Journal of Water & Health

Journal of Water and Health Vol 19 No 5, 762 doi: 10.2166/wh.2021.119

Microbial source tracking of fecal contamination in Laguna Lake, Philippines using the library-dependent method, rep-PCR

Laurice Beatrice Raphaelle O. dela Peña, Kevin L. Labrador, Mae Ashley G. Nacario, Nicole R. Bolo and Windell L. Rivera*

Pathogen-Host-Environment Interactions Research Laboratory, Institute of Biology, College of Science, University of the Philippines Diliman, Quezon City 1101, Philippines

*Corresponding author. E-mail: wlrivera@science.upd.edu.ph

LBROdP, 0000-0003-2218-4798; KLL, 0000-0002-3309-8062; MAGN, 0000-0002-3723-0403; NRB, 0000-0002-7629-4020; WLR, 0000-0002-4659-6132

ABSTRACT

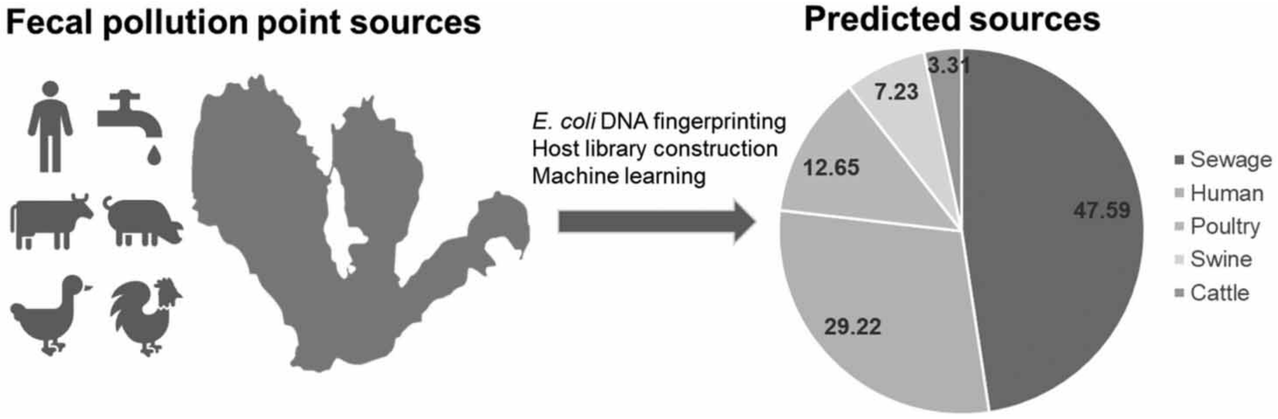
Laguna Lake is an economically important resource in the Philippines, with reports of declining water quality due to fecal pollution. Currently, monitoring methods rely on counting fecal indicator bacteria, which does not supply information on potential sources of contamination. In this study, we predicted sources of *Escherichia coli* in lake stations and tributaries by establishing a fecal source library composed of rep-PCR DNA fingerprints of human, cattle, swine, poultry, and sewage samples ($n = 1,408$). We also evaluated three statistical methods for predicting fecal contamination sources in surface waters. Random forest (RF) outperformed k-nearest neighbors and discriminant analysis of principal components in terms of average rates of correct classification in two- (84.85%), three- (82.45%), and five-way (74.77%) categorical splits. Overall, RF exhibited the most balanced prediction, which is crucial for disproportionate libraries. Source tracking of environmental isolates ($n = 332$) revealed the dominance of sewage (47.59%) followed by human sources (29.22%), poultry (12.65%), swine (7.23%), and cattle (3.31%) using RF. This study demonstrates the promising utility of a library-dependent method in augmenting current monitoring systems for source attribution of fecal contamination in Laguna Lake. This is also the first known report of microbial source tracking using rep-PCR conducted in surface waters of the Laguna Lake watershed.

Key words: DNA fingerprinting, Laguna Lake, microbial source tracking, random forest, rep-PCR

HIGHLIGHTS

- DNA fingerprinting of *E. coli*, coupled with machine learning algorithms, can be used to discriminate fecal pollution sources in Laguna Lake, Philippines.
- The majority of *E. coli* isolates can be attributed to sewage contamination, followed by human and agricultural sources.
- Source-tracking methods can empower local agencies responsible for water quality management to minimize public health and economic risks.

GRAPHICAL ABSTRACT



INTRODUCTION

Laguna Lake (Laguna de Bay) is the largest freshwater lake in the Philippines and is considered as the most economically important lake in the country (Santos-Borja & Nepomuceno 2006). In the last quarter of 2020, the overall water quality of the lake conformed with the Class B classification of the country's Department of Environment and Natural Resources (DENR) Water Quality Guidelines No. 2016-08 in terms of fecal coliform (Laguna Lake Development Authority 2020), making it utilizable for aquaculture, transportation, recreation, and water source for power generation, farm irrigation, industrial cooling, and even for domestic use (Vargas 2015).

However, despite the lake's relatively good water quality, the majority of its 28 tributary rivers failed to conform to the water quality guidelines for Classes A–D in terms of fecal coliform (Laguna Lake Development Authority 2019). Pollution brought about by population growth, industrialization, and land use may cause the lake's deterioration, especially when polluted water from its tributary rivers flows directly into the lake. This poses risks to the estimated 16 million residents within its vicinity (Santos-Borja & Nepomuceno 2006). Strategies to mitigate the lake's degradation include regular monitoring of the water quality in terms of physicochemical properties as well as fecal coliform count, which is performed by the Laguna Lake Development Authority (LLDA). However, to efficiently prevent the water bodies' further degradation, identifying the origin of contamination is likewise essential.

In the Philippines, microbial source tracking (MST) is a relatively new discipline for identifying the sources of fecal contamination in the environment. MST studies involve library-dependent methods (LDMs) which include a range of fecal indicator source-tracking techniques and require the construction of a library of known source profiles that are used for comparison with environmental isolates (Stoeckel & Harwood 2007; Ahmed *et al.* 2008, 2009). DNA fingerprinting or rep-PCR of fecal indicator bacteria (FIB), such as *Escherichia coli*, is one of the most used LDM for MST (Mohapatra *et al.* 2007; Edge *et al.* 2010; Lyautey *et al.* 2010; Carlos *et al.* 2012) and was proven to be effective in a preliminary study conducted by Labrador *et al.* (2020). The said study aimed to optimize the most effective combination among three rep-PCR primers (REP, BOX, and GTG) for MST in Laguna Lake. The discriminant analysis of principal component (DAPC) was used as the classification tool to construct the library composed of the combination of BOX and GTG fingerprints of *E. coli* isolated from different known domestic and agricultural sources. A relatively high average rate of correct classification (ARCC) was achieved. However, further expansion of the library yielded a disproportionate dataset, possibly creating a bias toward sources which have more representatives than the others.

To address this limitation, other classification tools can also be considered for the construction of a DNA fingerprinting library. It has been reported that machine learning models such as k-nearest neighbors (kNN) and random forest (RF) classifiers are able to give more accurate or balanced predictions when using imbalanced source-tracking libraries (Robinson *et al.* 2007; Smith *et al.* 2010; Carlos *et al.* 2012). In this study, three statistical models, DAPC, kNN, and RF, were evaluated for the classification of environmental samples into different domestic and agricultural sources using the constructed rep-PCR library. The models were then used to predict sources of fecal contamination in Laguna Lake and eight of its tributary rivers. Results of this research can help in evaluating public health risks, identifying sources of pollution, and as science-based supporting information for policy-making and implementation for the rehabilitation and improvement of Laguna Lake and its tributaries.

MATERIALS AND METHODS

Site description

Laguna Lake has a total surface area of 900 km^2 and an estimated holding capacity of 2.19 billion m^3. Its watershed area of 3,820 km^2 expands along the provinces of Rizal and Laguna, some towns in Batangas and Cavite, and some cities in Metro Manila (Santos-Borja & Nepomuceno 2006). Stations in Laguna Lake and its tributaries were selected using the monitoring information of the LLDA (Figure 1). From a total of nine lake stations and 27 tributaries monitored by the agency, three lake stations and eight tributaries were shortlisted on the basis of having (1) the highest coliform count and (2) the greatest concentration of possible host sources (e.g., farms near the tributaries). For the lake stations, these were the East Bay (LS2), Northwest Bay (LS5), and South Bay (LS8). These stations are near the mouths of the tributaries, hence, the high levels of fecal contamination. For the tributaries, the following were selected to represent regions adjacent to the lake: Bagumbayan (TR1), Mangangate (TR2), and Tunasan (TR4) in the National Capital Region (NCR); Sapang Baho (TR3) in Rizal; and Biñan (TR5), Pila (TR6), San Cristobal (TR7), and Sta. Rosa (TR8) in Laguna. Land-use assessment in the vicinity of the lake reveals

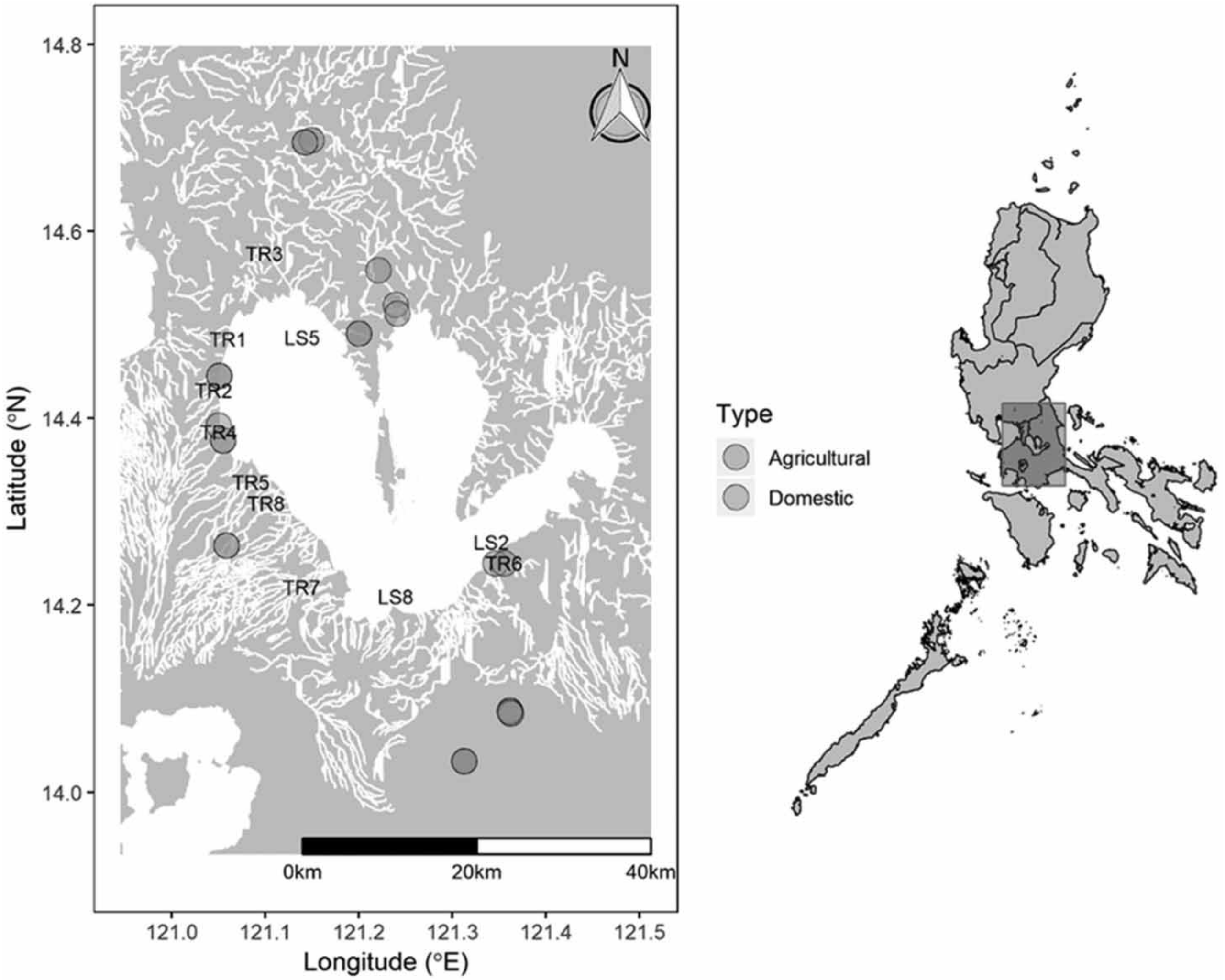


Figure 1 | Sampling location. Agricultural (red; feces from chickens, cattle, and swine) and domestic (blue; sewage and human feces) sources were used to construct the host source library which was used to source track thermotolerant *E. coli* isolates obtained from the environment (LS, Laguna Lake stations; TR, tributaries). Inset shows the geographical location relative to the island of Luzon. Please refer to the online version of this paper to see this figure in colour: doi: http://dx.doi.org/10.2166/wh.2021.119.

wide built-up areas around the northern tributaries (TR1–TR4) as they are situated in Metro Manila. The southern tributaries (TR5–8) have fewer industrialized areas and are surrounded by agricultural lands but are increasingly being converted to residential and commercial areas (Tanganco *et al.* 2019).

After the environmental sites were selected, nearby locations of possible host sources were identified. Domestic host sources were obtained from sewage treatment facilities of Manila Water and Laguna Water, as well as from sewage waters of Metro Manila draining toward Laguna Lake. In addition, human feces were collected from municipalities near the tributaries. These were Barangay Cupang, Muntinlupa, and the rural health unit of the municipality of Pila, Laguna. Human feces were obtained with clearance from the University of the Philippines Manila Research Ethics Board (UPMREB code: 2018-356-01). Meanwhile, agricultural host sources were collected in regions that focused on agricultural activities. These were Rizal and Laguna, located in the northern and southern boundaries of the lake, respectively. Fecal samples were collected from small backyard farms, specialized farms (e.g., piggeries, poultry farms, and cattle farms), and pastures.

Sample collection

Samples were categorized as (1) environmental (i.e., water from lake and tributaries), (2) agricultural (feces from chickens, ducks, cattle, and swine), and (3) domestic (sewage and human feces). Water and sewage samples (1 L) were obtained from the surface using grab sampling and were transferred to sterile wide-mouth water bottles (Nalgene, USA). Meanwhile, fecal samples were collected using a spatula and then transferred into either a stool container or a polypropylene bag. All samples were stored in ice and were transported to the laboratory for processing within 24 h after collection. Sample collection was done monthly from July 2017 to July 2019 during the wet and dry seasons.

Isolation and characterization of thermotolerant *E. coli*

Samples were serially diluted in sterile conical tubes using 0.9% saline solution as the diluent. Prior to serial dilution, fecal samples from each host source at a given site were pooled and 10 g were aseptically transferred into a sterile flask containing the diluent. After vigorous mixing, the mixture was serially diluted up to 10^{-7}. Selected dilutions (human and animal feces = 10^{-7} to 10^{-10}, sewage = 10^{-4} to 10^{-7}, tributary water = 10^{-4} to 10^{-7}, lake water = undiluted), done in duplicates, were filtered through a GN-6 Metricel membrane (47 mm diameter, 0.45 μm pore size; Pall Corp., USA) using a vacuum pump (Millipore, USA). The membrane filters were placed on modified membrane-thermotolerant *E. coli* (mTEC) agar (BD Difco, USA) and incubated at 37 °C for 2 h, then to 42 °C for 18–24 h. Presumptive *E. coli* isolates, characterized by blue to violet colonies on mTEC plates, were further streaked in eosin methylene blue agar (EMBA; BD BBL, USA) for confirmation. All isolates that exhibited a green metallic sheen on EMBA were selected for DNA extraction and molecular fingerprinting.

DNA extraction

DNA from thermotolerant *E. coli* was extracted using the boil lysis extraction method (Garcia *et al.* 2015). Briefly, *E. coli* grown on trypticase soy broth (TSB; BD BBL, USA) for 18–24 h were harvested through centrifugation (10,000 × g, 10 min). The harvested pellet was washed with sterile distilled water (1,000 μL), eluted (100 μL), and then heated (100 °C, 15 min). Afterwards, the sample was centrifuged (10,000 × g, 15 min), and the supernatant (50 μL) which served as a template for fingerprinting was collected in a fresh microtube.

DNA fingerprinting

A library of composite profiles from BOX-A1R (5′-CTACGGCAAGGCGACGCTGACG-3′) and $(GTG)_5$ (5′-GTGGTGGTGGTGGTG-3′) (hereafter referred to as BOX and GTG, respectively) was reported to have the highest ARCC for source-tracking fecal contamination in Laguna Lake (Labrador *et al.* 2020); hence, these two markers were used for fingerprinting. The PCR mixture (10 μL) for each primer was composed of the following: GoTaq Green Master Mix (1 × , Promega, USA), primer (1 μM), template DNA (1 μL), and appropriate amount of PCR water. The PCR was performed with similar cycling conditions for both primers (Kheiri & Akhtari 2017): initial denaturation (94 °C, 5 min); 30 cycles of denaturation (94 °C, 20 s), annealing (52 °C, 30 s), and extension (72 °C, 1 min); and final extension (72 °C, 10 min). A no-template control was included in every run. The resulting amplicons were subjected to agarose gel electrophoresis (2%, 190 V, 70 min), and the gels were visualized using a gel documentation system (Bio-print ST4, Vilber Lourmat, UK). Analysis of banding patterns was done with an imaging system (SuperMega ST4 v.16.08 g, Vilber Lourmat, UK). Band positions were normalized using a 1 kb molecular ladder (Hyperladder, Bioline, USA) as an external reference. Only a single observer performed the gel analysis to minimize variability attributed to multiple observers.

Source tracking

Analysis of DNA fingerprints, construction of host library, and source tracking were done using the programming language, R v.3.6.3 (R Core Team 2020), following Labrador *et al.* (2020) with modifications. Band positions were binned by rounding up their molecular weight to the nearest 50 bp. These were then converted into a binary sequence based on their presence (1) or absence (0) across samples. Binary sequences from both primers were integrated to form a composite profile to be used in the analysis. Isolates from a specific host source having identical profiles were collapsed into a single observation. The resulting data were partitioned into two sets: (1) the training dataset, which contained profiles from agricultural and domestic sources; and (2) the unknown dataset, which contained profiles from environmental sources. The training dataset was used to create the library. This was prepared depending on how the sources were categorized: (1) a two-way split had agricultural–domestic; (2) a three-way split had agricultural–human–sewage; and lastly, (3) a five-way split had cattle–poultry–swine–human–sewage.

Models for the source classifier were constructed using three different statistical techniques, namely, DAPC, kNN, and RF. Library accuracy was externally assessed by holdout of 20% of the library into a 'challenge' or test dataset which was excluded from model training. Samples were chosen using stratified sampling in order to preserve overall class distribution. A 10-fold cross-validation technique repeated five times was adopted for model training. In order to improve the prediction bias resulting from the class imbalance in the library, an oversampling technique was used, as implemented in the package *caret*. For comparative evaluation of classifiers, the subsampling techniques were maintained across all three methods; thus, the exact same data points were used for training.

DAPC was performed as implemented in the package, *adegenet* (Jombart 2008). The number of PCs with the least mean square error was carried over for discriminant analysis. Both kNN and RF were implemented using the package *caret* (Kuhn 2008). For RF, forest sizes ranging from 500 to 2,500 were explored, but accuracy rates were not significantly different (data not shown), thus, all analyses subsequently used 500 trees. The number of randomly selected predictors (mtry) for RF and the number of neighbors to consider (k) for kNN were optimized at each categorical split scheme using accuracy as the metric.

A confusion matrix between the observed and predicted categories was then created to calculate for the ARCC of the library. We explored how the disproportionate library affects the ARCC by calculating the percentage of known samples incorrectly predicted (% IP) as sewage for each method using the five-way split. To see how class imbalance affects the % IP in each model, sewage isolates were sampled at intervals of 100, starting with 200 up to 600.

Afterwards, the library was used to categorize the isolates in the unknown dataset. The posterior probability of each isolate to belong to a defined category was calculated. The category with the highest posterior probability was considered as the most probable identity of an isolate. Once environmental isolates were categorized, the percent composition of contamination of each source was determined.

RESULTS

Library performance

The reference library consisted of 1,408 thermotolerant *E. coli* isolates, 444 of which were from agricultural sources while 964 were from domestic sources (Table 1). These isolates were placed on pre-defined categories, and DAPC, RF, and kNN were used to assess library performance in classification.

Overall, good classification was obtained using the BOX–GTG composite profiles in either a two-way or three-way categorical split using DAPC and RF (Table 2). In terms of library accuracy, ARCC decreased as the number of categories increased. Comparably high accuracies were calculated for both two-way (RF ARCC = 84.35%; DAPC ARCC = 82.92%) and three-way (RF ARCC = 82.45%; DAPC ARCC = 81.77%) libraries using RF and DAPC. Meanwhile, accuracy was lower for the five-way library (RF ARCC = 74.77%; DAPC ARCC = 74.55%). Moreover, kNN yielded the lowest ARCC for all categorical schemes (two-way = 80.41%, three-way = 73.12%, five-way = 68.48%). Human sources had a consistently high relative rate of correct classification where it was used as a category in DAPC and RF classification (RCC = 92.13–93.19%). This was followed by sewage (RCC = 77.92–84.9%). Agricultural isolates were successfully classified when lumped together in a single category (RCC = 73.32–79.32%), further partitioning them into exclusive subcategories increased the probability of misclassification.

Table 1 | A total of 1,408 thermotolerant *E. coli* isolates from five different sources were used to construct the library

Category/source	Sample code	No. of isolates
Agricultural	**AGR**	**444**
Chicken/Duck	POU	167
Cattle	CAT	135
Swine	SWI	142
Domestic	**DOM**	**964**
Human	HUM	344
Sewage	SWG	620
Total		**1,408**

Bold indicates total number of samples in each category.

Table 2 | ARCC decreased as the number of categories increased, using DAPC, kNN, and RF

	Rate of correct classification (%) ± 95% CI		
Categorical split	DAPC	kNN	RF
Two-way ARCC	**82.92 ± 1.94**	**80.41 ± 2.05**	**84.35 ± 1.87**
DOM	86.00	81.96	86.65
AGR	75.19	77.05	79.32
Three-way ARCC	**81.77 ± 1.99**	**73.12 ± 2.31**	**82.45 ± 1.96**
SWG	81.90	75.65	80.68
HUM	93.19	55.1	92.13
AGR	73.32	83.64	77.5
Five-way ARCC	**74.55 ± 2.27**	**68.48 ± 2.43**	**74.77 ± 2.26**
SWG	77.92	90.75	84.9
HUM	92.68	70.26	93
CAT	49.44	31.97	45.9
SWI	50.49	29.93	45.26
POU	54.24	43.09	47.51

Bold indicates overall ARCC in each categorical scheme.

Library validation

An external validation of the library was conducted using a library 'challenge' which was composed of known source isolates ($n = 278$) excluded from the training set. The accuracy obtained for the training set (80%) was comparable to ARCCs derived when using the full library (Table 3). Similar to the whole library models, DAPC and RF yielded high ARCCs for both two- and three-way splits. McNemar's test shows no significant difference between DAPC and RF in test classification at all three splits ($p > 0.05$). In contrast, kNN held significantly lower ARCC at all categorical splits (McNemar's test RF vs. kNN, $p < 0.05$). RF had the highest ARCCs for test set results, ranging from 75.90 to 83.87%.

To see how class imbalance affects the accuracy of classification, we looked at the trend of the percentage of samples incorrectly predicted (% IP) as sewage as we populated the library with an increasing number of sewage samples (Figure 2). In general, a higher % IP can be seen in agricultural categories when increasing the sewage library using DAPC or kNN. A lower % IP was achieved by RF compared to DAPC and kNN, showing more balanced composition. Human classification tends to be more robust and has lower % IP which may indicate sufficient distinction from sewage, supporting the high RCC rates previously achieved.

Table 3 | ARCC achieved using DAPC, kNN, and RF using 80% of the library (training set) and the remaining 20% (test set)

	ARCC (%) ± 95% CI					
	DAPC		**kNN**		**RF**	
Categorical split	**Training set**	**Test set**	**Training set**	**Test set**	**Training set**	**Test set**
Two-way ARCC	**83.84 ± 2.11**	**79.93**	**78.57 ± 2.37**	**76.34**	**83.84 ± 2.11**	**83.87**
DOM	86.92	87.96	80.21	78.53	76.42	86.91
AGR	76.31	62.50	75.00	71.59	87.24	77.27
Three-way ARCC	**82.77 ± 2.17**	**75.27**	**71.96 ± 2.62**	**64.52**	**82.41 ± 2.19**	**78.49**
SWG	83.23	78.05	69.78	58.84	80.93	75.61
HUM	91.79	79.41	60.00	58.82	91.27	86.76
AGR	75.35	68.18	84.38	77.27	77.56	76.14
Five-way ARCC	**74.84 ± 2.52**	**71.58**	**65.92 ± 2.78**	**70.14**	**73.42 ± 2.56**	**75.90**
SWG	77.94	82.93	91.48	83.74	81.95	81.30
HUM	92.51	88.24	65.09	80.88	93.09	92.65
CAT	47.06	45.83	18.37	54.17	41.84	66.67
SWI	56.98	14.81	29.09	22.22	49.09	44.44
POU	52.90	61.11	40.69	50.00	46.90	55.56

Bold indicates overall ARCC in each categorical scheme.

Source-tracking environmental isolates

The environmental dataset was made up of 332 isolates from various locations, 220 of which were isolated during the wet season (May–October) and the remaining 112 during the dry season (November–April). Source prediction of unknown isolates using kNN differed greatly from DAPC and RF, which had similar percentages in all categorical schemes (Table 4). The kNN method was also not able to supply consistent source attribution across the three different schemes, resulting in varying percentages in each source category. Based on this finding, as well as the generally lower ARCC obtained from kNN, we focused the study on using RF and DAPC models.

Overall, fecal contamination in the environment was heterogeneous in that they were identified to have originated from agricultural, human, and sewage sources across all models used (Table 4). In all classification schemes, sewage was the dominant source of pollution (37.65–49.4%). This was followed by human sources (24.4–29.22% in DAPC and RF). Among the agricultural sources, poultry was the highest (12.65%) contributor of contamination, followed by swine (7.23–9.94%), and lastly, cattle (3.31–3.61%).

DISCUSSION

Fecal contamination is among the factors that cause deterioration in the water quality of Laguna Lake. Total coliform counts monitored by the LLDA provide information on the extent of contamination. However, there is a need to discern the origin of contamination for appropriate mitigation measures to be implemented and protect public health (Griffith *et al.* 2003; McLellan *et al.* 2003). Here, we utilized rep-PCR, a library-dependent MST method, to identify the dominant source of fecal contamination in Laguna Lake and its tributaries.

MST fingerprint library

Several rep-PCR markers for source-tracking *E. coli* are available in literature. Each generates different fingerprint profiles that ultimately affect library accuracy (Mohapatra *et al.* 2007). Labrador *et al.* (2020) assessed the performance of three of these markers and proposed combining profiles of BOX and GTG to attain the highest accuracy for source tracking in Laguna Lake. By combining the banding patterns of the two markers, more variability was introduced in the dataset that allowed for the isolates to be classified to their respective host sources. Improvement in the overall discriminatory power of the source-tracking library was also reported after band profiles from different primers were combined (Yoke-Kqueen

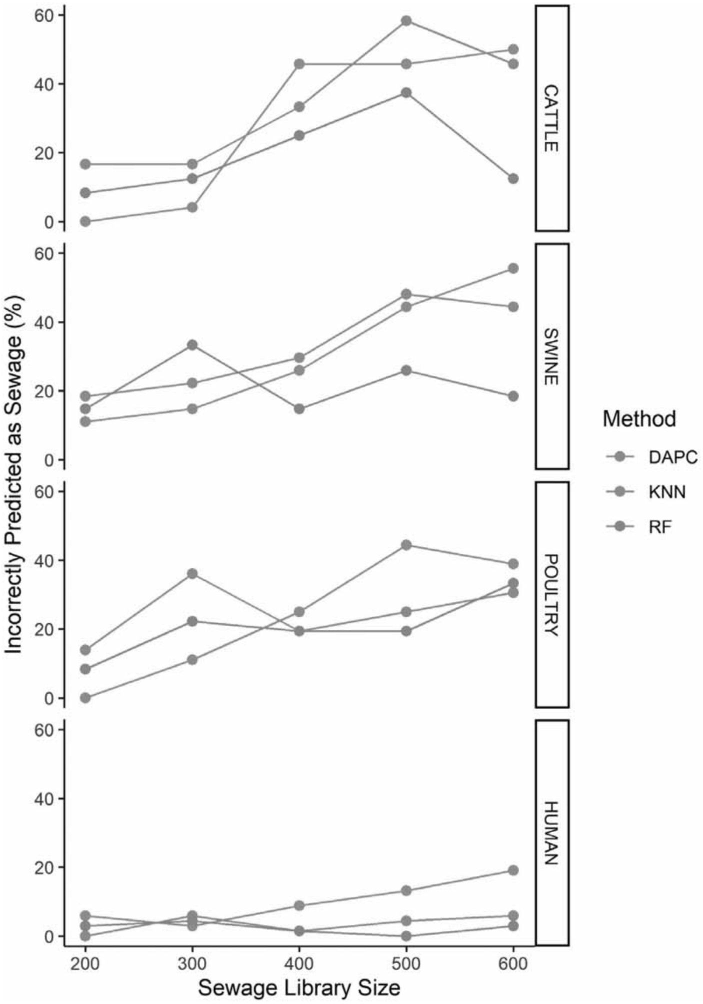


Figure 2 | Trend of the percentage of samples incorrectly predicted as sewage. Class size was based on (1) sample class and (2) statistical model used.

Table 4 | Comparison of predictions of unknown isolates using DAPC, kNN, and RF in two-, three-, and five-way splits

Categorical split	DAPC	kNN	RF
Two-way			
DOM	74.70	55.42	74.70
AGR	25.30	44.58	25.30
Three-way			
SWG	41.57	31.02	37.65
HUM	24.4	3.92	25.90
AGR	34.04	65.06	36.45
Five-way			
SWG	49.4	58.73	47.59
HUM	24.4	9.04	29.22
CAT	3.61	5.12	3.31
SWI	9.94	9.34	7.23
POU	12.65	17.77	12.65

et al. 2013; Sukhumungoon *et al.* 2016). However, partitioning the library into more categorical splits lowered the accuracy, an observation that was concordant with previous reports (Carson *et al.* 2003; Mohapatra *et al.* 2007; Mott & Smith 2011).

The library used as the training set can also be noted to have disproportionate sizes for each point source. The sewage samples are the most abundant with as much as four times those of the fecal libraries. LDMs of MST often rely on the representation of each source candidate in the library, as over- and under-representation may lead to biases in classification. In order to overcome biases, we integrated clonal isolates from the training set to improve prediction and representativeness (Hassan *et al.* 2005).

Our library had more difficulty in classifying the different agricultural host sources: chickens/ducks, cattle, and swine. In contrast, other genotypic libraries had higher accuracies when dealing with nonhuman host sources (Dombek *et al.* 2000; Mohapatra *et al.* 2007; Somarelli *et al.* 2007). Although the addition of representative isolates for each host source was reported to improve the library (Harwood *et al.* 2003; Wiggins *et al.* 2003; Mott & Smith 2011), the fingerprint profiles of the agricultural sources used in this study may not be variable enough to be classified accordingly. Such limitation can be attributed to the inherent fingerprint profiles rather than the library size. Therefore, the addition of more isolates would no longer improve classification. In this case, different profiling methods such as antibiotic resistance assays, or direct detection using library-independent methods (LIMs), can be further explored to assess their efficiency in identifying fecal contamination from agricultural sources.

Another limitation that needed to be addressed was the frequent false positives that plague genotypic libraries (Griffith *et al.* 2003; Myoda *et al.* 2003). Although library-based genotypic methods are able to correctly identify dominant sources of contamination in a given sample, they tend to incorrectly identify absent host sources as being present. A threshold percentage of $\geq$15% was suggested to minimize the severity of false positives (Griffith *et al.* 2003; Harwood *et al.* 2003). Here, we opted not to impose a single threshold for our classification, as the appropriate level of probability of the isolate belonging to the same category may vary with the three different methods. Furthermore, omitting unknown sources in the prediction may be problematic when paired with our goal of determining percentages of fecal contribution in the environmental samples (Ritter *et al.* 2003; Robinson *et al.* 2007). The application of an unclassified category for our samples while comparing the different classification methods may be further explored in the future.

We recommend further assessments of the library including (1) temporal stability of the library, (2) applicability to other water bodies, including other watersheds in the Philippines, (3) addition of wildlife sources, and (4) replicability of results in other laboratories, as a precedent for the use of MST as a regular monitoring tool.

Prediction models

Harwood *et al.* (2000) proposed that ARCCs ranging around 60–70% are suitable for MST studies with the objective of pollution control. Our model was able to predict source categories with an accuracy of 74.77% for a five-way split using RF, which is comparable to other reported ARCCs obtained through rep-PCR of *E. coli*. Lyautey *et al.* (2010) reported an accuracy of 77% using BOX and ERIC libraries, while Mohapatra *et al.* (2007) achieved 79.89% in GTG(5) library. Other gel fingerprinting studies have used different classifiers such as discriminant analysis (Dombek *et al.* 2000; Mohapatra *et al.* 2007; Somarelli *et al.* 2007), support vector machines (Garabetian *et al.* 2020), kNN and neural networks (Carlos *et al.* 2012) with lower or similar accuracy rates.

Robinson *et al.* (2007) suggested kNN as a compromise between the strengths of maximum similarity (MS) and discriminant analysis in terms of accuracy and prediction bias when dealing with disproportionate libraries. However, our results show that compared with DAPC and RF, kNN is the least suitable for source classification in our library and is more prone to incorrectly predicting isolates as sewage-derived as sewage library increases. This is also consistent with Lyautey *et al.* (2010) reporting kNN as being less sensitive and specific than another model, MS.

Comparison of different statistical algorithms shows that DAPC and RF yield similar results in terms of prediction accuracy, while kNN may not be as suitable for our fingerprint library. RF was observed to be more robust when tackling unbalanced datasets and can be used in the quantification of source attribution of fecal pollution. Smith *et al.* (2010) also reported the outperformance of DA by RF using antibiotic resistance profiles for bacterial source tracking. The agreement between the prediction from DAPC and RF lends more support and increases confidence in our result.

It can be observed that for DAPC, the test set ARCC is much lower compared with the training set. This is perhaps due to overfitting, wherein the model learns the training set well but underperforms when looking at unknowns. The major source of incorrect predictions during test set validation came from isolates wrongly classified as sewage, particularly those from agricultural sources (cattle, swine, and poultry). The prediction bias toward sewage may be caused by the inherent class imbalance present in the fecal source library, with sewage samples populating most of the dataset at almost four times as much as other fecal sources (Table 1). This resulted in lower RCCs for these classes, particularly in DAPC and kNN. In contrast, RF exhibited more balanced RCC rates for the agricultural sources by minimizing incorrect assignments without compromising accuracy in sewage prediction. RF has been said to be less prone to overfitting (Breiman 2001), making it useful for our study. RF has also been implemented in several source-tracking studies using other library-dependent methodologies such as antibiotic resistance profiling (Smith *et al.* 2010), microarray (Dubinsky *et al.* 2016), and 16S amplicon sequences (Roguet *et al.* 2018, 2020).

Fecal source attribution in Laguna Lake

All three statistical methods used in this study unanimously show that sewage is the dominant source of fecal pollution in Laguna Lake. The contamination of lake water by sewage poses a health risk to the populace residing around the vicinity. This is because sewage is reported to contain chemical contaminants, such as heavy metals and organic compounds (Lamastra *et al.* 2018), and biological contaminants, such as pathogens, contaminants of emerging concerns, and antimicrobial resistant genes (Alygizakis *et al.* 2020). Fecal pollution due to sewage may be due to the inadequate wastewater treatment system in the Philippines. Specifically, in Metro Manila, just about 15% of households are linked to sewage systems, and of this only half are being treated before release (Palanca-Tan 2017; Jalilov 2018). The majority of the wastewater in the megacity flows into septic tanks, which are not government-regulated but privately maintained by each household (Palanca-Tan 2013). A more alarming situation is the case of informal settlements, where wastewater is being directly dumped into drainage without any treatment (Palanca-Tan 2017). According to Santo Domingo & Edge (2010), untreated municipal wastewater may also contain fecal contamination derived from animal sources. Similarly, Griffith *et al.* (2003) reported that sewage is not purely composed of human contamination; rather, it contains fecal contamination from other sources that infiltrate sewage systems.

In several studies (Wiggins *et al.* 1999; Griffith *et al.* 2003; Ahmed *et al.* 2009; Kon *et al.* 2009), MST often used FIB from sewage as a substitute for human-related contamination. However, our study demonstrated that our sewage samples can be successfully discriminated from human sources compared with other agricultural isolates (Figure 2). This revealed several pieces of information: first, human contribution to sewage may not be as high as was previously expected (Labrador *et al.* 2020), and that the former is not a good representative of the latter. Secondly, animal-derived contamination may have a higher contribution to sewage contamination than previously known. Lastly, the distinct sewage category suggested that it was a mixture of contamination from various sources, such as industrial wastes and non-point sources, that were not accounted for in this study.

Although the use of library-independent host-specific markers has gained more popularity in recent source-tracking studies, the advantage of the LDM is that it can supplement the methodologies that monitoring systems already have in place. Our LDM utilizes indicator bacteria that the LLDA is also testing, which allows for congruence in data gathering, ease of sampling, and comparability of results (Mott & Smith 2011). Furthermore, the established methods, which include bacterial culture and PCR, are more easily adapted in laboratories which may lack the equipment and expertise required for LIM studies.

Overall, our results show that there is evidence of fecal contamination in Laguna Lake. This implies that management guidelines should be improved and strictly implemented to improve the water quality of the lake. Moreover, we identified the dominant sources of pollution in the watershed, which can aid government units and monitoring agencies in focusing their efforts and enacting policies for water quality management, such as establishing more stringent standards for wastewater disposal and total maximum daily loads of *E. coli* and other bacteria. Proper rehabilitation of bodies of water is important because it can prevent outbreaks of human and ecosystem diseases (Santo Domingo *et al.* 2007), leading to a positive impact in the field of public health.

CONCLUSION

MST of fecal contamination is an important aspect of water quality management. We demonstrate that DNA fingerprinting of *E. coli* coupled with RF is a promising method for source attribution in Laguna Lake. The validation of the prediction algorithm showed that RF yields the highest accuracy compared with DAPC and kNN and is shown to be more reliable when dealing with disproportionate datasets. RF found the dominant source of fecal contamination to be from sewage, a result that was also supported by DAPC. This may lead to public health implications; thus, more stringent measures are recommended to improve the lake's water quality. Further studies on MST should explore whether the outcome is scalable in spatially and temporally distinct settings.

ACKNOWLEDGEMENTS

This study was financially supported by the Philippine Council for Industry, Energy, and Emerging Technology Research and Development (PCIEERD) of the Department of Science and Technology (DOST). We thank our collaborating agency, Laguna Lake Development Authority (LLDA), for the technical assistance. We also acknowledge the following individuals for their valuable contributions to the study: Jeffrey Thomas V. Beldia, Joseth Jermaine M. Abello, Kenneth T. Co, Gicelle T. Malajacan, Sherwina S. Paiso, Mark Raymond A. Vejano, and Jerro Paolo I. Lunas.

CONFLICT OF INTEREST

The authors declare that they have no conflicts of interest.

DATA AVAILABILITY STATEMENT

All relevant data are included in the paper or its Supplementary Information.

REFERENCES

Ahmed, W., Hargreaves, M., Goonetilleke, A. & Katouli, M. 2008 Population similarity analysis of indicator bacteria for source prediction of faecal pollution in a coastal lake. *Marine Pollution Bulletin* **56** (8), 1469–1475.

Ahmed, W., Goonetilleke, A., Powell, D., Chauhan, K. & Gardner, T. 2009 Comparison of molecular markers to detect fresh sewage in environmental waters. *Water Research* **43**, 4908–4917.

Alygizakis, N. A., Urík, J., Beretsou, V. G., Kampouris, I., Galani, A., Oswaldova, M., Berendonk, T., Oswald, P., Thomaidis, N. S., Slobodnik, J., Vrana, B. & Fatta-Kassinos, D. 2020 Evaluation of chemical and biological contaminants of emerging concern in treated wastewater intended for agricultural reuse. *Environment International* **138**, 105597. https://doi.org/10.1016/j.envint.2020.105597.

Breiman, L. 2001 Random forests. *Machine Learning* **45**, 5–32.

Carlos, C., Alexandrino, F., Stoppe, N. C., Sato, M. I. Z. & Ottoboni, L. M. M. 2012 Use of *Escherichia coli* BOX-PCR fingerprints to identify sources of fecal contamination of water bodies in the State of São Paulo, Brazil. *Journal of Environmental Management* **93** (1), 38–43.

Carson, C. A., Shear, B. L., Ellersieck, M. R. & Schnell, J. D. 2003 Comparison of ribotyping and repetitive extragenic palindromic-PCR for identification of fecal *Escherichia coli* from humans and animals. *Applied and Environmental Microbiology* **69** (3), 1836–1839.

Dombek, P. E., Johnson, L. K., Zimmerley, S. T. & Sadowsky, M. J. 2000 Use of repetitive DNA sequences and the PCR to differentiate *Escherichia coli* isolates from human and animal sources. *Applied and Environmental Microbiology* **66** (6), 2572–2577.

Dubinsky, E. A., Butkus, S. R. & Andersen, G. L. 2016 Microbial source tracking in impaired watersheds using PhyloChip and machine-learning classification. *Water Research* **105**, 56–64.

Edge, T. A., Hill, S., Seto, P. & Marsalek, J. 2010 Library-dependent and library-independent microbial source tracking to identify spatial variation in faecal contamination sources along a Lake Ontario beach (Ontario, Canada). *Water Science and Technology* **62** (3), 719–727.

Garabetian, F., Vitte, I., Sabourin, A., Moussard, H., Jouanillou, A., Mornet, L., Lesne, M. & Lyautey, E. 2020 Uneven genotypic diversity of *Escherichia coli* in fecal sources limits the performance of a library-dependent method of microbial source tracking on the southwestern French Atlantic coast. *Canadian Journal of Microbiology* **66** (12), 698–712.

Garcia, B. C. B., Dimasupil, M. A. Z., Vital, P. G., Widmer, K. W. & Rivera, W. L. 2015 Fecal contamination in irrigation water and microbial quality of vegetable primary production in urban farms of Metro Manila, Philippines. *Journal of Environmental Science and Health, Part B* **50** (10), 734–743.

Griffith, J. F., Weisberg, S. B. & McGee, C. D. 2003 Evaluation of microbial source tracking methods using mixed fecal sources in aqueous test samples. *Journal of Water and Health* **1** (4), 141–151.

Harwood, V. J., Whitlock, J. & Withington, V. 2000 Classification of antibiotic resistance patterns of indicator bacteria by discriminant analysis: use in predicting the source of fecal contamination in subtropical waters. *Applied and Environmental Microbiology* **66** (9), 3698–3704.

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. *Journal of Water and Health* **1** (4), 153–166.

Hassan, W. M., Wang, S. Y. & Ellender, R. D. 2005 Methods to increase fidelity of repetitive extragenic palindromic PCR fingerprint-based bacterial source tracking efforts. *Applied and Environmental Microbiology* **71** (1), 512–518.

Jalilov, S. M. 2018 Value of clean water resources: estimating the water quality improvement in Metro Manila, Philippines. *Resources* **7** (1), 1.

Jombart, T. 2008 adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics* **24** (11), 1403–1405.

Kheiri, R. & Akhtari, L. 2017 Genetic diversity and antimicrobial resistance of *Escherichia coli* as microbial source tracking tools of Karaj River, Iran. *Water Science and Technology: Water Supply* **17** (5), 1468–1478.

Kon, T., Weir, S. C., Howell, E. T., Lee, H. & Trevors, J. T. 2009 Repetitive element (REP)-polymerase chain reaction (PCR) analysis of *Escherichia coli* isolates from recreational waters of southeastern Lake Huron. *Canadian Journal of Microbiology* **55** (3), 269–276.

Kuhn, M. 2008 Building predictive models in R using the caret package. *Journal of Statistical Software* **28** (5), 1–26.

Labrador, K. L., Nacario, M. A. G., Malajacan, G. T., Abello, J. J. M., Galarion, L. H., Rensing, C. & Rivera, W. L. 2020 Selecting rep-PCR markers to source track fecal contamination in Laguna Lake, Philippines. *Journal of Water and Health* **18** (1), 19–29.

Laguna Lake Development Authority 2019 *LLDA Quarterly Water Monitoring Report Laguna Lake and Tributary Rivers October to December, 2019*. Available from: https://llda.gov.ph/wp-content/uploads/dox/waterqualityrpt/2019/2019q4.pdf.

Laguna Lake Development Authority 2020 *LLDA Quarterly Water Monitoring Report Laguna Lake and Tributary Rivers October to December, 2020*. Available from: https://llda.gov.ph/wp-content/uploads/dox/waterqualityrpt/2020/2020q4.pdf.

Lamastra, L., Suciu, N. A. & Trevisan, M. 2018 Sewage sludge for sustainable agriculture: contaminants' contents and potential use as fertilizer. *Chemical and Biological Technologies in Agriculture* **5** (10). https://doi.org/10.1186/s40538-018-0122-3.

Lyautey, E., Lu, Z., Lapen, D. R., Berkers, T. E., Edge, T. A. & Topp, E. 2010 Optimization and validation of rep-PCR genotypic libraries for microbial source tracking of environmental *Escherichia coli* isolates. *Canadian Journal of Microbiology* **56** (1), 8–17.

McLellan, S. L., Daniels, A. D., Alissa, K. & Salmore, A. K. 2003 Genetic characterization of *Escherichia coli* populations from host sources of fecal pollution by using DNA fingerprinting. *Applied and Environmental Microbiology* **69** (5), 2587–2594.

Mohapatra, B. R., Broersma, K. & Mazumder, A. 2007 Comparison of five rep-PCR genomic fingerprinting methods for differentiation of fecal *Escherichia coli* from humans, poultry and wild birds. *FEMS Microbiology Letters* **277** (1), 98–106.

Mott, J. & Smith, A. 2011 Library-dependent source tracking methods. In: *Microbial Source Tracking: Methods, Applications, and Case Studies* (Hagedorn, C., Blanch, A. & Harwood, V., eds). Springer, New York.

Myoda, S. P., Carson, C. A., Fuhrmann, J. J., Hahm, B. K., Hartel, P. G., Yampara-Iquise, 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. *Journal of Water Health* **1** (4), 167–180.

Palanca-Tan, R. 2013 Efficiency and environmental effects of privatizing waterworks and sewerage (MWSS) in Metro Manila, Philippines. *IAMURE International Journal of Ecology and Conservation* **5** (1), 57–71. https://doi.org/10.7718/ijec.v5i1.507.

Palanca-Tan, R. 2017 Health and water quality benefits of alternative sewerage systems in Metro Manila, Philippines. *Environment and Urbanization* **29** (2), 567–580. https://doi.org/10.1177/0956247817718402.

R Core Team 2020 *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ritter, K. J., Carruthers, E., Carson, C. A., Ellender, R. D., Harwood, V. J., Kingley, K., Nakatsu, C., Sadowsky, M., Shear, B., West, B., Whitlock, J. E., Wiggins, B. A. & Wilbur, J. D. 2003 Assessment of statistical methods used in library-based approaches to microbial source tracking. *Journal of Water and Health* **1** (4), 209–223.

Robinson, B. J., Ritter, K. J. & Ellender, R. D. 2007 A statistical appraisal of disproportional versus proportional microbial source tracking libraries. *Journal of Water and Health* **5** (4), 503–509.

Roguet, A., Eren, A. M., Newton, R. J. & McLellan, S. L. 2018 Fecal source identification using random forest. *Microbiome* **6**, 185. https://doi.org/10.1186/s40168-018-0568-3.

Roguet, A., Esen, Ö. C., Eren, A. M., Newton, R. J. & McLellan, S. L. 2020 FORENSIC: an online platform for fecal source identification. *Msystems* **5** (2). https://doi.org/10.1128/mSystems.00869-19.

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 Research* **41** (16), 3539–3552.

Santo Domingo, J. W. & Edge, T. A. 2010 Identification of primary sources of faecal pollution. In: *Safe management of shellfish and harvest waters* (G. Rees, K. Pond, D. Kay, J. Bartram & J. Santo Domingo, eds). World Health Organization, IWA Publishing, London, pp. 51–90.

Santos-Borja, A. & Nepomuceno, D. N. 2006 Laguna de Bay: institutional development and change for lake basin management. *Lakes & Reservoirs: Research & Management* **11** (4), 257–269.

Smith, A., Sterba-Boatwright, B. & Mott, J. 2010 Novel application of a statistical technique, Random Forests, in a bacterial source tracking study. *Water Research* **44** (14), 4067–4076.

Somarelli, J. A., Makarewicz, J. C., Sia, R. & Simon, R. 2007 Wildlife identified as major source of *Escherichia coli* in agriculturally dominated watersheds by BOX A1R-derived genetic fingerprints. *Journal of Environmental Management* **82** (1), 60–65.

Stoeckel, D. M. & Harwood, V. J. 2007 Performance, design, and analysis in microbial source tracking studies. *Applied and Environmental Microbiology* **73** (8), 2405–2415.

Sukhumungoon, P., Tantadapan, R. & Rattanachuay, P. 2016 Repetitive sequence based-PCR profiling of *Escherichia coli* O157 strains from beef in Southern Thailand. *Southeast Asian Journal of Tropical Medicine and Public Health* **47** (1), 55–65.

Tanganco, L. J. U., Alberto, M. A. J. & Gotangco, C. K. Z. 2019 Forecast of potential areas of urban expansion in the Laguna de Bay basin and its implications to water supply security. *Philippine Journal of Science* **148** (4), 715–724.

Vargas, V. 2015 *Laguna de Bay, Philippines: Environmental Literacy*. Integration and Application Network. Available from: https://ian.umces.edu/blog/2015/07/21/laguna-de-bay-philippines-environmental-literacy/.

Wiggins, B. A., Andrews, R. W., Conway, R. A., Corr, C. L., Dobratz, E. J., Dougherty, D. P., Eppard, J. R., Knupp, S. R., Limjoco, M. C., Mettenburg, J. M., Rinehardt, J. M., Sonsino, J., Torrijos, R. L. & Zimmerman, M. E. 1999 Use of antibiotic resistance analysis to identify nonpoint sources of fecal pollution. *Applied and Environmental Microbiology* **65** (8), 3483–3486.

Wiggins, B. A., Cash, P. W., Creamer, W. S., Dart, S. E., Garcia, P. P., Gerecke, T. M., Han, J., Henry, B. L., Hoover, K. B., Johnson, E. L., Jones, K. C., McCarthy, J. G., McDonough, J. A., Mercer, S. A., Noto, M. J., Park, H., Phillips, M. S., Purner, S. M., Smith, B. M., Stevens, E. N. & Varner, A. K. 2003 Use of antibiotic resistance analysis for representativeness testing of multiwatershed libraries. *Applied and Environmental Microbiology* **69** (6), 3399–3405.

Yoke-Kqueen, C., Teck-Ee, K., Son, R., Yoshitsugu, N. & Mitsuaki, N. 2013 Molecular characterisation of *Vibrio parahaemolyticus* carrying tdh and trh genes using ERIC-, RAPD- and BOX-PCR on local Malaysia bloody clam and Lala. *International Food Research Journal* **20** (6), 3299–3305.

First received 7 April 2021; accepted in revised form 20 August 2021. Available online 1 September 2021