratory conditions. J Appl Microbiol 103, 1122–1127. doi:10.1111/j.1365-2672.2007.03347.x

Walters, S.P., Field, K.G., 2009. Survival and persistence of human and ruminant-specific faecal Bacteroidales in freshwater microcosms. Environ.Microbiol. 11, 1410–1421.

Walters, W.A., Xu, Z., Knight, R., 2014. Meta-analyses of human gut microbes associated with obesity and IBD. FEBS Lett. 588, 4223–4233. doi:10.1016/j.febslet.2014.09.039

Wong, K., Fong, T.T., Bibby, K., Molina, M., 2012. Application of enteric viruses for fecal pollution source tracking in environmental waters. Environ. Int. doi:10.1016/j.envint.2012.02.009

Wyn-Jones, A.P., Carducci, A., Cook, N., D'Agostino, M., Divizia, M., Fleischer, J., Gantzer, C., Gawler, A., Girones, R., Holler, C., de Roda Husman, A.M., Kay, D., Kozyra, I., Lopez-Pila, J., Muscillo, M., Nascimento, M.S., Papageorgiou, G., Rutjes, S., Sellwood, J., Szewzyk, R., Wyer, M., 2011. Surveillance of adenoviruses and noroviruses in European recreational waters. Water Res 45, 1025–1038. doi:10.1016/j.watres.2010.10.015

Yahya, M., Blanch, A.R., Meijer, W.G., Antoniou, K., Hmaied, F., Ballesté, E., 2017. Comparison of the Performance of Different Microbial Source Tracking Markers among European and North African Regions. J. Environ. Qual. 46, 760. doi:10.2134/jeq2016.11.0432

Fig 1. Schematic representation of the computational process used to generate and validate microbial source tracking models with Ichnaea®.

Fig 2. Histograms of sample projections onto the linear discriminant (projection vector) given by linear discriminant analysis according to human and non-human sources. This discriminant represents the linear combination of variables that best separate the sources. Shown are four different scenarios: A) point source data matrix using all variables, B) point source data matrix using molecular variables only, C) diluted and aged sample data matrix using all variables and D) diluted and aged data matrix using molecular variables only. These scenarios show a variable degree of source separability (or, alternatively, source overlap), from non-existent in A to a significant one in D.

Fig 3. 3D plots showing sample projections onto the three linear discriminants given by linear discriminant analysis, according to the four pollution sources considered (red: human, green: pig, black: cow, blue: poultry). These discriminants (named LD1 to LD3) represent different linear combinations of variables that best separate the sources, LD1 being the discriminant achieving the highest separability, followed by LD2 and then LD3. Shown are two different scenarios: A) point source data matrix using all variables and B) diluted and aged data matrix using all variables. Again, the plots reflect two very diverse situations of source separability. In A, the four sources form compact and cleanly separated data clusters. As the samples are progressively diluted and aged, separability slowly decreases until it becomes impossible: the information content in the sample vanishes and the data sample converges towards the data origin, regardless of the source.

940 **Table 1**. List of 30 initially selected parameters (microbial indicators and MST markers) for the
941 definition of single and derived variables (ratios) in the statistical and machine learning methods
942 of this study. Their labels are indicated.

Label	Parameter	Method
EC	*Escherichia coli*	(ISO, 2001a) ISO 16649-1:2001
FE	Faecal enterococci	(ISO, 2000a) ISO 7899-2:2000
CP	*Clostridium perfringens* spores	ISO/DIS 14189
SOMCPH	Somatic coliphages	(ISO, 2000b) ISO 10705-2:2000
GA17PH	Human-specific *Bacteroides* phages	(Gómez-Doñate et al., 2011; ISO, 2001b)
CWPH	Cow-specific *Bacteroides* phages	(Gómez-Doñate et al., 2011; ISO, 2001b)
PGPH	Pig-specific *Bacteroides* phages	(Gómez-Doñate et al., 2011; ISO, 2001b)
PLPH	Poultry-specific *Bacteroides* phages	(Gómez-Doñate et al., 2011; ISO, 2001b)
BifSorb	Human *Bifidobacterium* Sorbitol Agar (HBSA yellow colonies)	(Bonjoch et al., 2005)
BifTot	Total *Bifidobacterium* Sorbitol Agar (HBSA total colonies)	(Bonjoch et al., 2005)
HMBif	Human-specific Bifidobacteria by qPCR	(Gomez-Donate et al., 2012)
CWBif	Cow-specific Bifidobacteria by qPCR	(Gomez-Donate et al., 2012)
PGNeo	Pig-specific *Neoscardovia* by qPCR	(Gomez-Donate et al., 2012)
PLBif	Poultry-specific Bifidobacteria by qPCR	(Gomez-Donate et al., 2012)
TLBif	Total Bifidobacteria by qPCR	(Gomez-Donate et al., 2012)
BacR	Ruminant-specific Bacteroidetes by qPCR	(Reischer et al., 2006)
Pig2Bac	Pig-specific Bacteroidetes by qPCR	(Mieszkin et al., 2009)
AllBac	All Bacteroidetes by qPCR	(Layton et al., 2006)
HF183	Human-specific Bacteroidetes by qPCR	(H C Green et al., 2014)
FEqPCR	Faecal enterococci by qPCR	(Haugland et al., 2005)
HMMit	Human-specific Mitochondrial marker by qPCR	(Schill and Mathes, 2008)
CWMit	Cow-specific Mitochondrial marker by qPCR	(Schill and Mathes, 2008)
PGMit	Pig-specific Mitochondrial marker by qPCR	(Schill and Mathes, 2008)
PLMit	Poultry-specific Mitochondrial market by qPCR	(Schill and Mathes, 2008)
Acesulfame	Artificial sweetener	(Scheurer et al., 2009)
Cyclamate	Artificial sweetener	(Scheurer et al., 2009)
Saccharin	Artificial sweetener	(Scheurer et al., 2009)
Sucralose	Artificial sweetener	(Scheurer et al., 2009)
HAdV	Human-specific Adenovirus by qPCR	(Hernroth et al., 2002)
NoV	Norovirus (GI and GII) by qPCR	ISO/TS 15216-1 (Oristo et al., 2018)

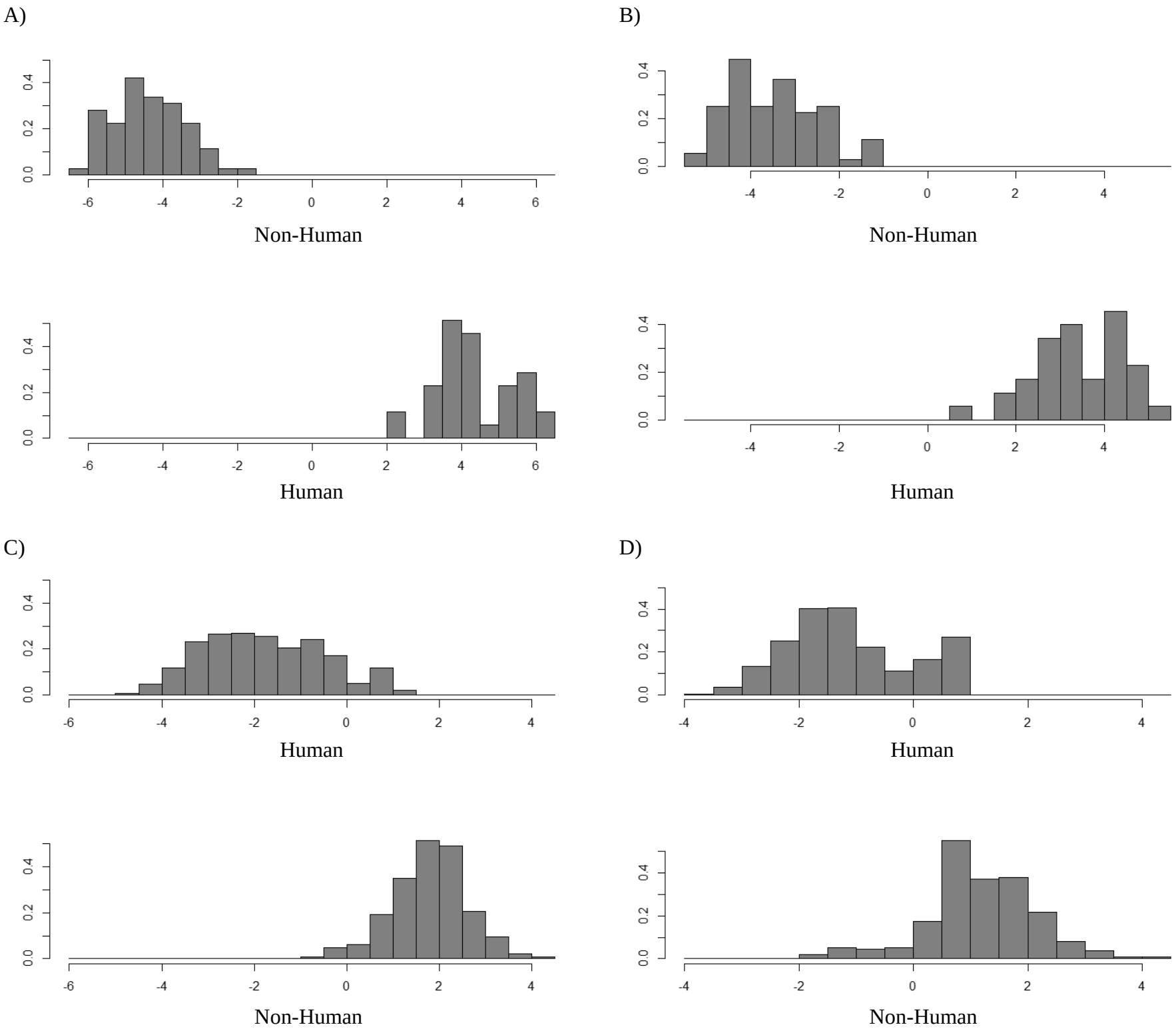
Table 2. Selected subsets of parameters providing the best prediction models using 4, 2 or 1 variable for the different scenarios: distinguishing between human (HM) and non-human (Non-HM) pollution or four pollution sources (human, bovine, poultry and porcine) analyzing the *Point Source Training* Matrix with linear discriminant analysis.

	All Markers			Molecular	
	No. of variables	**Variables**	**LOOCV Accuracy**	**Variables**	**LOOCV Accuracy**
HM vs Non-HM	4			HF183 + HMBif + PLMit + 1 variable (BacR or NoV)	100 %
	3	EC + HF183 + HMBif	100 %	BacR + HF183 + HMBif or PGNeo or PLMit	99.06 %
	2	HF183 + HMBif	98.11 %	HF183 + HMBif	98.11 %
	1	GA17PH	92.45 %	HF183	84.91 %
4 Sources	5	CWMit + GA17PH + Pig2Bac + PLBif + BacR or HF183	100 %	BacR + CWMit + Pig2Bac + PLMit + 1 variable (i. e.: HF183, HMBif, NoV)	99.06 %
	4	CWMit + GA17PH + Pig2Bac + PLBif	99.06 %	CWMit + Pig2Bac + PLMit + 1 variable (i. e.: BacR, HF183, HMBif, NoV)	98.11 %
	3	CWMit + Pig2Bac + PLMit	97.17%	CWMit + Pig2Bac + PLMit	97.17%
	2	Pig2Bac + CWMit or PLMit	76.42%	Pig2Bac + CWMit or PLMit	76.42%

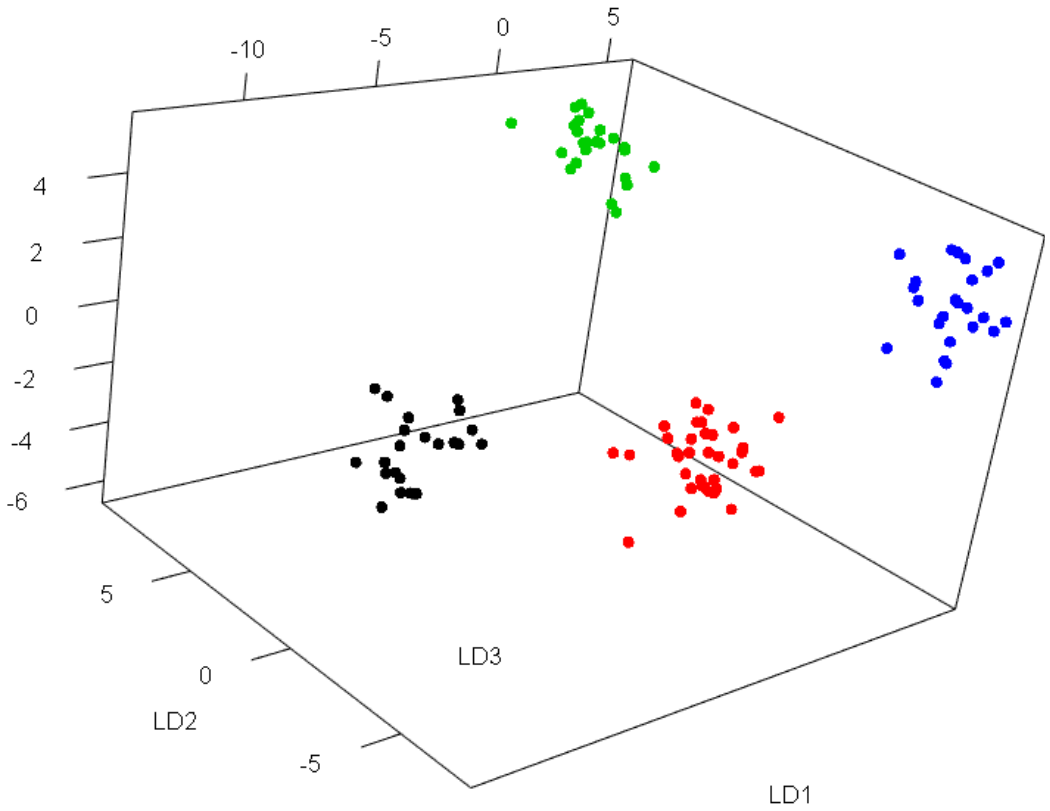
948 Table 3. Selected prediction models obtained by linear discriminant analysis using different numbers of variables. Models to evaluate the different scenarios:
949 distinguishing between human (HM) and non-human (Non-HM) pollution or four pollution sources (human, bovine, poultry and porcine) analyzing the *Aged-*
950 *Diluted Training* Matrix.
951

		All Markers		Molecular	
	No. of variables	**Variables**	**LOOCV Accuracy**	**Variables**	**LOOCV Accuracy**
HM vs Non-HM	4			HF183 + HMBif + PLBif + Pig2Bac	99.14%
	3	EC + HF183 + HMBif	99.59%	HF183 + HMBif + TLBif	93.34 %
	2	GA17 + HF183	95.71%	HF183 + HMBif	91.98 %
4 Sources	5	BacR + GA17PH + HF183 + Pig2Bac + PLBif	96.04%	HMBif + NoV + PGMit + PLBif + PLMit	98.8%
	4	BacR + GA17PH + Pig2Bac + PLBif	93.88%	HF183 + CWMit + Pig2Bac + PLBif	92.26 %
	3	BacR + GA17PH + PLBif	87.04 %	HF183 + Pig2Bac + PLBif	87.07 %
	2	Pig2Bac + PLBif	69.83 %	Pig2Bac + PLBif	69.83 %

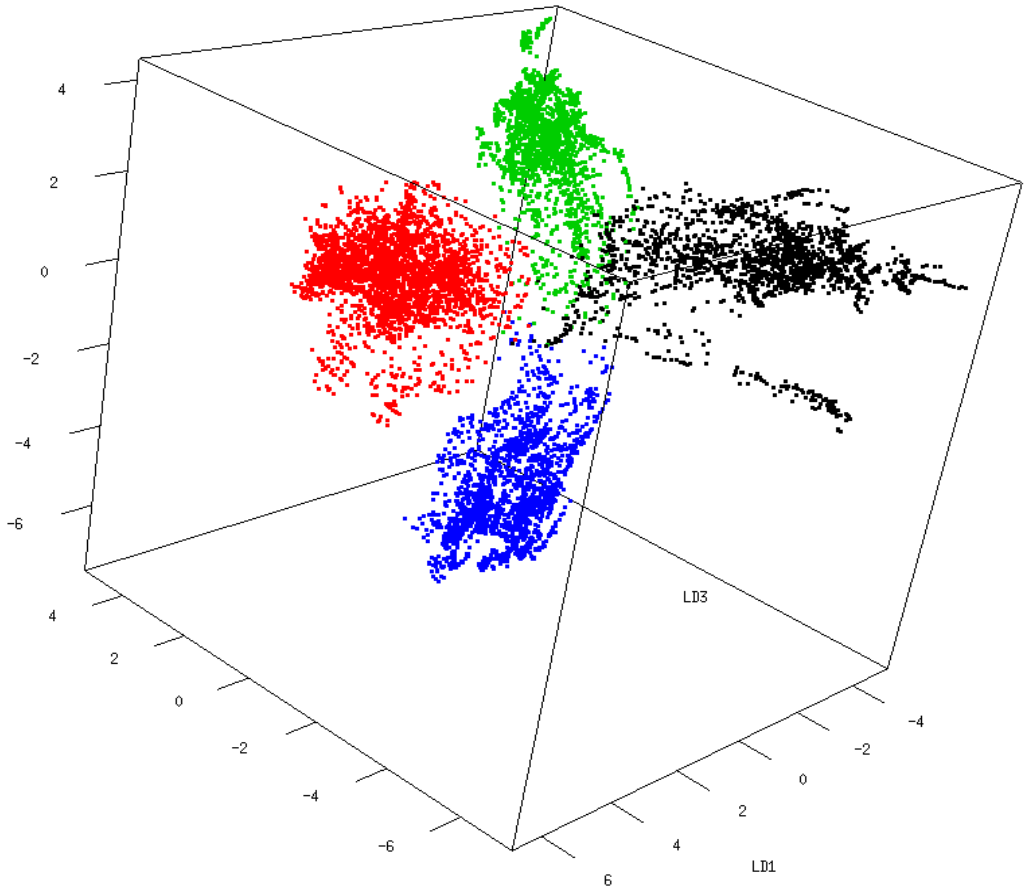
952

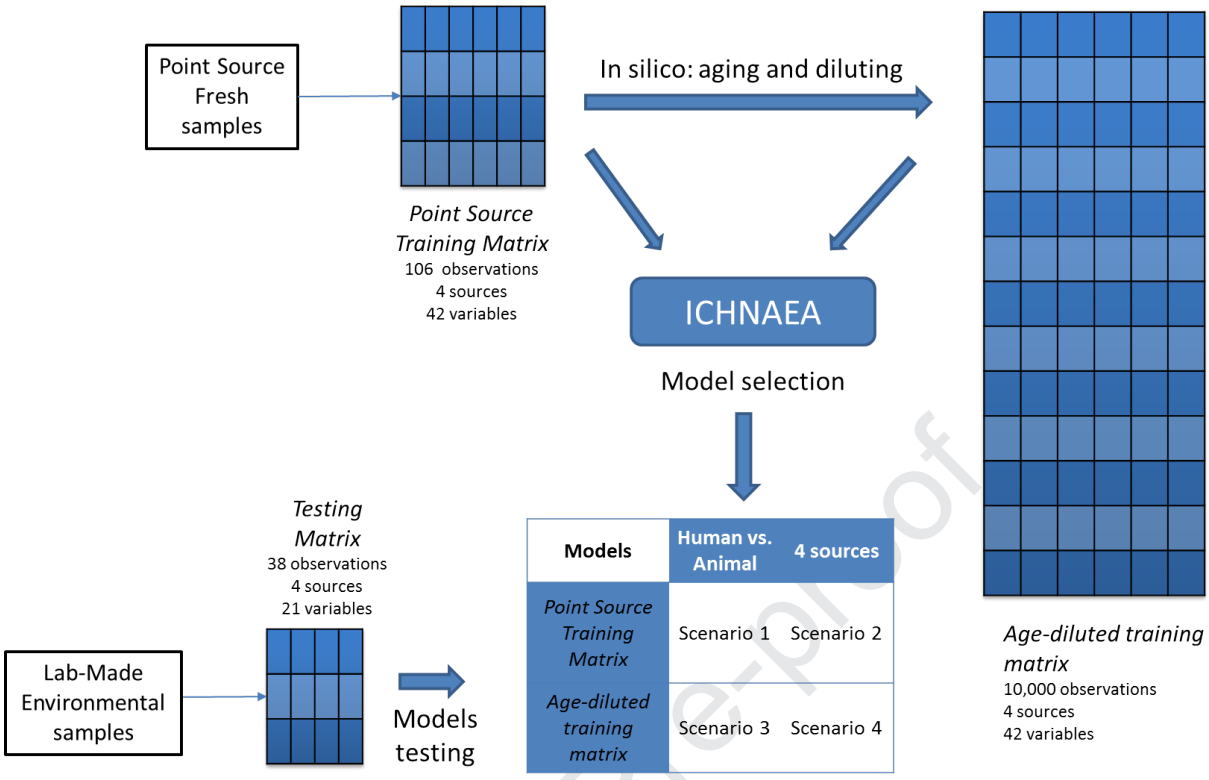

A)
B)
C)
D)
Non-Human
Human
Human
Non-Human
Non-Human
Human
Human
Non-Human

A)



B)





Highlights

- Samples from 5 geographical sources were analysed with 30 faecal markers and indicators.
- A machine learning software was used to develop faecal source discriminant models.
- An *in-silico* matrix was generated using faecal samples, adding dilution and inactivation.
- LDA models' output was a combination of markers able to improve the accuracy of classification.
- Models using between 2 and 5 source tracking markers can achieve LOOCV accuracies of over 95%.

AUTHOR DECLARATION TEMPLATE

We wish to confirm that there are no known conflicts of interest associated with this publication and there has been no significant financial support for this work that could have influenced its outcome.

We confirm that the manuscript has been read and approved by all named authors and that there are no other persons who satisfied the criteria for authorship but are not listed. We further confirm that the order of authors listed in the manuscript has been approved by all of us.

We confirm that we have given due consideration to the protection of intellectual property associated with this work and that there are no impediments to publication, including the timing of publication, with respect to intellectual property. In so doing we confirm that we have followed the regulations of our institutions concerning intellectual property.

We understand that the Corresponding Author is the sole contact for the Editorial process (including Editorial Manager and direct communications with the office). She is responsible for communicating with the other authors about progress, submissions of revisions and final approval of proofs. We confirm that we have provided a current, correct email address which is accessible by the Corresponding Author and which has been configured to accept email from eballeste@ub.edu.

Elisenda Ballesté