Predicting fecal sources in waters with diverse pollution loads using general and molecular host-specific indicators and applying machine learning methods

Arnau Casanovas-Massana[1*▲], Marta Gómez-Doñate[1*], David Sánchez[2], Lluís A. Belanche-Muñoz[2], Maite Muniesa[1] and Anicet R. Blanch[1#]

[1]Department of Microbiology, University of Barcelona, Av. Diagonal 643, Barcelona, Catalonia, Spain.

[2]Department of Software, Technical University of Catalonia, Jordi Girona 1-3, Barcelona, Catalonia, Spain

*These authors contributed equally to this work.

[#]Corresponding author

Mailing address:

Avinguda Diagonal, 643, 08028 Barcelona, Catalonia, Spain

Phone: (+34) 93 402 9012

Fax: (+34) 93 403 9047

Email: ablanch@ub.edu

▲Present address: Department of Epidemiology of Microbial Diseases, Yale School of Public Health, New Haven, Connecticut, USA.

ABSTRACT

In this study we use a machine learning software (Ichnaea) to generate predictive models for water samples with different concentrations of fecal contamination (point source, moderate and low). We applied several MST methods (host-specific *Bacteroides* phages, mitochondrial DNA genetic markers, *Bifidobacterium adolescentis* and *B. dentium* markers, and bifidobacterial host-specific qPCR), and general indicators (*E. coli*, enterococci and somatic coliphages) to evaluate the source of contamination in the samples. The results provided data to the Ichnaea software, that evaluated the performance of each method in the different scenarios and determined the source of the contamination. Almost all MST methods in this study determined correctly the origin of fecal contamination at point source and in moderate concentration samples. When the dilution of the fecal pollution increased (below 3 $\log_{10}$ CFU *E. coli*/100 ml) some of these indicators (bifidobacterial host-specific qPCR, some mitochondrial markers or *B. dentium* marker) were not suitable because their concentrations decreased below the detection limit. Using the data from source point samples, the software Ichnaea produced models for waters with low levels of fecal pollution. These models included some MST methods, on the basis of their best performance, that were used to determine the source of pollution in this area. Regardless the methods selected, that could vary depending on the scenario, inductive machine learning methods are a promising tool in MST studies and may represent a leap forward in solving MST cases.

Keywords: fecal pollution, bacteria, bacteriophages, *Bifidobacterium*, *Bacteroides*, machine learning, microbial source tracking

1. INTRODUCTION

Fecal pollution of water poses public health risks and leads to economic losses and environmental deterioration throughout the world. Effluents from municipal and slaughterhouse wastewater treatment, combined storm-water and sewer overflows, leakage of septic systems, runoff from manure and fecal slurries deposited in fields and grazing pastures, uncontrolled discharge of fecal waste and droppings from wildlife may contaminate surface or ground waters (Ritter et al., 2002; Simpson et al., 2002). Directly monitoring microbial pathogens is generally expensive and technically complex. In addition, pathogens are only present intermittently in water bodies and usually at low concentrations (Field and Samadpour, 2007; Savichtcheva and Okabe, 2006). Because of these limitations, water quality regulations are mainly based on the enumeration of indicator microorganisms such as total coliforms, fecal coliforms, *Escherichia coli* and enterococci. However, using these indicators does not provide information about the source of the fecal contamination. Determination of this source or sources is a key parameter to improve the management of fecal contamination at the origin, by increasing the efficiency of remediation efforts and resolving the legal responsibilities for remediation.

For this purpose, numerous microbial source tracking (MST) methods have been proposed in recent years, although various authors have reported that no single source-tracking method correctly determines the source of fecal pollution in all scenarios (Blanch et al., 2006; Field et al., 2003; Griffith et al., 2003; Harwood et al., 2003; Moore et al., 2005; Myoda et al., 2003; Noble et al., 2003; Samadpour et al., 2005; Stoeckel et al., 2004). A particularly critical issue is the generally poor performance of most of these techniques in

low fecal load matrices (Hagedorn et al., 2011). Therefore, simultaneously combining different methods might be a better approach to identifying the sources of fecal pollution (Blanch et al., 2006; Gourmelon et al., 2007; Griffith et al., 2003; Santo Domingo et al., 2007).

Machine learning is a branch of artificial intelligence concerned with the design and study of algorithms that enable machines, i.e. computers, to learn from data and, in particular, construct models based on empirical data collected from a modeled phenomenon. Thus, it could be used to develop predictive models based on a combination of several MST methods to determine the sources of fecal pollution in water (Belanche-Muñoz and Blanch, 2008; Brion and Lingireddy, 2003). To this end, an integrated open platform machine learning software program, called Ichnaea, was developed by Sánchez et al., 2011. This machine learning approach was originally designed to process a specific data matrix consisting of 103 fecal samples of human, cow, pig, poultry, horse, and sheep origin, described by 26 microbial and chemical fecal indicators obtained in a European study (Blanch et al., 2006). The machine learning method that was developed has been adapted to accept MST data matrices containing samples with fluctuating fecal concentration levels and fecal indicators with different environmental persistence. Moreover, it can be applied to different geographical and climatic areas.

In the present study, Ichnaea was used to determine the main source of fecal pollution in two real scenarios in which the water presented low levels of fecal pollution. To obtain the data required by Ichnaea for the construction of the predictive models, three general fecal indicators (*E. coli*, enterococci and somatic coliphages) were selected, along with four

library-independent MST methods, namely the detection of *Bacteroides* phages specific for human, cattle, pig and poultry fecal pollution (Gómez-Doñate et al., 2011); the PCR analysis of mitochondrial DNA genetic markers associated with cattle, pig and poultry (Kortbaoui et al., 2009; Martellini et al., 2005); a multiplex PCR with *Bifidobacterium adolescentis* and *B. dentium* markers specific for human pollution (Bonjoch et al., 2004); and four bifidobacterial qPCRs specific for human, cattle, pig and poultry pollution (Gómez-Doñate et al., 2012). All these methods were evaluated in unique point source samples and the results used as training matrix for the Ichnaea software. Then, the results from the low fecal pollution scenarios were introduced in the software which generated thousands of predictive models using different combinations of indicators and proposed the potential main source of fecal pollution in each environmental sample. To our knowledge, this is the first attempt to use machine learning methods to determine the origin of fecal pollution in water in a real scenario, which could represent a step forward in the solution of the MST problem.

2. MATERIALS AND METHODS

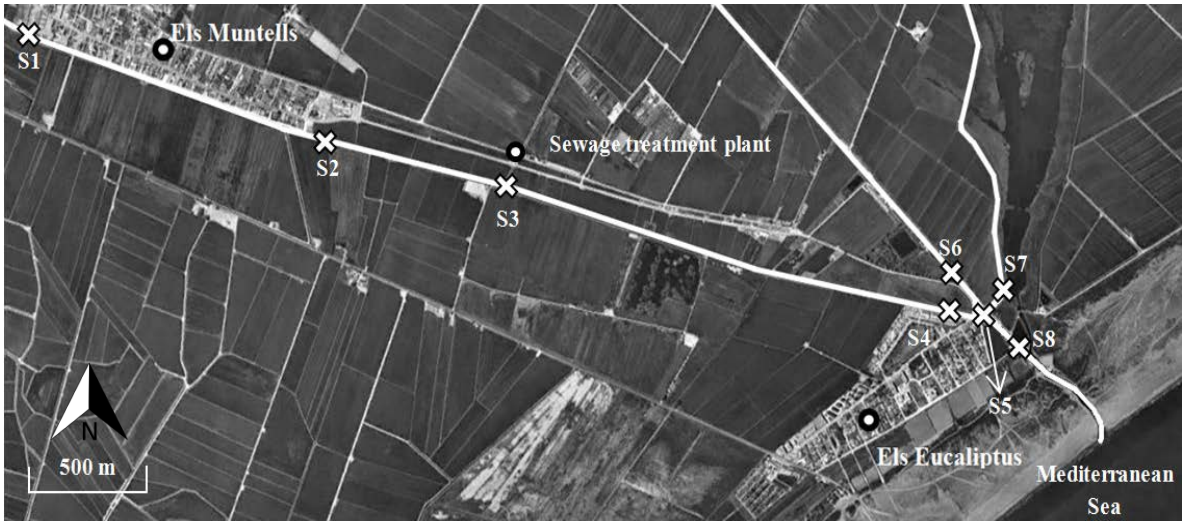
2.1 Study sites and sample collection

The main site of study was a system of water irrigation channels potentially exposed to fecal contamination in the Ebre Delta (Fig 1). The Ebre Delta is a natural delta with an area of 320 km^2 located in Catalonia (Northeastern Spain), where the Ebre River spreads out and drains into the Mediterranean Sea. It is one of the largest wetlands in Western Europe, containing several fresh and saltwater lagoons with a high diversity of bird and fish species. More than 75% of the delta drainage area is composed of agricultural fields, most of which

contain rice crops. The rice cultivation is cyclic: from April to early September, fresh water flows continuously from the river through different channels and floods the rice fields. The water then flows by gravity into collecting channels that finally discharge into the sea. After the harvest in early September, the water circulation is stopped and the fields dry up through evaporation during the winter. In addition to the agricultural fields, there are some scattered cattle-grazing patches in the area under study, and some poultry/livestock facilities nearby.

One sample was collected from each of the sampling sites in May, July, September and November 2010 (Figure 1). The first and second sites (S1, S2) were located in the main channel immediately upstream and downstream of the town of Els Muntells (500-550 inhabitants), respectively. The third sampling site (S3) was downstream from Els Muntells sewage treatment plant. Site four (S4) was located near the Els Eucaliptus complex (250 vacation apartments) just upstream from where the main channel drains into a small lagoon. Samples were taken from the lagoon itself, which was site number five (S5). Samples were collected from two more secondary channels draining into the lagoon (sites S6 and S7). Finally, water was taken from the channel that carries the water from the lagoon to the sea (site S8). Samplings were performed bimonthly over an eight month period to capture seasonal variation and evaluate different water flows related to the irrigation dynamic.

Figure 1. Location of Ebre Delta sampling sites (S1-S8). Aerial photograph courtesy of http://www.Earth.Google.com and the Institut Cartogràfic de Catalunya



Additionally, five freshwater samples were collected monthly from January to May 2010 in the lower course of the Llobregat River (LLOB) at a sampling site located 7 km from the point at which it drains into the sea. According to the regional water authority (http://aca-web.gencat.cat), this river runs through a heavily urbanized area and is mainly subjected to the influence of several effluents from sewage treatment plants. However, the diffuse and irregularly distributed contamination from animals (pets, wildlife and livestock) cannot be ruled out (Lucena et al., 1988; Rubiano et al., 2012).

Inductive machine learning methods require a training matrix of data from samples of known and unique fecal origin in order to develop predictive models for further application in real samples. To this end, human and animal fecal samples were collected from July 2011 to February 2012. Four human sewage samples (HM) were obtained from the influent sewer entering an urban wastewater treatment plant (population equivalent of 400,000) after the bar screen. Five animal samples were obtained from each of the following sources: slurry and wastewater from a pig slaughterhouse (PG); poultry slaughterhouse wastewater

effluent (PL); and pooled feces and process water from a ruminant slaughterhouse (CW). All slaughterhouses were located in Catalonia (Northeastern Spain) and had separate pipes for animal wastewater and the human wastewater from employees' lavatories and showers.

In all samplings, two liters of water were collected in polyethylene containers and refrigerated at 4°C for up to 6 h before being processed in accordance with standardized protocols (Clescerl et al., 1998; International Organization for Standardization, 1994)

2.2 *E. coli*, enterococci and somatic coliphages enumeration

All water samples were filtered through a 0.44 µm nitrocellulose filter (Millipore, Molsheim, France). Filters were then placed on Chromocult® agar plates (Merck, Darmstadt, Germany) at 44.5ºC for 24 h to enumerate *E. coli* or on m-*Enterococcus* agar plates (Difco, Sparks, MD, USA) at 37ºC for 48 h followed by Bile Esculin Agar (Scharlau, Barcelona, Spain) for 3 h at 44ºC to enumerate the enterococci colonies, based on the hydrolysis of esculin (Figueras et al., 1998; Manero and Blanch, 1999). Somatic coliphages were counted after filtration of the sample through low protein-binding 0.22-µl pore size membrane filters (Millex-GP; Millipore, Molsheim, France), and the phages were analyzed by the double agar layer technique using *E. coli* strain WG5 in accordance with the standardized procedure (International Organization for Standardization, 2000). The volumes analyzed varied according to the expected contamination of the sample. For point-source samples several ten-fold dilutions in phosphate buffer saline were prepared and at least three of them were analyzed in duplicate. For environmental water samples three different volumes were tested in duplicate (1, 10 and 25 ml).

2.3 Detection of host-specific *Bacteroides* phages

Bacteroides thetaiotaomicron strains GA17 and CW18, and *B. fragilis* strains PG76 and PL122 were used as host strains for phage detection. These strains have been reported to be specific for the detection of phages from human, cattle, pig and poultry fecal sources, respectively (Gómez-Doñate et al., 2011; Payan et al., 2005). Phages infecting these *Bacteroides* strains were enumerated in all samples using the double agar layer plaque assay, as previously described (International Organization for Standardization, 2001). Additionally, the ratios between somatic coliphages and each one of the host-specific *Bacteroides* phages were calculated and used as indicators as they has been proposed as good candidates to discriminate the sources of fecal pollution (Muniesa et al., 2012)

2.4 Bacterial DNA extraction

DNA was directly extracted from 200 µl of slaughterhouse and wastewater samples using the QIAamp® DNA Blood Mini Kit (Qiagen GmbH, Hilden, Germany) in accordance with the manufacturer's instructions. Water samples from the Ebre Delta were first concentrated by centrifuging 200 ml of water at 16,000 *g* for 10 min. The pellet was resuspended in 10 ml of Ringer ¼ solution (Oxoid, Basingstoke, United Kingdom) and centrifuged again at 16,000 *g* for 10 min. The pellet was then resuspended in 200 µl of Ringer ¼ solution and extracted with the kit as described above. Water samples from the Llobregat River were concentrated by centrifuging 50 ml of water at 16,000 *g* for 10 min and the pellet was resuspended in 200 µl of Ringer ¼ solution and extracted with the kit as described above. Controls of DNA extraction with ultrapure water were performed to exclude contamination during the DNA extraction process in every batch of samples.

2.5 PCR analysis of mitochondrial markers

Three nested PCR assays were performed on each sample to detect mitochondrial host-specific DNA associated with cattle, pigs and chicken (BOMITO, POMITO and CKMITO, respectively). Primers and the conditions used were the same as those described previously (Kortbaoui et al., 2009; Martellini et al., 2005) with the following modifications: 2 µl of the first PCR reaction was used in the nested PCR, and 0.4 µg μL^{-1} of bovine serum albumin (BSA; Madison, WI, USA) was added to all reactions. Negative and positive controls were performed in all sets of reactions by using ultrapure water or mitochondrial DNA extractions from cow, pig and chicken liver. The amplicons were separated after electrophoresis on 1.5% agarose gels and visualized by ethidium bromide staining. A list of all primers and probes used in this study can be found in Table 1.

2.6 *B. adolescentis* (ADO) and *B. dentium* (DEN) detection

A multiplex nested PCR (Table 1) was performed in each sample to detect *B. adolescentis* (ADO) and *B. dentium* (DEN) human fecal pollution markers following the same procedure described elsewhere (Bonjoch et al., 2004), with the addition of 0.4 µg μL^{-1} of bovine serum albumin (Madison, WI, USA) to all reactions. The amplicons were separated after electrophoresis on 2.5% agarose gels and visualized with ethidium bromide staining. Positive controls were performed using *B. adolescentis* DSM 20083^T and *B. dentium* DSM 20084^T DNA extracts and negative controls were included in all sets of reactions.

Table 1. PCR primers and probes used in this study

Primer/Probe	Sequence 5'–3'	Reference	
Mitochondrial markers		**Development of method**	**Specificity**
Pomito3-G	GGCCACATTAGCACTACTCAACATC	Martellini et al., 2005	Ballesté eta l. 2010
Pomito3-D	AGATCCGATGATTACGTGCAAC		
Pomito11-G	CTCTATACTCTTACTATCTCTAGGAC		
Pomito11-D	ATACGCCTAGTGCAATGGTGATGGA		
Bomito1-G	ACATACCCTTGATTGGACTAGCAT		
Bomito1-D	ATCACTACCCCTCAAACGCCTTCAAG		
Bomito11-G	GATTGGACTAGCATTAGCTGCAACC		
Bomito11-D	CTTGAAGGCGTTTGAGGGGTAGTGAT		
Ckmito1-G	ACCCTATTTGACTCCCTCAA	Kortbaoui et al., 2009	Kortbaoui et al., 2009
Ckmito1-D	ATGTCGACCAGGGGTTTATG		
CkmitoN1-G	CCCCCACACTAACAAGCAAT		
CkmitoN1-D	GGTTGTAAGGTGGTCGTGAT		
***B. adolescentis* and *B. dentium* multiplex**			
lm26	GATTCTGGCTCAGGATGAACG	Kaufmann et al., 1997	Kaufmann et al., 1997
lm3	CGGGTGCTICCCACTTTCATG		
Bi-ADO 1	CTCCAGTTGGATGCATGT	Bonjoch et al., 2004	Ballesté et al al. 2010
Bi-ADO 2	CGAAGGTTGCTCCCAGT		
Bi-DEN 1	ATCCCGGGGGTTCGCCT		
Bi-DEN 2	GAAGGGCTTGCTCCCGA		
Bifidobacterial host-specific qPCR			
Bif-F	TTCGGGTTGTAAACCGCTTTT	Gómez-Doñate et al., 2012	Gómez-Doñate et al., 2012
Bif-R	TACGTATTACCGCGGCTGCT		
HMprobe	VIC-TCGGGGTGAGTGTACCT-MGB		
PLprobe	FAM-GAGAGTGAGTGTACCCGTT-MGB		
PGprobe	FAM-CGCAAGTGAGTGTACCTT-MGB		
CWprobe	FAM-TTCGGCCGTGTTGAGT-MGB		

2.7 Bifidobacterial host-specific qPCR analysis

Human-, poultry-, cattle- and pig-specific bifidobacteria were quantified using four TaqMan® assays targeting different regions of the 16S rRNA gene using common primers Bif-F and Bif-R and specific probes HMprobe, PLprobe, CWprobe, and PGprobe (Table 1) and following the procedure described previously (Gómez-Doñate et al., 2012). Plasmid constructs containing the different four targets were prepared as described and used to prepare standard curves with concentrations ranging from 10^9 to 10^0 genomic copies per reaction, based on the construct size. All the samples and standards were run in duplicate and non-template controls were included in all the plates to discard the presence of contaminating DNA. Inhibition was minimized by using the Environmental Real-Time PCR Master Mix 2.0 (Applied Biosystems, Spain) specifically designed to reduce the impact of inhibitors in complex environmental samples (Cao et al., 2012). Additionally, dilutions for some negative samples were also tested to discard any inhibition effect. Inhibition for this qPCR assay was previously monitored in heavily polluted samples by using aliquots spiked with dilutions of the standards (Gómez-Doñate et al., 2012).

2.8 Statistical analysis

Analysis of variance tests (ANOVA) and the F-test were performed to compare the average concentrations of *E. coli*, enterococci and somatic coliphages between all samples. When the standard deviation of the compared samples was found to be statistically different according to Cochran's C test, Bartlett's test, Hartley's test and Levene's test, the comparison was performed using the Kruskal–Wallis test with the median concentrations.

All statistical analysis was performed using Statgraphics Plus software (version 5.1; Rockville, MD, USA).

2.9 Machine learning methods: Ichnaea modeling

The starting point for the Ichnaea system was a data matrix containing the set of MST indicators described above and the values obtained for those indicators in the samples of known fecal origin. The main characteristic of this data matrix (called the training matrix) was that all samples were collected at point source, i.e. high fecal pollution, and with no delay in the time between discharge and sampling (recent samples). Since each indicator presents differential decay in front of natural inactivation processes, and the relationship between indicators changes if the pollution occurred time ago and/or if the samples have suffered dilution, we considered therefore these aging factors and estimated "aged" samples versus "recent" samples. Given the training matrix, the system computed a number of predictive models for different dilution and aging factors (Bonjoch et al., 2009; Lasobras, 1997; Martellini et al., 2005), using several standard classifiers such as linear discriminant analysis, support vector machine, nearest neighbors and multinomial regression (Duda et al., 2000). Ichnaea created a bag of models (given by different fecal indicators and modeling technique) out of the training set. The confidence of each model was assessed by means of ten-fold cross-validation (Geisser, 1975). The system also estimated the importance of each indicator in the training matrix at different concentration levels, generating a heat map that presents all MST indicators sorted by the frequency with which they were selected by the best predictive models at each level of aging and dilution. Since persistence in the environment is dependent on season, the importance of indicators and MST methods in two different seasons (summer and winter) was taken into account

(Ballesté and Blanch, 2010; Bonjoch et al., 2011, 2009; Muniesa et al., 1999). Bifidobacterial host-specific qPCR indicators were not included in the generation of the heat maps because their concentration was below their limit of detection in all samples (Ebre Delta).

At prediction time, the new samples (in this case, the data matrices from the deltas of the Llobregat and Ebre rivers) were introduced in the system, which selected for prediction only the feasible models from the previously computed set of models. For every sample to be predicted, a model was feasible if every indicator required by the model had been measured (i.e., supplied) in that particular sample. An indicator was considered to be measured in a sample, regardless of the result obtained (positive or negative).

Once all of the selected models yielded their prediction, an overall prediction was determined by a simple majority vote. The output delivered included a prediction of the fecal source of the sample and a measure of the system's confidence in this prediction based on the percentage of models that predicted the majority vote.

3. RESULTS AND DISCUSSION

3.1 Samples of known fecal origin

In total, 15 samples from slaughterhouses were collected: five from poultry, five from cattle and five from pigs. Four more samples were collected from raw urban sewage (Table 2). The samples from pigs and humans showed the highest *E. coli*, enterococci and somatic coliphages concentrations ($p<0.05$). The concentrations in poultry samples were significantly lower than the human and pig samples ($p<0.05$). Cattle samples taken from

pooled feces (CW3 and CW4) showed high concentrations, statistically equivalent to those of the human and pig samples (p>0,005). However, the cattle samples taken from the slaughterhouse process water (CW1, CW2 and CW5) showed the lowest concentrations. Regarding the MST methods, the host-specific *Bacteroides* phages GA17, PL122 and CW18 strains were highly specific, recovering phages exclusively in human, poultry and cattle samples, respectively. On the other hand, PG76 presented false positive reactions in human samples, although the concentrations reported were significantly lower than in pig samples (p=0.016). In previous studies, PG76 strain was also detected in human samples at low concentrations (Gómez-Doñate et al., 2011) which confirms that this strain is not univocally related to pig fecal pollution. ADO-DEN multiplex was positive in all human samples but also in two poultry samples. The combination of ADO and DEN has been reported as strongly human specific, although some ADO false negatives have also been found in other poultry polluted samples (Blanch et al., 2006). The mitochondrial DNA markers correctly identified all samples of animal origin, although human samples 1 and 2 gave false positive results for pig and cattle markers. It has been reported that these mitochondrial markers could present false positives due to mitochondrial DNA contamination from non-fecal sources (skin, fur, hair, etc.), the potential meat carryover in human feces (Caldwell et al., 2011), or the PCR reagents (Leonard et al., 2007). With respect to the bifidobacterial host-specific qPCRs, they all showed high specificity and their concentrations were always higher than those of the general fecal indicators. Overall, apart from specificity issues in PG76 strain and ADO marker, most of the methods correctly identified the origin of fecal pollution in point-source samples. Positive and negative controls showed the expected results for all the indicators analyzed.

Table 2. Results of general fecal indicators and MST methods in samples of known fecal origin at source point. Host-specific markers for human: GA17, ADO-DEN and HMprobe; for poultry: PL122, CKMITO and PLprobe; for cow: CW18, BOMITO and CWprobe; and for pig: PG76, POMITO and PGprobe. "<" values represent the limits of quantification.

	General fecal indicators (CFU or PFU/100 ml)			Host-specific *Bacteroides* phages (PFU/100 ml)				ADO-DEN multiplex[a]		Mitochondrial markers[a]			*Bifidobacterial* host-specific qPCR (genomic copies per 100 ml)			
Sample	***E. coli***	**Enterococci**	**Somatic coliphages**	**GA17**	**PL122**	**CW18**	**PG76**	**ADO**	**DEN**	**POMITO**	**BOMITO**	**CKMITO**	**HMprobe**	**PLprobe**	**CWprobe**	**PGprobe**
Pig 1	2.34×10^{6}	1.45×10^{5}	3.33×10^{5}	<10	<10	<10	1.67×10^{3}	–	–	+	–	–	$<6.42\times10^{5}$	$<3.57\times10^{5}$	$<6.57\times10^{5}$	4.10×10^{9}
Pig 2	1.20×10^{7}	1.25×10^{6}	5.63×10^{7}	<10	<10	<10	1.40×10^{5}	–	–	+	–	–	$<6.42\times10^{5}$	$<3.57\times10^{5}$	$<6.57\times10^{5}$	1.00×10^{10}
Pig 3	1.35×10^{6}	5.50×10^{4}	2.84×10^{6}	<10	<10	<10	1.76×10^{4}	–	–	+	–	–	$<6.42\times10^{5}$	$<3.57\times10^{5}$	$<6.57\times10^{5}$	1.10×10^{9}
Pig 4	1.50×10^{7}	1.20×10^{5}	1.48×10^{7}	<10	<10	<10	3.00×10^{4}	–	–	+	–	–	$<6.42\times10^{5}$	$<3.57\times10^{5}$	$<6.57\times10^{5}$	3.00×10^{7}
Pig 5	1.85×10^{6}	1.00×10^{5}	9.30×10^{6}	<10	<10	<10	6.10×10^{4}	–	–	+	–	–	$<6.42\times10^{5}$	$<3.57\times10^{5}$	$<6.57\times10^{5}$	3.70×10^{7}
Poultry 1	5.50×10^{4}	2.30×10^{4}	5.50×10^{4}	<10	2.90×10^{3}	<10	<10	+	–	–	–	+	$<6.42\times10^{5}$	1.20×10^{8}	$<6.57\times10^{5}$	$<3.00\times10^{5}$
Poultry 2	7.40×10^{4}	3.00×10^{3}	7.40×10^{4}	<10	1.00×10^{3}	<10	<10	–	–	–	–	+	$<6.42\times10^{5}$	9.70×10^{7}	$<6.57\times10^{5}$	$<3.00\times10^{5}$
Poultry 3	2.34×10^{3}	7.70×10^{2}	2.34×10^{3}	<10	1.27×10^{2}	<10	<10	–	–	–	–	+	$<6.42\times10^{5}$	1.30×10^{8}	$<6.57\times10^{5}$	$<3.00\times10^{5}$
Poultry 4	2.70×10^{4}	1.87×10^{4}	2.70×10^{4}	<10	2.60×10^{3}	<10	<10	–	–	–	–	+	$<6.42\times10^{5}$	9.20×10^{7}	$<6.57\times10^{5}$	$<3.00\times10^{5}$
Poultry 5	7.40×10^{4}	4.30×10^{4}	7.40×10^{4}	<10	1.00×10^{3}	<10	<10	+	–	–	–	+	$<6.42\times10^{5}$	8.80×10^{7}	$<6.57\times10^{5}$	$<3.00\times10^{5}$
Cow 1	2.12×10^{2}	1.87×10^{2}	2.00×10^{1}	<10	<10	2.00×10^{1}	<10	–	–	–	+	–	$<6.42\times10^{5}$	$<3.57\times10^{5}$	1.30×10^{6}	$<3.00\times10^{5}$
Cow 2	9.97×10^{3}	7.65×10^{3}	6.22×10^{3}	<10	<10	2.00×10^{1}	<10	–	–	–	+	–	$<6.42\times10^{5}$	$<3.57\times10^{5}$	1.60×10^{5}	$<3.00\times10^{5}$
Cow 3	1.52×10^{7}	6.52×10^{6}	1.06×10^{7}	<10	<10	3.00×10^{1}	<10	–	–	–	+	–	$<6.42\times10^{5}$	$<3.57\times10^{5}$	9.80×10^{8}	$<3.00\times10^{5}$
Cow 4	1.02×10^{7}	8.60×10^{6}	7.40×10^{6}	<10	<10	2.86×10^{1}	<10	–	–	–	+	–	$<6.42\times10^{5}$	$<3.57\times10^{5}$	1.10×10^{9}	$<3.00\times10^{5}$
Cow 5	8.72×10^{3}	7.53×10^{3}	1.00×10^{2}	<10	<10	3.00×10^{1}	<10	–	–	–	+	–	$<6.42\times10^{5}$	$<3.57\times10^{5}$	3.10×10^{5}	$<3.00\times10^{5}$
Human 1	4.50×10^{6}	3.50×10^{6}	8.80×10^{5}	4.80×10^{3}	<10	<10	1.50×10^{3}	+	+	+	+	–	4.50×10^{7}	$<3.57\times10^{5}$	$<6.57\times10^{5}$	$<3.00\times10^{5}$
Human 2	6.00×10^{6}	3.75×10^{6}	1.90×10^{6}	2.35×10^{4}	<10	<10	3.00×10^{2}	+	+	–	–	–	2.40×10^{7}	$<3.57\times10^{5}$	$<6.57\times10^{5}$	$<3.00\times10^{5}$
Human 3	4.50×10^{6}	2.90×10^{6}	1.77×10^{6}	3.90×10^{4}	<10	<10	3.05×10^{3}	+	+	–	+	–	2.50×10^{7}	$<3.57\times10^{5}$	$<6.57\times10^{5}$	$<3.00\times10^{5}$
Human 4	4.25×10^{6}	2.80×10^{6}	2.52×10^{6}	9.20×10^{3}	1.00×10^{2}	<10	5.75×10^{3}	+	+	–	–	–	1.10×10^{7}	$<3.57\times10^{5}$	$<6.57\times10^{5}$	$<3.00\times10^{5}$

[a] +, positive signal; −, negative signal

3.2 Llobregat River samples

The results for Llobregat River are presented in Table S1. Somatic coliphages concentrations were significantly lower than those of the point-source samples (p=0.0001), except for the cattle slaughterhouse process water samples. GA17 phages were present in all samples, with concentrations one to three $\log_{10}$ units less than in urban sewage (p=0.017). Furthermore, the ADO marker was detected in all samples and the DEN marker was present in two out of five. The bifidobacterial qPCR human-specific target was also detected in all samples, but with concentrations two or three $\log_{10}$ units lower than in raw urban sewage. Pig-, cattle- and poultry-specific *Bacteroides* phages were detected at low concentrations in some samples. BOMITO mitochondrial DNA marker was detected in samples 1 and 3, and CKMITO in sample 2.

The Llobregat River mainly receives fecal contamination inputs of urban origin (treated sewage from the surrounding towns) (Lucena et al., 1988) and the indicators tested in this study confirmed that the main source of fecal pollution in the Llobregat River was human. Firstly, the ratio somatic coliphages / human-specific *Bacteroides* phages was higher than the ratios of other host-specific *Bacteroides* phages and somatic coliphages, as shown previously (Muniesa et al., 2012). These ratios have been proposed as a potential tool for identifying the origin of fecal pollution when multiple sources of fecal pollution are present in a sample (Muniesa et al., 2012). Secondly, all samples showed a positive PCR amplification of the human-associated indicator *B. adolescentis* (Field and Samadpour, 2007). Thirdly, the human-specific bifidobacterial qPCR was positive in all samples, but there was no signal for poultry-, pig-, or cattle-specific samples. However, the qualitative detection of some of mitochondrial indicators for cattle and poultry suggest that other

sources should not be ruled out, despite the potential limitations of mitochondrial markers described above.

3.3 Ebre Delta samples

All sampling sites presented similar concentrations of *E. coli*, enterococci and somatic coliphages ($p>0.05$) regardless of the sampling month, which indicate that the levels of fecal pollution were relatively homogeneous in the whole water irrigation channel system throughout the year (Table S2). The concentrations of these three fecal indicators were lower in all Ebre sites than in the point-source samples and the Llobregat River samples ($p<0.05$), with the exception of the cattle slaughterhouse process water samples. Interestingly most samples had concentrations of *E. coli* and enterococci within the values established by European Directive 2006/7/EC on bathing water quality that permit recreational human bathing (Anonymous, 2006).

As for the MST methods assayed, the ADO marker was detected in 13 out of 32 samples, while the DEN marker was only present in one, which was also positive for ADO marker. Cattle, pig and poultry mitochondrial DNA markers could be detected only in one sample each. Human- and poultry-specific *Bacteroides* phages were each detected in seven samples, but cattle-specific phages were only present in two samples and no pig-specific *Bacteroides* phages were detected. Finally, the bifidobacterial host-specific qPCRs gave negative results in all samples. In general, most MST indicators do not perform effectively because the concentration of their targets is below their limit of detection.

The Ebre Delta scenario is a good example of cases where the dilution or aging of the fecal pollution in water makes it difficult to determine the fecal source based on only one MST indicator, since many host-specific MST indicators individually, including the ones used in this study, do not provide enough information to form a conclusion. In these circumstances, a combination of different indicators is essential for determining the source of fecal pollution. The decision of which MST methods should be applied in each case should be taken under an objective basis, and for it machine learning methods may play a major role (1).

3.4 Machine learning modeling

The models generated by Ichnaea indicated that the main source of fecal pollution in the Llobregat River was human (Table 3). The confidence in this prediction (based on the percentage of models that predicted the majority vote) was well above 60% in all but one of the samples. In order to evaluate confidences, it has to be considered that there were four possible sources of fecal pollution in our approach, which means that the minimum theoretical possible confidence was 25%, and thus, an acceptable prediction should exceed this proportion. Predictions with confidences higher than 50% are considered to be good since more than half of the models generated agree with the same predicted source (Belanche-Muñoz and Blanch, 2008).

Table 3. Ichnaea predictions of the fecal pollution source of samples. Confidence is based on the percentage of models generated that predicted the origin of the majority of the pollution. LLOB: Llobregat River. S1-S8, sampling sites in the Ebre Delta.

Sample	Month	Predicted source	Confidence
LLOB1	January	Human	62.5%
LLOB2	February	Human	77.4%
LLOB 3	March	Human	53.9%
LLOB 4	April	Human	71.0%
LLOB 5	May	Human	78.6%
Site 1	May	Poultry	51.3%
	July	Poultry	34.1%
	September	Cattle	38.4%
	November	Poultry	45.9%
Site 2	May	Poultry	45.9%
	July	Cattle	32.9%
	September	Cattle	36.4%
	November	Poultry	58.9%
Site 3	May	Poultry	43.4%
	July	Poultry	30.4%
	September	Poultry	33.4%
	November	Human	56.5%
Site 4	May	Cattle	31.6%
	July	Poultry	53.7%
	September	Poultry	33.7%
	November	Poultry	47.1%
Site 5	May	Poultry	44.3%
	July	Poultry	36.5%
	September	Human	32.0%
	November	Poultry	60.5%
Site 6	May	Poultry	48.7%
	July	Poultry	46.4%
	September	Poultry	65.3%
	November	Poultry	60.2%
Site 7	May	Poultry	51.7%
	July	Poultry	57.6%
	September	Poultry	47.8%
	November	Poultry	71.1%
Site 8	May	Poultry	67.2%
	July	Poultry	63.2%
	September	Poultry	39.6%
	November	Poultry	61.1%

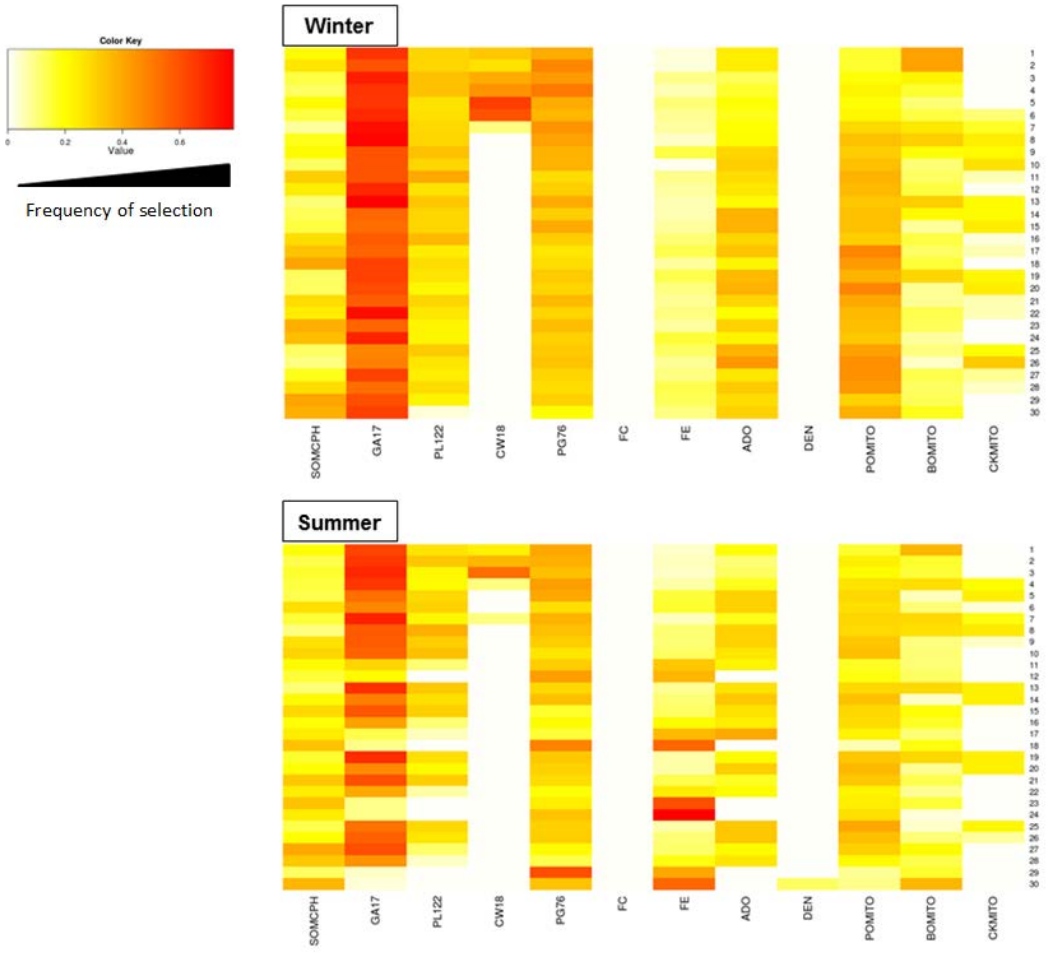
The outcome of Ichnaea in the Llobregat River system did not differ to the conclusions that would be reached by direct observation of the data. To test the suitability of the software, it is necessary to apply it in a more complicated system, in terms of low levels of contamination and diversity of the fecal sources. The next step was to test Ichnaea with the Ebre Delta samples, selected because they represent a complex system in which low-levels of fecal inputs from different sources will hamper a correct determination of the fecal source if directly evaluated.

In the Ebre Delta (Table 3), poultry was suggested to be the main fecal pollution source in 26 out of 32 samples. The remaining samples were predicted to contain mainly cattle (four out of 32) or human (two out of 32) fecal pollution. These results were consistent with the presence of a sewage treatment plant and a poultry farm nearby. The confidences in the prediction ranged from 30.4% to 67.2% which in general, were slightly lower compared with the Llobregat River. These predictions are valuable in a scenario of low fecal pollution because most of the MST methods individually could not provide interpretable information to indicate an origin of fecal pollution in the samples. Since the selection of the most informative MST methods should be taken, Ichnaea can produce objective predictions, even when the MST indicators are negative in a given sample. A negative result is as valid as any other result and can be used to classify the sample whenever the system has been properly trained. Nevertheless, it should be considered that a negative result is strongly dependent on the limit of detection of a given method, that if too low would make this method inapplicable. The results will also be strongly influenced by the volume of sample analyzed or whether the sample has been or not concentrated. If concentrated, the effectiveness of the concentration method will also be considered to evaluate the final

performance of the method. The choice of the right MST method-s, that has been controversial in most MST studies, appears therefore critical for a good MST assessment. The software Ichnaea intends to provide objective assessment for the selection of the methodology.

In this particular case, the heat maps generated by Ichnaea indicated that at high fecal pollution concentrations, the predictive models with the highest confidence values were based mostly on the phages infecting *Bacteroides* strains GA17, PG76, CW18 and BOMITO. However, the best models for waters with moderate or low levels of fecal pollution selected primarily GA17, PG76, ADO and POMITO (Figure 2).

Figure 2. Heat maps estimating the frequency of selection of fecal indicators in the training matrix at different concentration levels. The Y-axis values shows the different levels of increasing dilution and aging of samples calculated from the decay values of each indicator. Values being the top rows correspond to fresh, highly concentrated samples and the bottom ones to more diluted and aged samples. Color key: red corresponds to those indicators more frequently selected by the predictive models generated for the training matrix in a given level of sample dilution and aging. White denotes less frequently selected indicators.



It should be noted that these results showing the most suitable methods would have been different if other MST methods would have been selected to provide data to Ichnaea. Therefore, the information provided by Ichnaea will be strongly dependent on the sort and the amount of data used to train the system, the limit of detection of the method used and the volumes analyzed.

One limitation of the approach applied is that at the moment Ichnaea can only point out a main origin of fecal pollution in a given sample. In scenarios such as the Llobregat River, where a particular origin is the major contributor to the fecal load, the confidences in the

predictions are high and coincident with a direct evaluation of the data. However, in the Ebre Delta, or similar systems, when multiples sources in different proportions might be contributing to the total fecal pollution load, the confidences are substantially lower and represent the real problem for MST assessment. In this scenario Ichnaea could provide an objective prediction of the main origin based in the data provided and generate models that determine the selection of the most suitable MST methods. In these scenarios, however, the presence of other important contributions in addition to the main fecal source could be relevant. In further studies Ichnaea should be improved to include the possibility of simultaneous multiple origins in a sample.

Another limitation of this study is that the system could only predict those origins for which it has been trained, but other fecal pollution sources might be present in the scenario (i.e. wild birds in the Ebre Delta). The MST methods used covered the most common fecal pollution sources potentially present in the area, but unfortunately, no MST method specific for wild birds was available, and thus, it could not be included in our panel. One of the future directions is the design new indicators for wild animals, particularly birds to improve the range of fecal origins to be predicted by Ichnaea.

3.5 Conclusions

The concentration of MST indicators at point source, their die-off in the environment and the detection limits of their enumeration methodology are key factors in their feasibility and reliability in MST. Most culture-dependent or molecular MST indicators proposed over recent years present major limitations in waters with low levels of fecal pollution (below 3 $\log_{10}$ CFU *E. coli*/100 ml) (Hagedorn et al., 2011). Under these conditions the software

Ichnaea performed effectively, using results obtained from samples of known fecal source as training data and combining several MST indicators to objectively propose with reasonably good confidences the main origin of fecal pollution. Therefore, despite some limitations that should be addressed in further studies, inductive machine learning methods are a promising tool and may represent a leap forward in solving the MST problem.

ACKNOWLEDGEMENTS

This work was supported by the Spanish government, under research project number CGL2011-25401, and the research programs of the Catalan Biotechnology Reference Network (XRB) and the Commission for Universities and Research of the Catalan Ministry of Innovation, Universities and Enterprise (DIUE), under grant reference 2009SGR1043. M. Gómez-Doñate is a recipient of an FI grant from the government of Catalonia under reference 2009SGR1043. The authors thank the slaughterhouses for providing the samples used in this study.

REFERENCES

Anonymous, 2006. Directive 2006/7/EC of the European Parliament and of the Council of 15 February 2006 concerning the management of bathing water quality and repealing Directive 76/160/EEC.

Ballesté, E., Blanch, A.R., 2010. Persistence of *Bacteroides* species populations in a river as measured by molecular and culture techniques. Appl. Environ. Microbiol. 76, 7608–16. doi:10.1128/AEM.00883-10

Balleste, E., Bonjoch, X., Belanche, L.A., Blanch, A.R., 2010. Molecular indicators used in the development of predictive models for microbial source tracking. Appl. Environ. Microbiol. 76:1789-95.

Belanche-Muñoz, L., Blanch, A.R., 2008. Machine learning methods for microbial source tracking. Environ. Model. Softw. 23, 741–750. doi:10.1016/j.envsoft.2007.09.013

Blanch, A.R., Belanche-Muñoz, L., Bonjoch, X., Ebdon, J., Gantzer, C., Lucena, F., Ottoson, J., Kourtis, C., Iversen, A., Kühn, I., Mocé, L., Muniesa, M., Schwartzbrod, J., Skraber, S., Papageorgiou, G.T., Taylor, H., Wallis, J., Jofre, J., 2006. Integrated analysis of established and novel microbial and chemical methods for microbial source tracking. Appl. Environ. Microbiol. 72, 5915–26. doi:10.1128/AEM.02453-05

Bonjoch, X., Ballesté, E., Blanch, A.R., 2004. Multiplex PCR with 16S rRNA gene-targeted primers of *Bifidobacterium spp*. to identify sources of fecal pollution. Appl. Environ. Microbiol. 70, 3171–5.

Bonjoch, X., García-Aljaro, C., Blanch, A.R., 2011. Persistence and diversity of faecal coliform and enterococci populations in faecally polluted waters. J. Appl. Microbiol. 111, 209–15. doi:10.1111/j.1365-2672.2011.05028.x

Bonjoch, X., Lucena, F., Blanch, A.R., 2009. The persistence of bifidobacteria populations in a river measured by molecular and culture techniques. J. Appl. Microbiol. 107, 1178–85. doi:10.1111/j.1365-2672.2009.04297.x

Brion, G.M., Lingireddy, S., 2003. Artificial neural network modelling: a summary of successful applications relative to microbial water quality. Water Sci. Technol. 47, 235–40.

Caldwell, J., Payment, P., Villemur, R., 2011. Mitochondrial DNA as Source Tracking Markers of Fecal Contamination, in: Hagedorn, C., Blanch, A.R., Harwood, V.J. (Eds.), Microbial Source Tracking: Methods, Applications, and Case Studies. Springer, New York, pp. 229–250.

Cao, Y., Griffith, J.F., Dorevitch, S., Weisberg, S.B., 2012. Effectiveness of qPCR permutations, internal controls and dilution as means for minimizing the impact of inhibition while measuring *Enterococcus* in environmental waters. J. Appl. Microbiol. 113, 66–75.

Clescerl, L.S., Greenberg, A.E., Eaton, A.D. (Eds.), 1998. Standard Methods for the Examination of Water and Wastewater, 20th editi. ed. American Public Health Association, Washington DC.

Duda, R.O., Hart, P.E., Stork, D.G., 2000. Pattern Classification, 2nd edition. ed. John Wiley and Sons, New York.

Field, K.G., Chern, E.C., Dick, L.K., Fuhrman, J., Griffith, J., Holden, P.A., LaMontagne, M.G., Le, J., Olson, B., Simonich, M.T., 2003. A comparative study of culture-independent, library-independent genotypic methods of fecal source tracking. J. Water Health 1, 181–94.

Field, K.G., Samadpour, M., 2007. Fecal source tracking, the indicator paradigm, and managing water quality. Water Res. 41, 3517–38. doi:10.1016/j.watres.2007.06.056

Figueras, M.J., Inza, I., Polo, F., Guarro, J., 1998. Evaluation of the oxolinic acid--esculin--azide medium for the isolation and enumeration of faecal streptococci in a routine monitoring programme for bathing waters. Can. J. Microbiol. 44, 998–1002.

Geisser, S., 1975. The Predictive Sample Reuse Method with Applications. J. Am. Stat. Assoc. 70, 320–328. doi:10.1080/01621459.1975.10479865

Gómez-Doñate, M., Ballesté, E., Muniesa, M., Blanch, A.R., 2012. New molecular quantitative PCR assay for detection of host-specific Bifidobacteriaceae suitable for microbial source tracking. Appl. Environ. Microbiol. 78, 5788–95. doi:10.1128/AEM.00895-12

Gómez-Doñate, M., Payán, A., Cortés, I., Blanch, A.R., Lucena, F., Jofre, J., Muniesa, M., 2011. Isolation of bacteriophage host strains of *Bacteroides* species suitable for tracking sources of animal faecal pollution in water. Environ. Microbiol. 13, 1622–31. doi:10.1111/j.1462-2920.2011.02474.x

Gourmelon, M., Caprais, M.P., Ségura, R., Le Mennec, C., Lozach, S., Piriou, J.Y., Rincé, A., 2007. Evaluation of two library-independent microbial source tracking methods to identify sources of fecal contamination in French estuaries. Appl. Environ. Microbiol. 73, 4857–66. doi:10.1128/AEM.03003-06

Griffith, J.F., Weisberg, S.B., McGee, C.D., 2003. Evaluation of microbial source tracking methods using mixed fecal sources in aqueous test samples. J. Water Health 1, 141–51.

Hagedorn, C., Blanch, A.R., Harwood, V.J. (Eds.), 2011. Microbial Source Tracking: Methods, Applications, and Case Studies. Springer New York, New York, NY. doi:10.1007/978-1-4419-9386-1

Harwood, V.J., Wiggins, B., Hagedorn, C., Ellender, R.D., Gooch, J., Kern, J., Samadpour, M., Chapman, A.C.H., Robinson, B.J., Thompson, B.C., 2003. Phenotypic library-based microbial source tracking methods: efficacy in the California collaborative study. J. Water Health 1, 153–66.

International Organization for Standardization, 1994. ISO 5667/3:1994 - Guidance on the Preservation and Handling of Samples. Geneva, Switzerland.

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

International Organization for Standardization, 2001. ISO 10705-4. Water quality. Detection and enumeration of bacteriophages. Part 4. Enumeration of bacteriophages infecting *Bacteroides fragilis*. Geneva, Switzerlan.

Kaufmann, P., Pfefferkorn, A., Teuber, M., Meile, L., 1997. Identification and quantification of *Bifidobacterium* species isolated from food with genus-specific 16S rRNA-targeted probes by colony hybridization and PCR. Appl. Environ. Microbiol. 63, 1268–73.

Kortbaoui, R., Locas, A., Imbeau, M., Payment, P., Villemur, R., 2009. Universal mitochondrial PCR combined with species-specific dot-blot assay as a source-tracking method of human, bovine, chicken, ovine, and porcine in fecal-contaminated surface water. Water Res. 43, 2002–10. doi:10.1016/j.watres.2009.01.030

Lasobras, J., 1997. Relationship between the morphology of bacteriophages and their persistence in the environment. Water Sci. Technol. 35, 129–132. doi:10.1016/S0273-1223(97)00247-3

Leonard, J.A., Shanks, O., Hofreiter, M., Kreuz, E., Hodges, L., Ream, W., Wayne, R.K., Fleischer, R.C., 2007. Animal DNA in PCR reagents plagues ancient DNA research. J. Archaeol. Sci. 34, 1361–1366.

Lucena, F., Bosch, A., Ripoll, J., Jofre, J., 1988. Fecal pollution in llobregat river: Interrelationships of viral, bacterial, and physico-chemical parameters. Water. Air. Soil Pollut. 39, 15–25. doi:10.1007/BF00250944

Manero, A., Blanch, A.R., 1999. Identification of *Enterococcus spp.* with a biochemical key. Appl. Environ. Microbiol. 65, 4425–30.

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

Moore, D.F., Harwood, V.J., Ferguson, D.M., Lukasik, J., Hannah, P., Getrich, M., Brownell, M., 2005. Evaluation of antibiotic resistance analysis and ribotyping for identification of faecal pollution sources in an urban watershed. J. Appl. Microbiol. 99, 618–28. doi:10.1111/j.1365-2672.2005.02612.x

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

Muniesa, M., Lucena, F., Jofre, J., 1999. Study of the potential relationship between the morphology of infectious somatic coliphages and their persistence in the environment. J. Appl. Microbiol. 87, 402–409. doi:10.1046/j.1365-2672.1999.00833.x

Myoda, S.P., Carson, C.A., Fuhrmann, J.J., Hahm, B.-K., Hartel, P.G., Yampara-Lquise, H., Johnson, L., Kuntz, R.L., Nakatsu, C.H., Sadowsky, M.J., Samadpour, M., 2003. Comparison of genotypic-based microbial source tracking methods requiring a host origin database. J. Water Health 1, 167–80.

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 contamination 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, 5659–62. doi:10.1128/AEM.71.9.5659-5662.2005

Ritter, L., Solomon, K., Sibley, P., Hall, K., Keen, P., Mattu, G., Linton, B., 2002. Sources, pathways, and relative risks of contaminants in surface water and groundwater: a perspective prepared for the Walkerton inquiry. J. Toxicol. Environ. Health. A 65, 1–142.

Rubiano, M.-E., Agulló-Barceló, M., Casas-Mangas, R., Jofre, J., Lucena, F., 2012. Assessing the effects of tertiary treated wastewater reuse on a Mediterranean river (Llobregat, NE Spain): pathogens and indicators [corrected]. Environ. Sci. Pollut. Res. Int. 19, 1026–32. doi:10.1007/s11356-011-0562-9

Samadpour, M., Roberts, M.C., Kitts, C., Mulugeta, W., Alfi, D., 2005. The use of ribotyping and antibiotic resistance patterns for identification of host sources of *Escherichia coli* strains. Lett. Appl. Microbiol. 40, 63–8. doi:10.1111/j.1472-765X.2004.01630.x

Sánchez, D., Belanche-Muñoz, L.A., Blanch, A.R., 2011. A Software System for the Microbial Source Tracking Problem. J. Mach. Learn. Res. 17, 56–62.

Santo Domingo, J.W., Bambic, D.G., Edge, T.A., Wuertz, S., 2007. Quo vadis source tracking? Towards a strategic framework for environmental monitoring of fecal pollution. Water Res. 41, 3539–52. doi:10.1016/j.watres.2007.06.001

Savichtcheva, O., Okabe, S., 2006. Alternative indicators of fecal pollution: relations with pathogens and conventional indicators, current methodologies for direct pathogen monitoring and future application perspectives. Water Res. 40, 2463–76. doi:10.1016/j.watres.2006.04.040

Simpson, J.M., Santo Domingo, J.W., Reasoner, D.J., 2002. Microbial source tracking: state of the science. Environ. Sci. Technol. 36, 5279–88.

Stoeckel, D.M., Mathes, M. V, Hyer, K.E., Hagedorn, C., Kator, H., Lukasik, J., O'Brien, T.L., Fenger, T.W., Samadpour, M., Strickler, K.M., Wiggins, B.A., 2004. Comparison of seven protocols to identify fecal contamination sources using *Escherichia coli*. Environ. Sci. Technol. 38, 6109–17.