Selecting rep-PCR markers to source track fecal contamination in Laguna Lake, Philippines

Kevin L. Labrador, Mae Ashley G. Nacario, Gicelle T. Malajacan, Joseth Jermaine M. Abello, Luiza H. Galarion, Christopher Rensing and Windell L. Rivera

ABSTRACT

Fecal contamination is one of the factors causing deterioration of Laguna Lake. Although total coliform levels are constantly monitored, no protocol is in place to identify their origin. This can be addressed using the library-dependent microbial source tracking (MST) method, repetitive element sequence-based polymerase chain reaction (rep-PCR) fingerprinting. Serving as a prerequisite in developing the host-origin library, we assessed the discriminatory power of three fingerprinting primers, namely BOX-A1R, $(GTG)_5$, and REP1R-1/2-1. Fingerprint profiles were obtained from 290 thermotolerant *Escherichia coli* isolated from sewage waters and fecal samples of cows, chickens, and pigs from regions surrounding the lake. Band patterns were converted into binary profiles and were classified using the discriminant analysis of principal components. Results show that: (1) REP1R-1/2-1 has a low genotyping success rate and information content; (2) increasing the library size led to more precise estimates of library accuracy; and (3) combining fingerprint profiles from BOX-A1R and $(GTG)_5$ revealed the best discrimination (average rate of correct classification (ARCC) $= 0.82 \pm 0.06$) in a two-way categorical split; while (4) no significant difference was found between the combined profiles (0.74 ± 0.15) and using solely BOX-A1R (0.76 ± 0.09) in a four-way split. Testing the library by identifying known isolates from a separate dataset has shown that a two-way classification performed better (ARCC $= 0.66$) than a four-way split (ARCC $= 0.29$). The library can be developed further by adding more representative isolates per host source. Nevertheless, our results have shown that combining profiles from BOX-A1R and $(GTG)_5$ is recommended in developing the MST library for Laguna Lake.

Key words | *E. coli*, Laguna Lake, microbial source tracking, rep-PCR

Kevin L. Labrador
Mae Ashley G. Nacario
Gicelle T. Malajacan
Joseth Jermaine M. Abello
Luiza H. Galarion
Windell L. Rivera (corresponding author)
Pathogen-Host-Environment Interactions Research Laboratory, Natural Sciences Research Institute,
University of the Philippines Diliman,
Quezon City,
Philippines
E-mail: *wlrivera@science.upd.edu.ph*

Mae Ashley G. Nacario
Gicelle T. Malajacan
Joseth Jermaine M. Abello
Windell L. Rivera
Institute of Biology, College of Science,
University of the Philippines Diliman,
Quezon City,
Philippines

Christopher Rensing
Fujian Provincial Key Laboratory of Soil Environmental Health and Regulation, College of Resources and Environment,
Fujian Agriculture and Forestry University,
Fuzhou,
China

INTRODUCTION

Laguna Lake, the largest lake in the Philippines, has experienced continued deterioration over the past years. The surrounding communities, estimated to have a population size of around 15 million people, benefit from the lake through various agricultural, industrial, and domestic uses (Laguna Lake Development Authority 2018). However, these communities introduce waste into the system, and the intensive demand coupled with high pollution load contributed to the decline in water quality (Santos-Borja & Nepomuceno 2006; WAVES 2016). In fact, monitoring activities by the Laguna Lake Development Authority, the agency that is primarily responsible for the lake's development, revealed fecal contamination in the lake and high total coliform counts in most of its tributaries. These pose serious public health problems since pathogens from infected sources can be introduced into the environment through feces, ultimately causing risks to public health and the economy (Ahmed *et al.* 2009; Ballesté *et al.* 2010).

doi: 10.2166/wh.2019.042

Although the extent of the contamination is known, there is no information pertaining to their origin. Knowledge on sources of fecal contamination is critical because it can (1) allow development of management schemes to minimize their input, (2) aid in restoration and remediation efforts, (3) evaluate health risks, and (4) reduce the danger of disease outbreaks (Hagedorn *et al.* 1999; Gourmelon *et al.* 2007; Graves *et al.* 2007; Okabe *et al.* 2007; Ballesté *et al.* 2010). One way to identify sources of fecal contamination is through microbial source tracking (MST).

MST is a collection of methods that utilize microorganisms to identify and quantify dominant sources of fecal contamination in a given system (Scott *et al.* 2002; Stoeckel & Harwood 2007). The methods are broadly categorized into either (1) library-dependent (LDM) or (2) library-independent (LIM), depending on the need to develop a culture library of known host sources that will serve as a database for identifying unknown isolates (Gomi *et al.* 2014). Genotypic methods, a subclass of LDM that utilizes genetic fingerprinting techniques, rely on the genetic variation of indicator bacteria to develop the library (Scott *et al.* 2002). Among these methods is repetitive element sequence-based polymerase chain reaction (rep-PCR), a technique that targets repetitive palindromic sequences that are widespread in bacterial genomes (Versalovic *et al.* 1991). Compared to other genetic fingerprinting methods, rep-PCR is easier, less technically demanding, faster, cheaper, and displays better reproducibility, while generating relatively more efficient and reliable results that are easy to interpret, making it a practical approach for source tracking fecal contamination (Dombek *et al.* 2000; Scott *et al.* 2002; Carson *et al.* 2003; Mott & Smith 2011; Kheiri & Akhtari 2017).

The success of MST-LDM depends on the culture library. Several factors must be considered during its development including the rep-PCR marker used, library representativeness and size (Mohapatra *et al.* 2007; Mott & Smith 2011). There are several rep-PCR markers that are commonly used for genotyping, and each tends to generate different fingerprint profiles thereby affecting library accuracy (Mohapatra *et al.* 2007). Meanwhile, there is a need to profile greater than 600 isolates per host source to account for the genetic diversity of the microbial indicator and refine the size and representativeness of the library (Mott & Smith 2011). Developing a library that meets the recommended size while using many primers demands considerable resources. Therefore, as a screening step, it is necessary to evaluate the discriminatory power of several primers on a small number of samples first. This is to ensure that resource allocation is maximized toward the development of an accurate fingerprint library.

Our objective was to assess the discriminatory power of three rep-PCR fingerprinting primers in source tracking fecal contamination in Laguna Lake. The primers, namely BOX-A1R, $(GTG)_5$ and REP1R-1/2-1, were selected since they had the highest average rate of correct classification (ARCC) based on a comparative study by Mohapatra *et al.* (2007). We also employed the composite analysis to assess the accuracy of the different primer combinations. Using thermotolerant *Escherichia coli* as the microbial indicator, we populated the reference library with isolates from sewage waters and fecal samples from cows, chickens, and pigs. Initial site surveys and consultations with stakeholders identified these host sources as major contributors of fecal contamination to the lake. Furthermore, we also assessed library performance based on its categorical split – that is the number of categories in the library in which the isolates will be categorized. In this regard, we have designed a two-way split (Agricultural–Domestic) and a four-way split (Chicken–Cow–Pig–Sewage) library. We aimed to answer the following questions: (1) How did an increase in library size affect its performance? (2) Was a two-way split better than a four-way split? (3) Which among the primers or primer combinations were best suited in source tracking fecal contamination to Laguna Lake? (4) How accurate was the library in source tracking known isolates? To the best of our knowledge, this is the first initiative to perform MST on the Philippines' largest lake.

METHODS

Sample collection

Fecal samples were collected from potential agricultural animal host sources (i.e., chicken, cows, and pigs) in backyard farms located in the provinces of Laguna and Rizal (Figure 1). Sewage samples representing human-derived contamination were also collected from sewage treatment

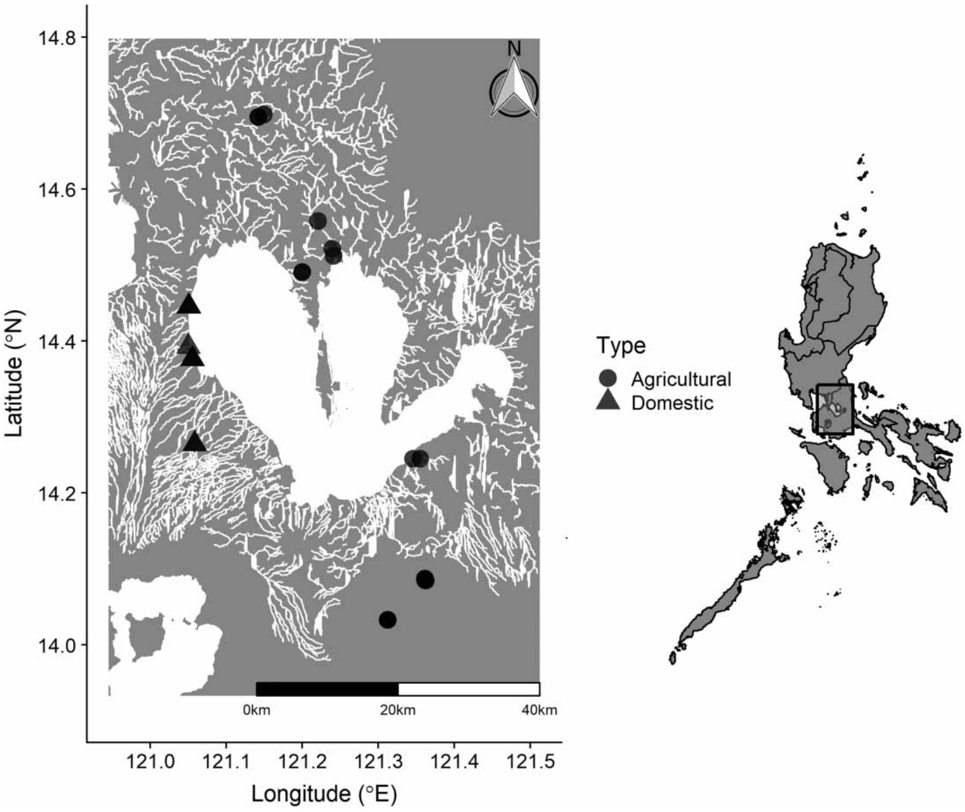

Figure 1 | Sample collection sites for constructing the reference library. Colors indicate the types of sample that were collected. Inset shows the geographical location relative to the island of Luzon. Please refer to the online version of this paper to see this figure in color: http://dx.doi.org/10.2166/wh.2019.042.

facilities of Manila Water and Laguna Water, as well as in sewage waters from Metro Manila draining toward Laguna Lake. Collected samples were stored in ice and were transported to the Pathogen–Host–Environment Interactions Research Laboratory (PHEIRL) of the Natural Sciences Research Institute (NSRI), the University of the Philippines Diliman for processing within 24 h after collection. Sample collection was done from July 2017 to July 2018.

Isolation and identification of thermotolerant *E. coli*

Fecal samples from each animal source at a given site were pooled and 10 g were aseptically transferred into a sterile flask containing 0.9% saline solution (90 mL). After vigorous mixing, the mixture was serially diluted up to 10^{-7} using 0.9% saline solution (30 mL) as a diluent. Meanwhile, sewage samples were serially diluted up to 10^{-10} using the same diluent. Serial dilutions, 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. Identities of presumptive *E. coli* isolates, characterized by blue to violet colonies on mTEC plates, were further confirmed using eosin methylene blue agar (EMBA; BD BBL, USA). Isolates that exhibited a green metallic sheen on EMBA were selected for DNA extraction and additional molecular identification based on the protocol by Garcia *et al.* (2015), with slight modifications.

DNA was extracted using the boil-lysis DNA extraction method. Briefly, *E. coli* grown on trypticase soy broth (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 supernatant (50 μL) was collected in a fresh microtube and was used as a template for molecular identification. This was done by amplifying the 75 bp fragment of the *E. coli uidA* gene using the primers ECN1254F (5′-GCAAGGTGCACGGGAATATT-3′) and ECN1328R (5′-CAGGTGATCGGACGCGT-3′). The 10 μL PCR mixture was composed of the following: GoTaq Green Master Mix (1×, Promega, USA), forward and reverse primers (0.50 μM each), template DNA (1 μL), and an appropriate amount of PCR water. The PCR was performed in a thermal cycler (C1000, Bio-Rad, USA) with the following cycling conditions: initial denaturation (98 °C, 2 min); 35 cycles of denaturation (95 °C, 30 s), annealing (63 °C, 1 min), and extension (72 °C, 1 min); and final extension (72 °C, 5 min). A no-template control (NTC) was included in every run. Templates that generated the expected band size were selected for DNA fingerprinting.

DNA fingerprinting

Three primers were employed for fingerprinting (Mohapatra *et al.* 2007): (1) BOX-A1R (5′-CTACGGCAAGGCGACGCTGACG-3′); (2) $(GTG)_5$ (5′-GTGGTGGTGGTGGTG-3′); and (3) REP1R-I (5′-IIIICGICGICATCIGGC-3′) and REP2-I (5′-ICGICTTATCIGGCCTAC-3′) (hereafter referred to as BOX, GTG, and REP, respectively). The 10 μL PCR mixture for each setup was composed of the following: GoTaq Green Master Mix (1×, Promega, USA), primer (1 μM), template DNA (1 μL), and the appropriate amount of PCR water. The PCR was performed with similar cycling conditions for all 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). The NTC was included in every run.

The resulting amplicons were subjected to agarose gel electrophoresis (2%, 190 V, 60 min), and the gels were visualized using a gel documentation system (Bio-print ST4, Vilber Lourmat, UK). The analysis of banding patterns was done with an imaging system (SuperMegaCapt 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.

Fingerprint analysis

The statistical analysis was performed using the programming language, R v.3.5.0 (R Core Team 2017). Band positions were adjusted by binning their molecular weight (MW) to the nearest 20 bp. The binned MW was then converted into a binary sequence based on their presence (1) or absence (0) across samples. Isolates from a particular host source having identical binary sequences were collapsed into a single observation; this decloning step has shown to improve prediction and library representativeness (Wiggins *et al.* 2003; Mott & Smith 2011). In addition, composite profiles that integrated the binary sequence from all primers were included in the analysis. The library was prepared depending on how the host sources were categorized: in a two-way split, the categories were (1) domestic (sewage isolates) and (2) agricultural (animal fecal isolates); in a four-way split, the categories were (1) sewage, (2) chicken, (3) cow, and (4) pig.

Library accuracy was assessed using discriminant analysis of principal components (DAPCs) as implemented in the package, *adegenet* (Jombart 2008). To prevent classification bias leading to either overfitting or underfitting of the model, cross-validation was done 100 times (training set = 0.9). The number of principal components (PCs) with the least mean square error was carried over for the discriminant analysis (DA). The rate of correct classification (RCC; the percentage of isolates from a given host source that were correctly classified) and the average rate of correct classification (ARCC; percentage of the isolates correctly classified in all host categories; Mohapatra *et al.* 2007) were calculated by creating a confusion matrix between the observed and predicted categories using the package, *caret* (Kuhn *et al.* 2018).

The trend in ARCC as a function of the library size was also considered. To remove biases associated with disproportionate libraries (Mott & Smith 2011), the equal number of isolates from each host source (n_{host}) was randomly selected from the sample pool before subjecting them to DAPC. The value of n_{host} was increased by multiples of ten and the classification was iterated 100 times. Afterwards, the mean and 95% confidence interval (CI) of the ARCC were calculated. Lastly, the variation was compared (1) among rep-PCR markers within a categorical split and (2) between categorical split within a rep-PCR marker. For the former, the Kruskal–Wallis test was used followed by the *post-hoc* pairwise Wilcoxon test (p-values were adjusted using Bonferroni correction). For the latter, the Wilcoxon rank-sum test was used. The alpha level of significance (α) was set to 0.05 for all statistical tests.

Library testing

Once the optimal marker combination was determined, we tested the constructed library using a separate dataset. The test dataset was composed of *E. coli* from known sources that were subjected to similar assays and data preparation used in library construction. We used the *predict.dapc* function in order to categorize the test dataset based on the fingerprint library. This function calculated the posterior probability of each isolate in the test dataset; the category with the highest posterior probability was considered as the most probable identity of an isolate. This was done on both the two-way and four-way categorical split.

RESULTS

DNA fingerprinting

Rep-PCR profiles of 290 thermotolerant *E. coli* from various host sources were obtained using three primers. The host sources were from domestic sewage ($n = 110$) and agricultural animals (chicken, cow, pig; $n = 60$ each) (Figure 2). Fingerprint profiles, specifically the number of bands and the range of their MWs, are summarized in Table 1. BOX and GTG had a high percentage of amplification success (98 and 99%, respectively). In contrast, fewer samples

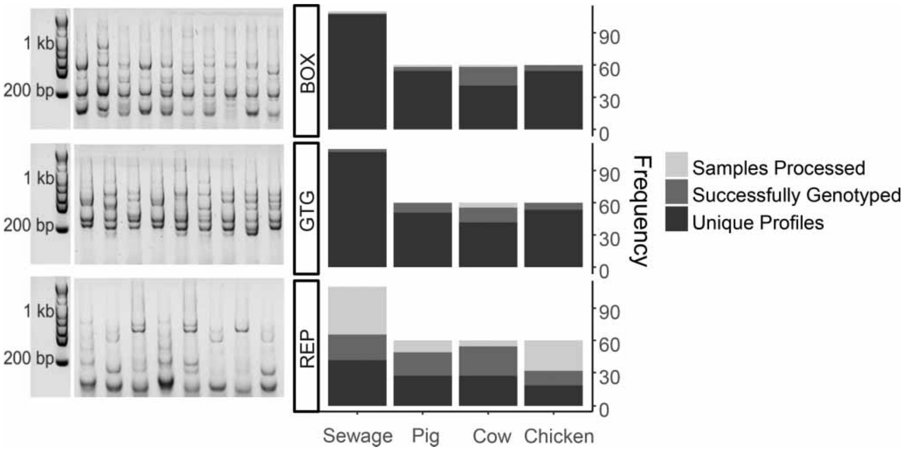

Figure 2 | Representative gel profiles and the number of isolates processed for primer assessment.

Table 1 | Rep-PCR fingerprint profiles obtained from the MST markers evaluated

			Range of band count		Range of band size (bp)	
MST marker	**Host source**	**Average band count**	**Minimum**	**Maximum**	**Minimum**	**Maximum**
BOX	Domestic	5	1	11	<20	1,800
	Agricultural	3	1	6	<20	1,440
GTG	Domestic	4	1	9	20	1,660
	Agricultural	3	1	5	20	960
REP	Domestic	2	1	5	20	1,040
	Agricultural	2	1	4	40	860

amplified using REP (70%). In addition, more complex profiles were generated by BOX and GTG, while only a few REP unique profiles remained after identical profiles were collapsed (40%). Since REP had a low genotyping success rate and minimal information content, it was considered ineffective and was omitted from the succeeding analyses.

Library performance

The library was assessed by increasing the number of representatives per host source (n_{host}) in multiples of ten and then subjecting the dataset to 100 iterations of the classification model. Given the categorical classification of the isolates in the current sample pool, greater n_{host} was evaluated for a two-way split (max $n_{\text{host}} = 100$) than the four-way split (max $n_{\text{host}} = 40$).

Two generalizations could be inferred from the results (Figure 3). First, increasing n_{host} led to decreasing mean and 95% CI of ARCC. Second, for the two-way categorical split, the mean and CI of ARCC stabilized as sample size increased. However, the same cannot be said for the four-way split since not enough n_{host} was obtained to observe stability in the distribution. This preliminary assessment provides an estimate of the ARCC of the rep-PCR marker used prior to maximizing the library.

In assessing the differences between the categorical split, the distribution at $n_{\text{host}} = 30$ was used to allow comparison among the markers used. Overall, a two-way split performed better than a four-way split. The Wilcoxon rank-sum test consistently showed significantly a higher ARCC in a two-way split for BOX (two-way = 0.83 ± 0.09, four-way = 0.77 ± 0.09; $p < 0.05$), GTG (two-way = 0.78 ± 0.10, four-way = 0.63 ± 0.12; $p < 0.05$), and BOX–GTG (two-way = 0.82 ± 0.09, four-way = 0.74 ± 0.15; $p < 0.05$).

Marker selection

In assessing the differences among markers, the ARCC distribution at maximum n_{host} was used. Based on the Kruskal–Wallis test, there was a statistically significant difference among the ARCC obtained from different marker types, both for two-way ($H_{(2)} = 109.75$, $p < 0.05$) and four-way ($H_{(2)} = 55.74$, $p < 0.05$) categorical split. For the two-way split (Table 2), the highest mean ARCC was from BOX–GTG (0.82 ± 0.06), followed by BOX (0.79 ± 0.04), and lastly, GTG (0.72 ± 0.05); the Wilcoxon rank-sum test revealed significant differences in all pairwise comparisons even after Bonferroni correction. For the four-way split (Table 3), no significant difference was found between BOX–GTG (0.74 ± 0.15) and BOX (0.76 ± 0.09). In

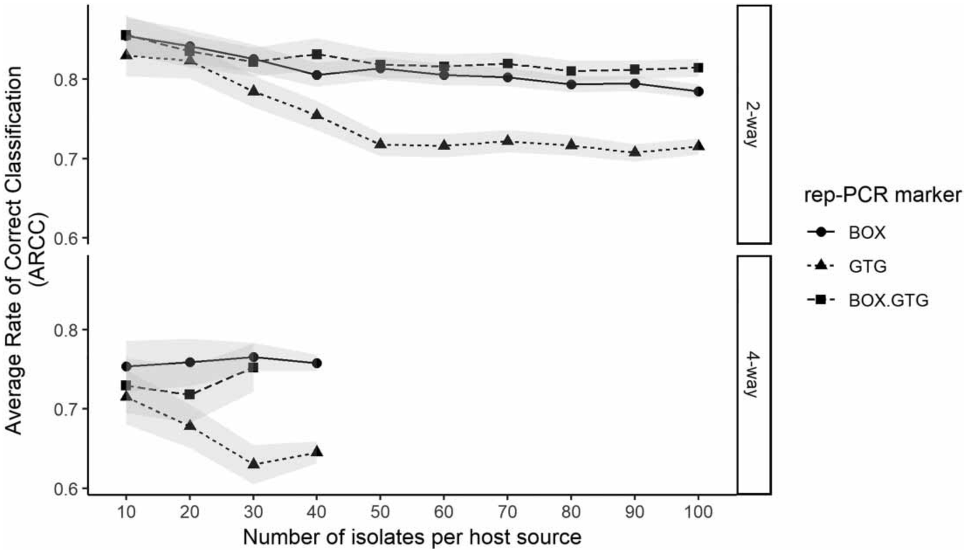

Figure 3 | Change in the ARCC as a function of the sample size. The dataset was partitioned based on the categorical split used. The shaded region represents the 95% CI obtained from 100 iterations of the classification model.

Table 2 | Assignment of *E. coli* isolates to host source in a two-way categorical split

			Rate of classification (%) of *E. coli* isolates in the assigned group		
Marker	**Host source**	***N***	**Agricultural**	**Domestic**	**ARCC (%)**
BOX	Agricultural	100	**82.00**	25.00	78.50
	Domestic	100	18.00	**75.00**	
GTG	Agricultural	100	**77.00**	33.00	72.00
	Domestic	100	23.00	**67.00**	
BOX–GTG	Agricultural	100	**84.00**	21.00	81.50
	Domestic	100	16.00	**79.00**	

Values shown are means obtained from 100 iterations of the DAPC. *N* indicates the number of isolates from each host source. The percentages of isolates that were correctly assigned are in bold.

contrast, GTG (0.63 ± 0.12) was significantly lower than the other two. These results suggest that for a two-way split, the best marker is BOX–GTG, whereas for a four-way split, it could either be BOX or BOX–GTG.

Library testing

To further assess its accuracy, the library was used in predicting the categories of a separate dataset consisting of composite BOX–GTG fingerprint profiles of *E. coli* from known sources. With the two-way categorical split, the library was able to correctly predict the isolates from agricultural and domestic sources with a classification rate of 67 and 64%, respectively (ARCC = 65.50%; Table 4). This accuracy decreased when using the four-way categorical split (ARCC = 29.17%; Table 5), with correct classification rates ranging from 13.33% (cow) to 53.33% (sewage).

DISCUSSION

In developing the culture library to source track fecal contamination in Laguna Lake, several factors were considered as suggested by Mott & Smith (2011). Firstly, *E. coli* was utilized as the microbial indicator since it is highly associated with the gut and feces of warm-blooded animals (Seurinck *et al.* 2003). Its presence also warns of the possible concurrent existence of other pathogenic microbes (Carson *et al.* 2003). This indicator has been utilized in many library-dependent MST, as reviewed by Mott & Smith (2011).

Secondly, the profiling method used here was rep-PCR. Discrimination of *E. coli* isolates based on the host origin using rep-PCR has been well documented (Dombek *et al.* 2000; Carson *et al.* 2003; McLellan *et al.* 2003; Seurinck *et al.* 2003; Mohapatra *et al.* 2007). This method has been utilized for source tracking fecal contamination in aquatic ecosystems such as beaches (Edge *et al.* 2010), natural ponds (Mohapatra *et al.* 2007), lakes (Kon *et al.* 2009), and watersheds (McLellan *et al.* 2003; Somarelli *et al.* 2007).

Table 3 | Assignment of *E. coli* isolates to host source in a four-way categorical split

			Rate of classification (%) of *E. coli* isolates in the assigned group				
Marker	**Host source**	***N***	**Chicken**	**Cow**	**Pig**	**Sewage**	**ARCC (%)**
BOX	Chicken	40	**67.50**	10.00	10.00	10.00	76.25
	Cow	40	17.50	**82.50**	5.00	10.00	
	Pig	40	12.50	5.00	**80.00**	7.50	
	Sewage	40	5.00	5.00	5.00	**75.00**	
GTG	Chicken	40	**70.00**	12.50	17.50	15.00	64.38
	Cow	40	7.50	**67.50**	12.50	12.50	
	Pig	40	17.50	15.00	**62.50**	15.00	
	Sewage	40	5.00	5.00	7.50	**57.50**	
BOX–GTG	Chicken	30	**73.33**	6.67	10.00	10.00	74.17
	Cow	30	10.00	**76.67**	10.00	10.00	
	Pig	30	10.00	10.00	**73.33**	6.67	
	Sewage	30	6.67	6.67	6.67	**73.33**	

Values shown are means obtained from 100 iterations of the DAPC. *N* indicates the number of isolates from each host source. The percentages of isolates that were correctly assigned are in bold.

Table 4 | Source tracking of *E. coli* isolates from known sources (test dataset) using the two-way library constructed from BOX–GTG fingerprint profiles

		Rate of classification (%) of *E. coli* isolates in the assigned group		
Test dataset	***N***	**Agricultural**	**Domestic**	**ARCC (%)**
Agricultural	100	**67.00**	33.00	65.50
Domestic	100	36.00	**64.00**	

N indicates the number of isolates from each host source. The percentages of isolates that were correctly assigned are in bold.

Table 5 | Source tracking of *E. coli* isolates from known sources (test dataset) using the four-way library constructed from BOX–GTG fingerprint profiles

		Rate of classification (%) of *E. coli* isolates in the assigned group				
Test dataset	***N***	**Chicken**	**Cow**	**Pig**	**Sewage**	**ARCC (%)**
Chicken	30	**20.00**	13.33	40.00	26.67	29.17
Cow	30	20.00	**13.33**	36.67	30.00	
Pig	30	16.67	26.67	**30.00**	26.67	
Sewage	30	16.67	0.03	26.67	**53.33**	

N indicates the number of isolates from each host source. The percentages of isolates that were correctly assigned are in bold.

The markers selected for assessment were BOX, GTG, and REP since these were reported to have the highest discriminatory power among the five rep-PCR methods commonly used for genotyping bacterial strains (Mohapatra *et al.* 2007; Mohapatra & Mazumder 2008). Among the three primers assessed in this study, our results show that REP performed poorly because of a low genotyping success rate and minimal information content. Similar observations were reported by Dombek *et al.* (2000) who generated REP fingerprint profiles that had a lower information content (25% fewer number of bands) compared to BOX and had isolates that were successful in generating fingerprint profiles with BOX but did not produce reliable fingerprints with REP.

Lastly, the statistical classification model employed was the DAPC developed by Jombart *et al.* (2010). In general, classification models are either unsupervised or supervised; the former requires no *a priori* grouping (e.g., cluster analysis and principal component analysis (PCA)), whereas the latter introduces the host source as a grouping variable (e.g., DA). No single approach has been shown to be superior to another (Scott *et al.* 2002; Graves *et al.* 2007; Stoeckel & Harwood 2007), and the appropriate statistical method depends on the nature of the library. Analyses performed on the dataset using two unsupervised classification methods, PCA and hierarchical clustering, revealed the absence of discrete clusters that could be useful for discriminating host sources (not shown). Given the binary nature of the data, the DA was also not usable since the correlation among variables was quite high. DAPC subjects the data to PCA to summarize the overall variability among individuals and then utilizes the PC scores for DA; this maximizes between-group variation while minimizing within-group variation, allowing for the best discrimination of samples to pre-defined groups. Although the classification model was originally developed for genetic data, it can be used with any dataset that is multivariate in nature (Jombart *et al.* 2010).

In terms of library performance, increasing library size improved the precision of estimates at the expense of accuracy. Although smaller libraries tend to have higher ARCCs, they suffer from lower representativeness which prevents the classification of isolates that are not in the library (Mott & Smith 2011). As much as possible, the library needs to consider all the isolates that are representative of a certain system for it to be effective in source tracking. Furthermore, the two-way performed better than the four-way categorical split. This is in concordance with previous reports showing that as the number of categorical splits increased, the accuracy decreased (Carson *et al.* 2003; Mohapatra *et al.* 2007). This is because a smaller number of categories leads to a greater probability that an isolate will be classified to its observed category by chance and that increasing the number of categories decreases this probability. For example, in a two-way split, a random classification will lead to 50% ARCC by chance, while for a four-way split, this becomes 25% (Mott & Smith 2011).

Testing the library using a separate dataset have shown its utility in actual source tracking. The two-way categorical split was able to identify the isolates correctly at a greater rate (ARCC = 65.50%) than that of the four-way categorical split (29.17%); this suggests that, in its current form, the library performs well only with the two-way split. However, we argue that the poor performance in the four-way split is a function of the number isolates representing each host source in the library. This underscores the importance of increasing the number of isolates per host source to further improve the library's accuracy.

Overall, fingerprint profiles generated by our isolates (Table 1) were different from other reports that used similar markers. Dombek *et al.* (2000) reported BOX and REP profiles containing 25–30 PCR bands on average with sizes ranging from less than 300 bp to about 4,500 bp. Carson *et al.* (2003) reported BOX profiles with 18–30 bands, while Seurinck *et al.* (2003) reported BOX profiles with 15–25 bands. McLellan *et al.* (2003) reported REP profiles with 13–22 bands ranging from 300 bp to 6 kb. This variability can be attributed to (1) the amplification protocol, (2) the host sources represented in the library, and (3) the spatiotemporal genetic variation of the indicator used. It was reported that the banding profile of rep-PCR differs with varying annealing temperatures (Korvin *et al.* 2014). In order to address such variability, we have performed an optimization where the annealing temperature was set to a gradient ranging from 47 to 57 °C; the median temperature of 52 °C was based on the protocol by Kheiri & Akhtari (2017) used in this study. The optimization was done for both BOX and GTG. Results showed that, despite the differences in the band intensity, consistent banding patterns from each representative host source isolate were recovered across the temperature gradient. This ruled out the possibility of banding pattern variation attributed to the amplification protocol. The spatiotemporal variability of profiles due to the genetic variation of the indicator is inherent to the MST-LDM (Ballesté *et al.* 2010; Mott & Smith 2011), thereby emphasizing the need to develop a library specific to the watershed of interest (McLellan *et al.* 2003).

Comparative studies of rep-PCR markers showed that different markers had different ARCC (Mohapatra *et al.* 2007; Mott & Smith 2011), while a composite analysis (i.e., combining band profiles among primers) improved overall discriminatory power (Yoke-Kqueen *et al.* 2013; Sukhumungoon *et al.* 2016), though that is not always the case (Dombek *et al.* 2000). Several papers also had different claims on which primer or primer combination generated the most reliable fingerprint data. Dombek *et al.* (2000) reported that BOX was considerably more accurate than REP in source tracking isolates from agricultural sources. Mohapatra *et al.* (2007) and Mohapatra & Mazumder (2008) reported GTG to have the greatest power in discriminating *E. coli* from various host sources (humans, poultry, and wild birds) as well as in differentiating populations from different aquatic environments. In concordance with our findings, Sukhumungoon *et al.* (2016) reported that either BOX or BOX–GTG combination was sufficient in identifying *E. coli* isolated from beef in southern Thailand. Variations in ARCC across different studies are to be expected since many factors affect library performance for the MST-LDM (Mott & Smith 2011).

CONCLUSION

We evaluated the discriminatory power of three rep-PCR markers on *E. coli* coming from various host sources that were identified to contribute fecal contamination in Laguna Lake. REP had a low genotyping success rate and information content and was thus dropped from the analyses. BOX had a higher discriminatory power than GTG; however, a combined profile from these markers had a higher ARCC. Our results indicate that BOX–GTG can be used as a rep-PCR marker in further developing the fingerprint library for MST in the lake. This future development includes increasing the number of isolates per host source as well as the inclusion of additional host sources when necessary. Since the most optimal primer combination has been selected, focus can now be given on assessing library representativeness, size, stability, and performance in identifying environmental isolates. It should be noted, however, that the LDM is not without limitations; MST should be performed using a toolbox approach; hence, there is a need to implement phenotypic fingerprinting methods (e.g., antibiotic resistance analysis) as well as LIMs.

ACKNOWLEDGEMENTS

This study was funded by the Department of Science and Technology – Philippine Council for Industry, Energy, and Emerging Technology Research and Development (DOST-PCIEERD) and is in collaboration with the Laguna Lake

Development Authority (LLDA). We thank the staff from the LLDA, Maynilad, Manila Water, and Laguna Water for their assistance in obtaining the samples and for providing valuable insights on this research. In addition, we thank the assistance provided by Jeffrey Thomas Beldia, Sherwina Paiso, Laurice Beatrice Raphaelle dela Peña, Nicole Bolo, Mark Raymond Vejano, and Jerro Paolo Lunas.

REFERENCES

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** (19), 4908–4917. https://doi.org/10.1016/j.watres.2009.09.047.

Ballesté, E., Bonjoch, X., Belanche, L. A. & Blanch, A. R. 2010 Molecular indicators used in the development of predictive models for microbial source tracking. *Applied and Environmental Microbiology* **76** (6), 1789–1795. https://doi.org/10.1128/AEM.02350-09.

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. https://doi.org/10.1128/AEM.69.3.1836-1839.2003.

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. https://doi.org/10.1128/AEM.66.6.2572-2577.2000.

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. https://doi.org/10.2166/wst.2010.335.

Garcia, B. C. B., Dimasupil, M. A. 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 Pesticides, Food Contaminants, and Agricultural Wastes* **50** (10), 734–743. https://doi.org/10.1080/03601234.2015.1048107.

Gomi, R., Matsuda, T., Matsui, Y. & Yoneda, M. 2014 Fecal source tracking in water by next-generation sequencing technologies using host-specific *Escherichia coli* genetic markers. *Environmental Science and Technology* **48** (16), 9616–9623. https://doi.org/10.1021/es501944c.

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. *Applied and Environmental Microbiology* **73** (15), 4857–4866. https://doi.org/10.1128/AEM.03003-06.

Graves, A. K., Hagedorn, C., Brooks, A., Hagedorn, R. L. & Martin, E. 2007 Microbial source tracking in a rural watershed dominated by cattle. *Water Research* **41** (16), 3729–3739. https://doi.org/10.1016/j.watres.2007.04.020.

Hagedorn, C., Robinson, S. L., Filtz, J. R., Grubbs, S. M., Angier, T. A. & Reneau, R. B. 1999 Determining sources of fecal pollution in a rural Virginia watershed with antibiotic resistance patterns in fecal streptococci. *Applied and Environmental Microbiology* **65** (12), 5522–5531.

Jombart, T. 2008 Adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics* **24** (11), 1403–1405. https://doi.org/10.1093/bioinformatics/btn129.

Jombart, T., Devillard, S. & Balloux, F. 2010 Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genetics*. https://doi.org/doi:10.1186/1471-2156-11-94.

Kheiri, R. & Akhtari, L. 2017 Clonal heterogeneity and efficacy of BOX and $(GTG)_5$ fingerprinting methods for molecular typing of *Escherichia coli* isolated from chickens in IRI. *Kafkas Universitesi Veteriner Fakultesi Dergisi* **23** (2), 219–225. https://doi.org/10.9775/kvfd.2016.16303.

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. https://doi.org/10.1139/W08-123.

Korvin, D., Graydon, C., McNeil, L. & Mroczek, M. 2014 Banding profile of rep-PCR experiments differs with varying extension times and annealing temperatures. *Journal of Experimental Microbiology and Immunology* **18**, 146–149.

Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., Cooper, T., Mayer, Z., Kenkel, B., Benesty, M., Lescarbeau, R., Ziem, A., Scrucca, L., Tang, Y., Candan, C. & Hunt, T. 2018 *caret: Classification and Regression Training. R Package Version 6.0-80.* Retrieved from: https://cran.r-project.org/package=caret.

Laguna Lake Development Authority 2018 *Existing Lake Uses*. Retrieved from: http://www.llda.gov.ph/index.php?option=com_content&view=article&id=110&Itemid=476 (accessed 5 February 2018).

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 Mcrobiology* **65** (5), 2587–2594. https://doi.org/10.1128/AEM.69.5.2587.

Mohapatra, B. R. & Mazumder, A. 2008 Comparative efficacy of five different rep-PCR methods to discriminate *Escherichia coli* populations in aquatic environments. *Water Science and Technology* **58** (3), 537–547. https://doi.org/10.2166/wst.2008.424.

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. https://doi.org/10.1111/j.1574-6968.2007.00948.x.

Mott, J. & Smith, A. 2011 Library-dependent source tracking methods. In: *Microbial Source Tracking Methods, Applications, and Case Studies* (C. Hagedorn, A. R. Blanch & V. J. Harwood, eds). Springer, New York, pp. 31–59.

Okabe, S., Okayama, N., Savichtcheva, O. & Ito, T. 2007 Quantification of host-specific Bacteroides-Prevotella 16S rRNA genetic markers for assessment of fecal pollution in freshwater. *Applied Microbiology and Biotechnology* **74** (4), 890–901. https://doi.org/10.1007/s00253-006-0714-x.

R Core Team 2017 *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from: https://www.r-project.org/.

Santos-Borja, A. & Nepomuceno, D. N. 2006 Laguna de Bay: institutional development and change for lake basin management. *Lakes & Reservoirs: Research and Management* **11** (4), 257–269. https://doi.org/10.1111/j.1440-1770.2006.00310.x.

Scott, T. M., Rose, J. B., Jenkins, T. M., Samuel, R., Lukasik, J. & Farrah, S. R. 2002 Microbial source tracking: current methodology and future directions microbial source tracking. *Applied and Environmental Microbiology* **68** (12), 5796–5803. https://doi.org/10.1128/AEM.68.12.5796.

Seurinck, S., Verstraete, W. & Siciliano, S. D. 2003 Use of 16S-23S rRNA intergenic spacer region PCR and repetitive extragenic palindromic PCR analyses of *Escherichia coli* isolates to identify nonpoint fecal sources. *Applied and Environmental Microbiology* **69** (8), 4942–4950. https://doi.org/10.1128/AEM.69.8.4942-4950.2003.

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. https://doi.org/10.1016/j.jenvman.2005.12.013.

Stoeckel, D. M. & Harwood, V. J. 2007 Performance, design, and analysis in microbial source tracking studies. *Applied and Environmental Microbiology* **73** (8), 2405–2415. https://doi.org/10.1128/AEM.02473-06.

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.

Versalovic, J., Koeuth, T. & Lupski, J. R. 1991 Distribution of repetitive DNA sequences in eubacteria and application to fingerprinting of bacterial genomes. *Nucleic Acids Research* **19** (24), 6823–6831.

WAVES 2016 *Pilot Ecosystem Account for Laguna de Bay Basin*. Retrieved from: https://www.wavespartnership.org/.

Wiggins, B. A., Cash, P. W., Creamer, W. S., Scott, E., Garcia, P. P., Gerecke, T. M., Han, J., Henry, B. L., Hoover, K. B., Johnson, E. L., Mccarthy, J. G., Mcdonough, J. A., Mercer, A., Noto, M. J., Park, H., Matthew, S., Purner, S. M., Smith, B. M., Erin, N., Varner, A. K., Dart, S. E., Jones, K. C., Mercer, S. A., Phillips, M. S. & Stevens, E. N. 2003 Use of antibiotic resistance analysis for representativeness testing of multiwatershed libraries. *Applied and Environmental Microbiology* **69** (6), 3399–3405. https://doi.org/10.1128/AEM.69.6.3399.

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 25 February 2019; accepted in revised form 30 November 2019. Available online 27 December 2019