Journal of Hazardous Materials 407 (2021) 124821

Contents lists available at ScienceDirect

Journal of Hazardous Materials

journal homepage: www.elsevier.com/locate/jhazmat

Research paper

Identification the source of fecal contamination for geographically unassociated samples with a statistical classification model based on support vector machine

Qiaowen Tan [a,b], Weiying Li [a,b,*], Xiao Chen [c]

[a] State Key Laboratory of Pollution Control and Resource Reuse, Tongji University, Shanghai 200092, China
[b] College of Environmental Science and Engineering, Tongji University, Shanghai 200092, China
[c] College of Defence Engineering, The Army Engineering University of PLA, Nanjing 210007, China

ARTICLE INFO

Editor: Dr. R. Debora

Keywords:
Machine learning
16S rRNA
Amplicon sequencing
Fecal source tracking
SVM

ABSTRACT

The bacterial diversity and corresponding biological significance revealed by high-throughput sequencing contribute massive information to source tracking of fecal contamination. The performances of classification models on predicting the fecal source of geographical local and foreign samples were examined herein, by applying support vector machine (SVM) algorithm. Random forest (RF) and Adaboost were applied for comparison as well. Discriminatory sequences were selected from *Clostridiale*, *Bacteroidales*, or *Lactobacillales* bacterial groups using extremely randomized trees (ExtraTrees). 1.51–12.64% of the unique sequences in the original library composed the representative markers, and they contributed 70% of the discrepancies between source microbiomes. The overall accuracy of the SVM model and the RF model on local samples was 96.08% and 98.04%, respectively, higher than that of the Adaboost (90.20%). As for the non-local samples, the SVM assigned most of the fecal samples into the correct category while several false-positive judgments occurred in closely related groups. The results in this paper suggested that the SVM was a time-saving and accurate method for fecal source tracking in contaminated water body with the potential capability of executing tasks based on geographically unassociated samples, and underlined the necessity of qPCR analysis for accurate detection of human source pollution.

1. Introduction

Fecal contamination of the natural water body has been recognized as a global environmental and sanitation problem due to the wide presence of pathogenic bacteria in feces. Many prevalent epidemics in history (such as cholera) were shown to be associated with fecal pollution of drinking water. In the recent coronavirus disease (COVID-19) pandemic, scientists found several cases whose stool specimens tested positive to the virus (Guan et al., 2020), indicating that a possible mode of transmission of this infectious disease is through the digestive system (Iacucci et al., 2020). Quick and accurate recognition of microbial pollution source is essential for environmental management and controlling public health risk since the severity of intestinal diseases caused by exposure of polluted water is closely related to the contamination source (Soller et al., 2010). For instance, wild birds and poultry have been shown to be the main contributors to *Campylobacter* pollution in surface water, which is the major pathogen accounting for human bacterial gastroenteritis (Mangen et al., 2015; Mulder et al., 2020).

Historically, researchers had been endeavoring to evaluate fecal contamination by fecal indicator bacteria (FIB) such as *Escherichia coli* and *Enterococcus* spp. However, due to the wide existence of FIB in the intestinal tract of warm-blooded animals, they do not provide sufficient information for fecal source tracking (Roguet et al., 2018). With the development of the next-generation sequencing (NGS) technology, a series of studies about community-based microbial source tracking (MST) have been carried out recently (Cao et al., 2013; Neave et al., 2014; Unno et al., 2012). In these studies, initial datasets were prepared by characterizing fecal communities in source samples using NGS of hypervariable regions of the 16S rRNA gene. Because of the differences in physiology characteristics and dietary habits, animals have developed distinctive intestinal microbiome during the long-term co-evolution with their intestinal microorganisms, and gradually formed unique

* Corresponding author at: State Key Laboratory of Pollution Control and Resource Reuse, Tongji University, Shanghai 200092, China.
E-mail address: 123lwyktz@tongji.edu.cn (W. Li).

https://doi.org/10.1016/j.jhazmat.2020.124821
Received 30 June 2020; Received in revised form 3 December 2020; Accepted 8 December 2020
Available online 11 December 2020

features for certain biotic populations.

Recently, many studies focused on both the application of SourceTracker software on MST tasks and its performance under various library configurations (Ahmed et al., 2015; Brown et al., 2019; O'Dea et al., 2019). SourceTracker is a machine learning classifier based on Bayesian theory, which was designed to estimate the contribution of microbial contamination in "sink" samples from "source" samples (Knights et al., 2011). Although researchers have proved that SourceTracker can be a useful tool for determining sources of aquatic bacterial contamination (Brown et al., 2017), and it can even be able to detect low-level signatures of sewage sources in a catchment (O'Dea et al., 2019), yet there are several limitations in the practical application of handling MST tasks: 1) Each new study using SourceTracker requires a large amount of time and resources to collect sufficient source and sink samples, thus restricting the feasibility to be widely utilized to some extent (Roguet et al., 2018). 2) SourceTracker cannot precisely analyze blinded freshwater samples spiked with source fecal material from a geographic region not represented in the initial library. In other words, local fecal samples (those collected from the same region/continent as the target samples) are required for the SourceTracker to build the initial library and assign host sources of new samples accurately (Staley et al., 2018). 3) Bayesian-based SourceTracker model is more like a "black--box". It means that this software cannot give further explanations to the contributions of pollution sources, resulting in a weak interpretability of the model.

For the above reasons, Roguet et al. (2018) firstly developed an alternative solution using random forest algorithm to classify fecal sources, offered a rapid and effective solution to differentiate host sources in environmental samples. However, it remained several questions to be investigated: 1) there are several widely-used supervised classification learning algorithms other than the random forest in the field of bioinformatics, including support vector machine (SVM), Adaboost, etc., and it is worth discussing whether these algorithms could perform better. 2) Whether these machine learning models can predict new samples (i.e. sink samples, which were collected from Australia in this study) using geographically non-local training set (i.e. source samples, which were collected from the USA in this study). Therefore, this work aims to evaluate the feasibility of a hybrid machine learning method (i.e. extremely randomized trees+ SVM) and compare the performances of SVM, RF and Adaboost algorithms for identifying local and non-local samples based on local training libraries.

2. Subjects and methods

2.1. Data collection

The raw sequencing files used for machine learning modeling and testing were obtained from the National Center for Biotechnology Information (NCBI). Samples from dataset A were collected from Florida, California & Minnesota, the USA. This dataset consisted fecal samples from goose (n = 19, n represents the number of samples, the same below), gull (n = 13), chicken (n = 15), dog (n = 14), cat (n = 13), dairycow (n = 15), beefcow (n = 5), deer (n = 15), swine (n = 17) as well as primary treated wastewater (n = 16) (Brown et al., 2017, 2019; Staley et al., 2018). Samples from beaver (n = 5) and rabbit (n = 6) were used as negative controls (Brown et al., 2017). Besides, blinded freshwater samples spiked with fecal materials (n = 29) were included to test the identification ability for blinded sources. The detailed information on the preparation methods of these spiked samples was documented in previous studies (Brown et al., 2019; Staley et al., 2018) and summarized in Tables S3 and S4. DNA was extracted from the samples of dataset A using the DNeasy PowerSoil DNA extraction kit (QIAGEN). Samples from dataset B were collected from Queensland, Australia, designed for testing the performances of models on predicting non-local samples. This dataset was comprised of chicken (n = 3), dog (n = 4), cow (n = 6), deer (n = 4), pig (n = 4) and untreated wastewater (n = 5) (O'Dea et al., 2019). DNA was extracted from the samples of dataset B using the QIAamp Power Fecal DNA Kit or the QIAamp Power Soil DNA Kit (QIAGEN). Amplicons of all the above samples were paired-end sequenced on the Illumina Hiseq 2500 (150 bp), and Miseq (300 bp). The sequencing results of dataset A are available under BioProject PRJNA377760 and PRJNA473286, and the accession numbers of the raw data of set B are SRP156322 & SRP118701.

2.2. Bioinformatics

The analysis of microbiome bioinformatics was performed with Qiime2 2019.7 (Bolyen et al., 2019). To ensure the consistency of length, the sequences obtained from the Illumina MiSeq platform runs were trimmed to 150 bp. Paired-end sequencing data was first demultiplexed by q2-demux plugin followed by primers removal using the cutadapt plugin (Martin, 2011) with a maximum allowable error rate of 0.2, all reads in which no primer was found were discarded. Low-quality base calls were filtered by q2-quality-filter (Bokulich et al., 2013). Clean sequences were processed by denoising with DADA2 (Callahan et al., 2016) to remove chimeras, join pair-end reads, and generate the high-resolution table of amplicon sequence variants (ASVs) (Eren et al., 2015) and representative sequences (Fig. S1).

After data pre-processing and denoising, ASVs aligned with mafft (Katoh et al., 2002) were used to construct a phylogeny with fasttree2 (Price et al., 2010). Rarefaction curves of alpha diversity indices (Figs. S2 and S3), beta diversity matrix by Jaccard dissimilarity, and Principal Coordination Analysis (PCoA) were produced or performed with q2-diversity plugin at the sampling depth of 32,000. Permutational multivariate analysis of variance (PERMANOVA) (Anderson, 2001) was conducted using R vegan. A sklearn Naïve Bayes machine learning classifier (q2-feature-classifier (Bokulich et al., 2018)) was used to assign taxonomy to each representative sequence. Firstly, the V5-V6 region of the 16 S rRNA gene (250 bp) was extracted from Greengenes 13_8 99% reference database (McDonald et al., 2012), then the Naïve Bayes classifier was trained using the reference reads just created. Finally, the classifier was used to predict the bacterial taxonomy of each ASV.

2.3. Modeling

2.3.1. Machine learning algorithms

Support vector machine (SVM) is a set of machine learning algorithms used for data mining tasks involved in classification and regression, developed by Vapnik (2013). SVM has been successfully utilized in many fields such as computer vision (Grauman and Darrell, 2005), and bioinformatics (Byvatov and Schneider, 2003; Dorff et al., 2010; Wang et al., 2019), etc. Compared to other methods, SVM 1) appears to be effective in handling high dimensional space, even where the number of dimensionalities (features) is greater than the number of samples, and 2) has a strong generalization ability, which means the SVM model could obtain satisfactory performance for predicting unknown samples (Ye et al., 2020). Based on the above pros, it could be hypothesized that the SVM model can achieve a good performance on community-based MST due to the presence of massive bacterial species in the fecal and environmental samples. Besides, the SVM only uses a subset of points in the training dataset (i.e. support vectors), so it is also a time-saving and memory-efficient algorithm (Vapnik, 2013; Zendehboudi et al., 2018).

For a binary classification task, it is intuitive to find a decision boundary that is "right in the middle" of the two types of sample points, because this boundary is most capable to tolerate the disturbances of samples. In other words, the classification performance of this boundary is hypothesized to be the most robust measure and has the strongest generalization ability for predicting new samples. According to the study via Boser et al. (1992), given training data$(x^i, y^i) \in D$, the basic idea of the SVM is to solve the following optimization problem:

$$\min_{w,b} \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{N}\xi^i \quad (1)$$
$$subject\ to: y^i(\omega^T\Phi(x^i)+b) \geq 1-\xi^i, i=1,2,\ldots,N$$

Where ω is the normal vector of the decision boundary, Φ maps x^i into a higher dimensionality using kernel trick, ξ and C are the relaxation variable and the regularization parameter, respectively. The boundary is only determined by several sample points (i.e. support vectors) in the outside of two classes.

The theory about random forest has been described in detail by Roguet et al. (2018) previously. In simple terms, when each decision tree in the forest growing, the split of nodes is determined by the maximum of mean decrease impurity (gini index). Therefore, an advantage of tree-based methods (including random forest) is that they could provide the degree of contribution for each feature (i.e. ASV) in differentiating two classes. In this study, extremely randomized trees (ExtraTrees) were practically used instead of random forest. When splitting nodes, random forest and ExtraTrees both use a subset of columns (features). However, the ExtraTrees randomly draw thresholds for each candidate feature and select the best of these thresholds as the splitting rule instead of directly selecting the most discriminative thresholds (like random forest does), to further introduce randomness (Geurts et al., 2006). This usually leads to an increase in the capability of generalization, but a slight increase in bias as expense (Geurts et al., 2006).

Adaboost is one of the most popular boosting ensemble algorithms, which was introduced by Freund and Schapire (1997). Simply speaking, the core idea of boosting is to fit a series of weak models (i.e., those are slightly better than random guessing) on modified versions of the data. The predictions from all of them are then combined through a weighted majority vote of these weak models to produce the final judgment. After each iteration, the algorithm will assign a larger weight to the misclassified samples in the previous iteration (so-called the modified version of data) to boost the performance of the ensemble model.

2.3.2. Supervised learning classification

A hybrid method of ExtraTreesClassifier + SVM Classifier was used herein. The modeling of supervised learning classification comprised two steps: Firstly, ExtraTrees were used to select the most important representative ASVs in discriminating one source from all other sources (Roguet et al., 2018). Simply speaking, for each candidate source, all samples were divided into two groups: samples belong to this candidate source and samples belong to other sources, a model was built to classify the two classes using ExtraTreesClassifier and generate importance indexes for all the ASVs, and selecting the most important ones (Fig. 1). After feature selection, every source was composed of ASVs that only appeared in samples of the corresponding source (host-specific) or ASVs of which abundances in these samples are significantly different from other sources (abundance-preferred). These are the most reliable indicators for differentiating two classes. Secondly, the SVM was used to build a classification model (base classifier) for each source using the representative ASVs mentioned above to generate a hyperplane (decision boundary) (Fig. 1). When the pollution source of an unknown sample A was to be predicted, the sequences matching the representative sequences in the classifier were extracted to compute the relative location of sample A in the sample space. There were two possibilities: 1) sample A located in the region of the corresponding source, 2) sample A

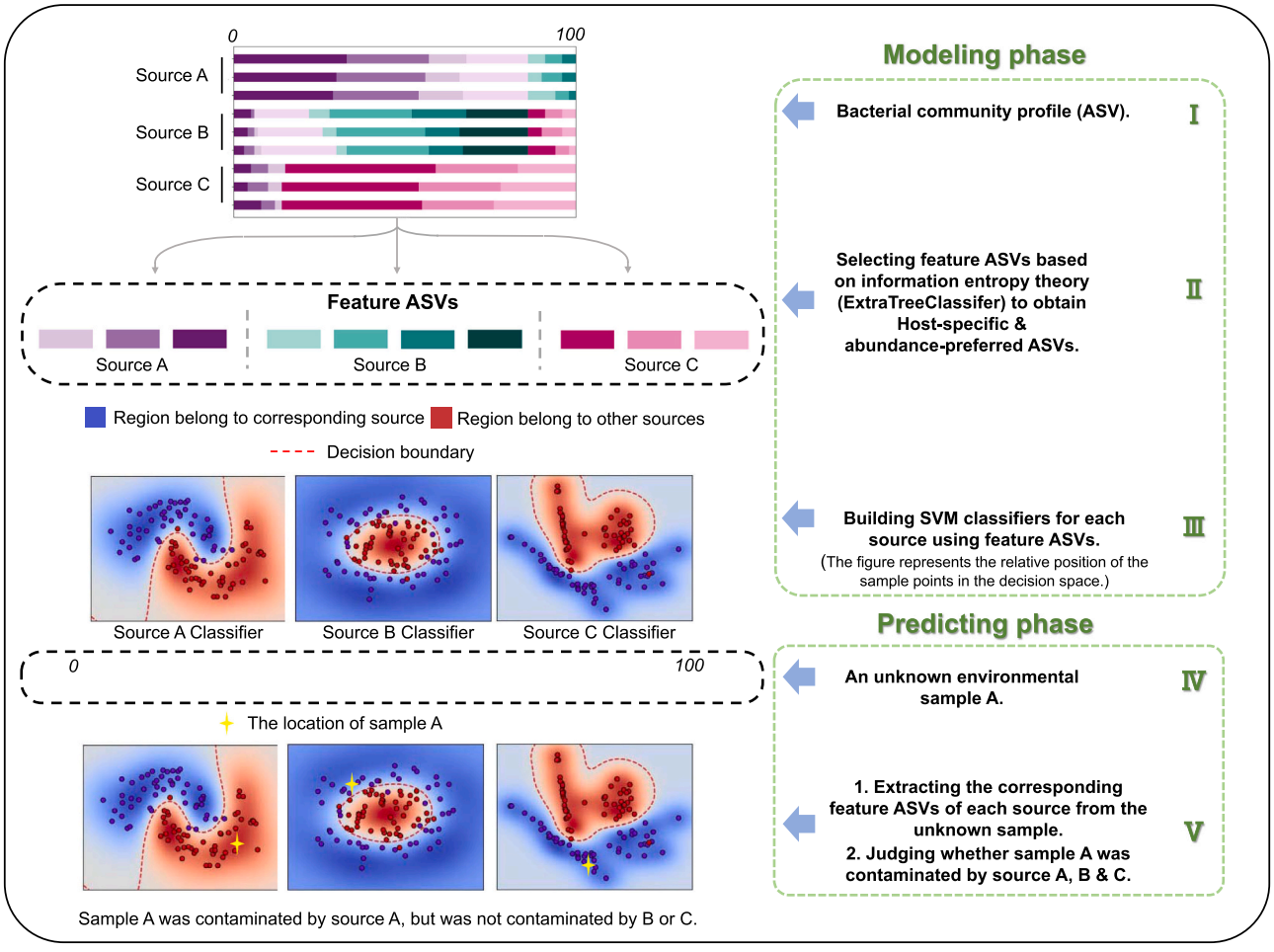


Fig. 1. Schematic diagram describing feature selection and SVM modeling processes.

located in the region of other sources. They corresponded to two judgments respectively: 1) sample A was polluted by this type of source, 2) sample A was not polluted by this type of source (Fig. 1). The distance from the decision boundary indicated the confidence of the judgment: the further the distance, the more likely it was to make a correct judgment. Each base classifier only focused on a subset of ASVs that was highly related to the corresponding source and independently judged the existence of candidate pollution sources. The overall result was composed of the outputs of all base classifiers (Fig. 1).

The models were all developed under Python Scikit-learn (sklearn) programming environment (Pedregosa et al., 2011). ExtraTrees model composed of 500 trees was created under default settings of *ExtraTreesClassifier* function in sklearn ensemble module to generate gini index (importance index), all the ASVs importance indices were sorted descendingly and then added up, the top 70% of the sum were selected (Fig. S2). Then the relative abundances of selected ASVs were calculated and the data were standardized using *scale* function in sklearn preprocessing module. The SVM model was built using sklearn svm.svc function, the rbf kernel (Gaussian kernel) was chosen. To estimate the probability for predicting unknown samples, the distance parameter generated by the API *descision_function* was mapped to 0–1 using the sigmoid function:

$$Sigmoid = \frac{1}{1 + e^{-x}} \tag{2}$$

The configuration of hyper-parameters of all the base classifiers was optimized by the *RandomizedSearchCV* function in the sklearn model-selection module, which used the accuracy rate as the estimator:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{3}$$

Where TP, FN, FP, TN represent the amount of true-positive, false-negative, false-positive and true-negative predictions, respectively. To evaluate the performances of base classifiers on the test dataset, metrics including precision, recall and F1 score were also adopted:

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

$$Percision = \frac{TP}{TP + FP} \tag{5}$$

$$F1 = \frac{2 \times Percision \times Recall}{Percision + Recall} \tag{6}$$

Besides, the *macro-average* strategy was considered to evaluate the overall performances of the models (i.e. the mean recall/precision rate of all base classifiers).

The running time of the modeling program of SVM, RF and Adaboost was listed in Table S5.

3. Results

3.1. Bacterial community profile and sample dissimilarity

The hosts of fecal samples included goose, chicken, gull, deer, cow (beef cow and dairy cow), swine, cat, dog, wildlife animals (beaver and rabbit) and urban wastewater, these samples were collected in Minnesota, the US and southeast Queensland, Australia. As for American samples, the taxonomic analysis for bacterial community composition revealed different characteristics of the microbial distributions from different host animals (on order level) (Fig. S5). The top 3 abundant orders were *Clostridiales* (41.03% ± 25.36%), *Bacteroidales* (16.21% ± 14.79%), and *Lactobacillales* (9.26% ± 18.73%). In general, *Clostridiales* spp. and *Bacteroidales* spp. were prevalent in all fecal samples except for gulls, part of samples from chicken and gulls were dominated by *Lactobacillales* spp. The abundance of the above three orders accounted for most of the overall bacterial community (66.50% ± 26.60%) in fecal samples (Fig. S5). Other bacterial orders only dominated in specific hosts, for example, sequences that were classified as *Erysipelotrichales* spp. were found to be abundant from pet samples (dogs and cats), and the dominance of *Turicibacterales* was only found in the feces of dogs.

Other than the results of taxonomy, PCoA using Jaccard dissimilarity clearly showed the separation and clustering of samples within and between host groups on the ASV level (Fig. 2). The first three coordinates accounted for 7.513%, 5.893%, and 5.016% of the total variance, which indicated that a certain part of the fecal samples was highly correlated considering that there were 153 sample points in the matrix. The samples from ruminant hosts including deer and cow were distributed alongside axis 1, whereas dogs, cat, and poultry (geese, gulls, and chickens) were linearly distributed in the plane consisting of axis 2 and axis 3. Also, the location of swine samples indicated that the microbial community composition of these samples was significantly different from other samples. Pairwise Adonis test confirmed the significant differences ($p < 0.05$) among host groups as well.

3.2. Feature selection for each classifier

Due to the prevalent and abundant presence of *Clostridiale*, *Bacteroidales*, and *Lactobacillales* in fecal samples, sequences belong to these orders were selected as candidates for feature selection. In consideration of the practical application, samples from nine fecal sources including dog, cow, cat, deer, gull, goose, chicken, swine, and wastewater were used to create eight classifiers (merging goose and chicken samples to create a "poultry" classifier). The number of ASV taxonomically classified as *Clostridiale*, *Bacteroidales*, or *Lactobacillales* was 2849, filtered from 51,099 ASVs of the original dataset. Each classifier was built by the most representative sequences selected by the extremely randomized trees. The importance index of each ASV and the decrease curves were plotted in Fig. S6, and the red points represented the position of the threshold of each classifier. The rapid decline of gini index, ranging from 0 to 500, indicated that the contribution of a small part of feature sequences was enough for differentiating a group of samples from others. The numbers of ASVs forming each classifier were 360 (poultry), 176 (cow), 170 (deer), 109 (dog), cat (66), swine (163), 43 (gull), and 53 (wastewater), respectively.

The PCA of all samples in each classifier was shown in Fig. 3. The total proportions of PC1 and PC2 in the total variance were between 18% and 83%, and the sample points from the corresponding source and other sources were completely linearly separable (except for the poultry classifier) in the plane consisted by the PC1 and PC2. Fig. 4b revealed the absolute abundances of ASVs (mean value) in each fecal source that were included in the corresponding classifier, in other classifiers, and not included in any classifier (explained by Fig. 4a), indicating that nearly half of the total abundance of the samples was responsible for discriminating a source from other sources. The abundances of sequences included in the corresponding classifier were higher than the sequences included in other classifiers. The distribution of shared ASVs between classifiers was visualized by Fig. 4c. The diagonal values represented the number of unique ASVs in each classifier, which showed that the major part (expect for gull and wastewater) of selected ASVs was host-specific. Other than these host-specific ASVs, Fig. 4b and c also suggested that the host-preferred ASVs occupied a certain proportion of the classifiers, these sequences represented those existed in multiple sources, but with different abundance patterns.

3.3. Performance of supervised learning algorithms on local and non-local samples

As described in Fig. 1, SVM models were applied after feature selection, RF and Adaboost classification models were also applied and tested for comparison. The original dataset obtained from the USA was randomly split into a training dataset ($n = 102$) and a test dataset

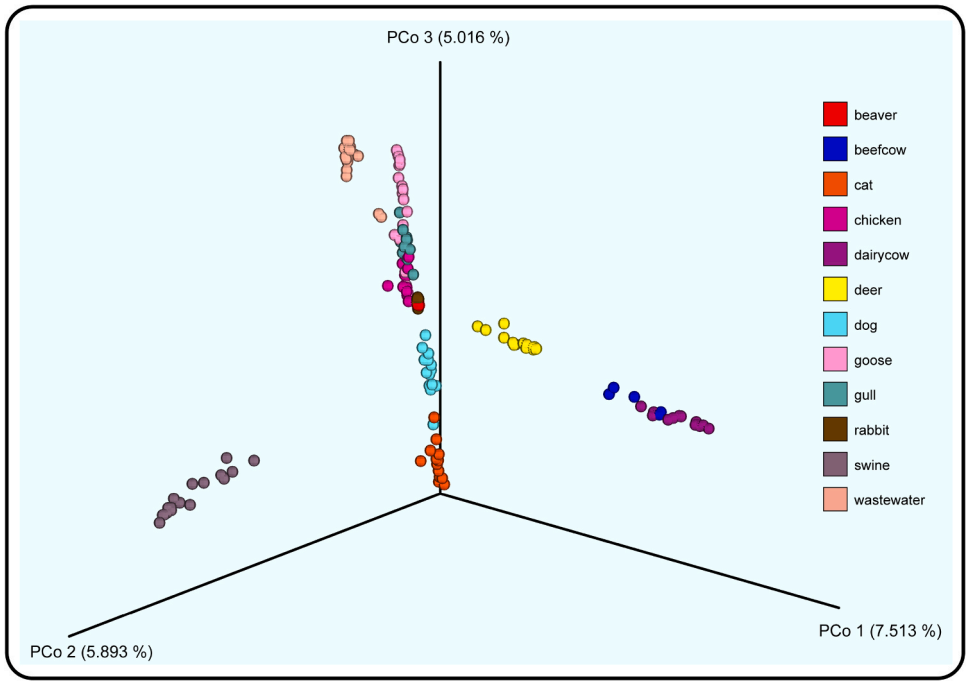


Fig. 2. Principal coordinates analysis using Jaccard dissimilarity matrix for all fecal samples collected in Minnesota, the United States of America.

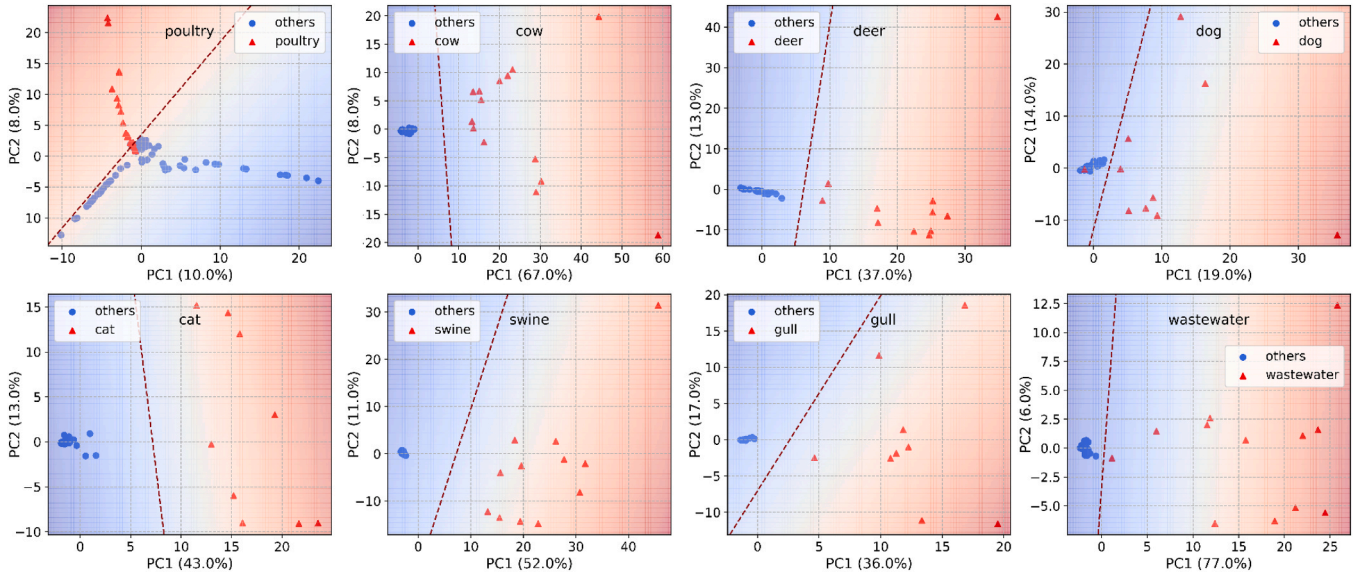


Fig. 3. PCA for the distribution of all samples in each classifier (the red line represents the linear decision boundary of two categories). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

($n = 51$) (Table S1). The randomness of splitting was proved by PCoA (Fig. S4) and Adonis test ($p = 0.991$, Table S2) between training and test datasets. They were used for the supervised learning process and the evaluation of the prediction performances, respectively. The counts of TP, FP and FN judgments by each classifier of each model were presented in Fig. 5a and recorded in Tables S9, S10 and S11 in detail. The overall performance of each model was shown in Fig. 5b. Generally speaking, the overall accuracy of the SVM model (96.08%) was close to that of the RF model (98.04%), exceeding that of the Adaboost model (90.20%). The difference between the SVM model and the RF model was that the SVM model was biased towards a higher false-positive rate, while the RF model was biased towards a higher false-negative rate. In other words, the SVM model had a higher recall rate while the RF model had a higher precision rate (Fig. 5b). Comparing the confidence levels of the SVM and the RF model for the prediction of the local U.S. samples (Tables S6 and S7), the probability values of correct judgments of the RF model were significantly higher than those of the SVM model (paired-sample T-test, $p < 0.05$), which further proved that the precision of RF model was higher, and it was more resistant to the false-positive judgments. In addition, beaver and rabbit samples were used as negative controls to test the specificity of the models. All 11 wildlife samples were correctly identified as "other sources" by SVM and RF models, three beaver samples were misclassified as *dog* by the Adaboost model (Tables S6, S7 and S8).

As for geographically non-local samples (from Australia), the threshold confidence of positive judgments was considered to be slightly

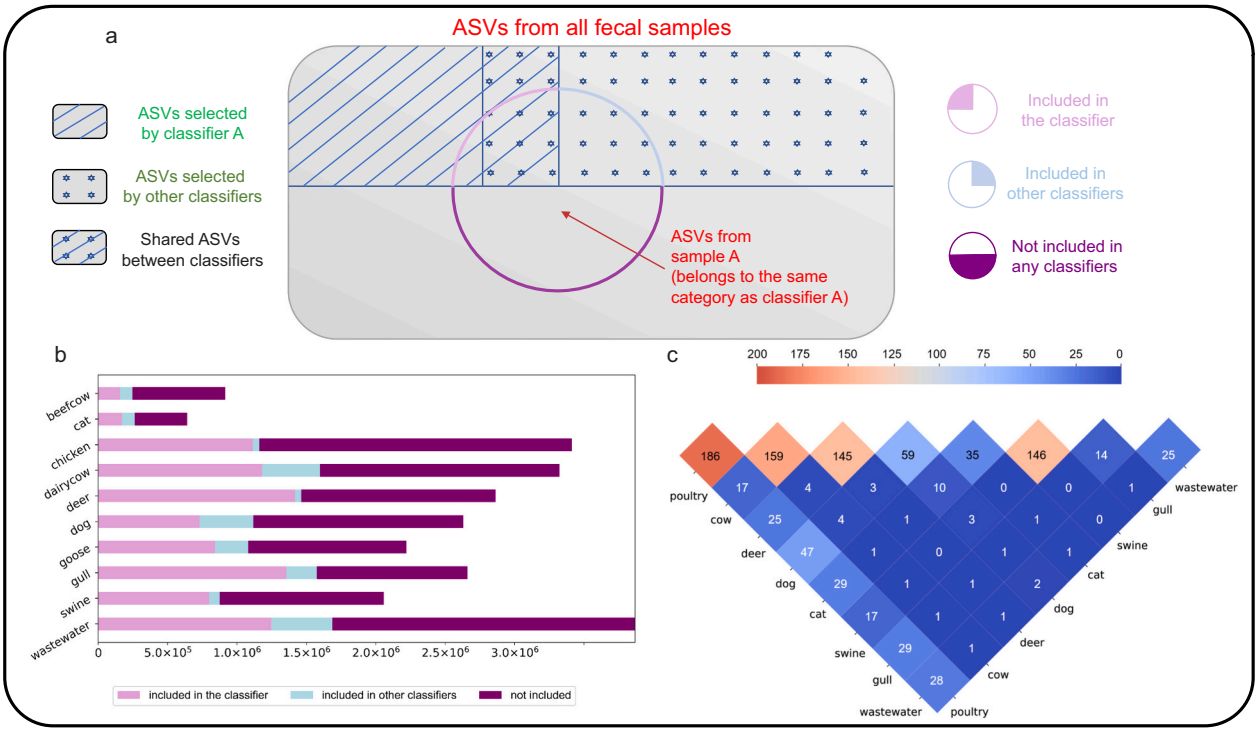
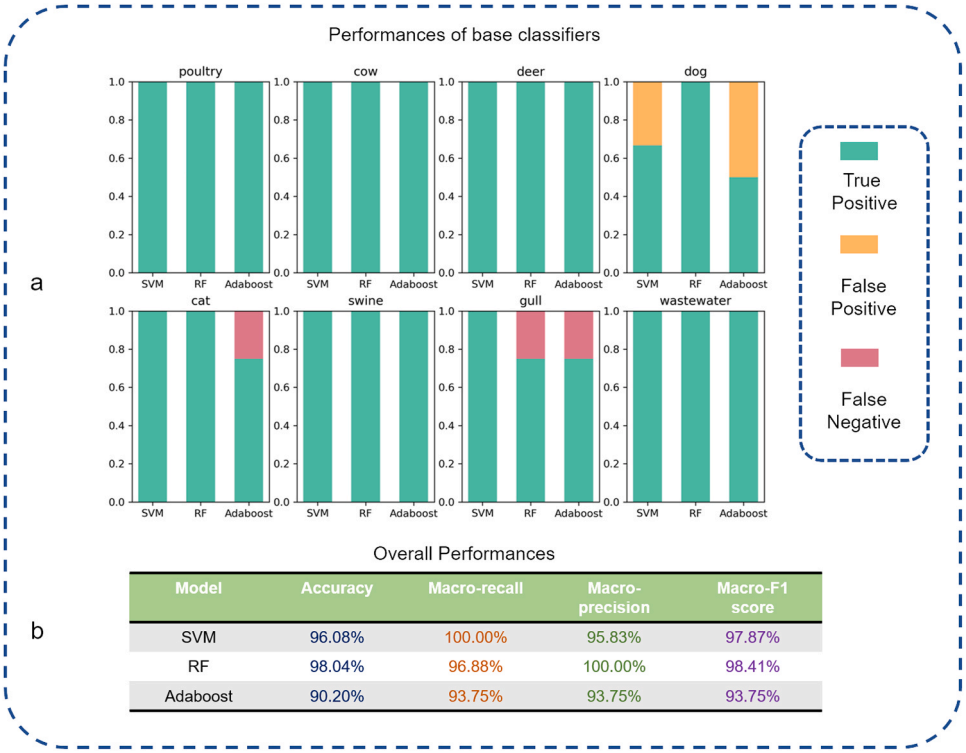


Fig. 4. (a) A Venn diagram that explains the meanings of "*included in the classifier*", "*included in other classifiers*", "*not included in any classifiers*" and "*shared ASVs*". (b) The absolute abundances of sequences belonging to the corresponding classifier, other classifiers or not belonging to any classifiers. (c) The numbers of shared ASVs between classifiers. The diagonal values represented the number of unique ASVs in each classifier.



Model	Accuracy	Macro-recall	Macro-precision	Macro-F1 score
SVM	96.08%	100.00%	95.83%	97.87%
RF	98.04%	96.88%	100.00%	98.41%
Adaboost	90.20%	93.75%	93.75%	93.75%

Fig. 5. (a) The performances of eight base classifiers of SVM, RF and Adaboost on the prediction of unknown fecal samples (local U.S.). To magnify the visual ratio of true positives, false negatives and false positives, the absolute numbers of true-negative judgments were excluded. (b) The overall performances of SVM, RF and Adaboost classifiers on the prediction of unknown fecal samples (local U.S.).

adjusted to 45%, in order to increase the recall rate of models. According to the prediction results, there were significant differences among the performances of SVM, RF, and Adaboost models on classifying these samples. Specifically, if cat and dog samples were considered to be in the same "pet" group, deer and cow samples were considered to be in the "ruminant animal" group, 20 (out of 21) animal fecal samples were classified correctly and/or classified as the source in the same group as the correct judgments by the SVM model (Table 1). Only five samples including three chicken samples and two dog samples were classified correctly by the RF samples, the rest were all false negatives. Misclassifications on three chicken samples and two cow samples occurred for the Adaboost model. Unfortunately, all three algorithms could not classify Australian sewage water samples correctly (Table 1). These results suggested that the SVM was most powerful among the three algorithms on the task of classifying geographical non-local samples.

3.4. Tracking the source of fecal pollution in environmental samples using the SVM model

The performances of the SVM model on environmental samples and artificial spiked water samples (simulating contaminated water) were tested and presented in Table 2. All freshwater samples showed no signature of contamination with animal feces or urban wastewater. Compared to other sources, Lakewater1, Lakewater4, Lakewater5, and Spikedwater02 were most likely to be polluted by poultry feces, but the probability values given by the model did not exceed 40% (Table 2). The contamination type of most of the artificial spiked water samples could be identified correctly by the SVM classifier, even when the proportion of contamination material was at a relatively low level (Table 2). As suggested by Spikedwater22, Spikedwater15, and Spikedwater28, when the animal sources existed, the pollution signature of them could cover the wastewater, causing the SVM to fail to identify sources of pollution from wastewater. But if wastewater was the only positive source of pollution, it could be accurately identified by the model (Table 2). In addition, the predictive results of the RF and Adaboost models on these samples were listed in Tables S12 and S13 for comparison. It can be seen that the RF and the Adaboost model were prone to the category of poultry (the phenomenon was extremely obvious with regard to the Adaboost model), resulting in more false-positive judgments in this category. Besides, the ability of the RF model on identifying the signal of cow and dog feces in water samples was weaker than that of the SVM model (Table S12).

Table 1
The prediction results of SVM, RF and Adaboost classifiers on unknown samples (Australia non-local)[c].

Sample id	SVM	RF	Adaboost
chickenAus1	Poultry (63.70%)[a]	Poultry (77.23%)[a]	Poultry (59.10%) + Dog (45.20%)
chickenAus2	Poultry (55.15%)[a]	Poultry (72.75%)[a]	Poultry (60.77%) + Dog (45.27%)
chickenAus3	Poultry (52.58%)[a]	Poultry (71.05%)[a]	Poultry (60.77%) + Dog (45.27%)
cowAus1	Cow (54.81%)[a]	—	Cow (73.10%)[a]
cowAus2	Cow (56.93%)[a]	—	Cow (73.10%)[a]
cowAus3	Cow (56.56%)[a]	—	Cow (73.10%)[a]
cowAus4	—	—	Poultry (47.27%) + Dog (45.20%)
cowAus5	Cow (56.16%)[a]	—	Cow (73.10%)[a]
cowAus6	Cow (56.07%)[a]	—	Dog (45.52%)
deerAus1	Deer (53.40%) + Cow (48.23%)[b]	—	Deer (73.10%)[a]
deerAus2	Deer (55.61%) + Cow (49.32%)[b]	—	Cow (73.10%)
deerAus3	Cow (52.02%) + Deer (47.59%)[b]	—	Cow (73.10%)
deerAus4	Cow (45.71%) + Deer (45.44%)[b]	—	Cow (73.10%) + Deer (73.10%)[b]
dogAus1	Dog (57.89%) + Cat (52.87%)[b]	Dog (71.06%)[a]	Dog (53.74%)[a]
dogAus2	Dog (58.68%)[a]	Dog (54.26%)[a]	Dog (56.54%)[a]
dogAus3	Dog (51.07%)[a]	—	Dog (53.52%)[a]
dogAus4	Dog (54.85%)[a]	—	Dog (56.34%)[a]
pigAus1	Swine (61.29%)[a]	—	Swine (73.10%)[a]
pigAus2	Swine (60.88%)[a]	—	—
pigAus3	Swine (59.71%)[a]	—	—
pigAus4	Swine (57.94%)[a]	—	—
sewageAus1	—	—	—
sewageAus2	—	—	—
sewageAus3	Cat (48.86%) + Wastewater (45.31%)	—	Dog (73.10%) + Poultry (46.88%)
sewageAus4	Cat (53.31%) + Wastewater (47.58%)	—	—
sewageAus5	—	—	Poultry (47.27%) + Dog (45.20)

[a] Judgments that were correct.
[b] Correct judgment and false-positive judgment occurred at the same time, but the false-positive judgments only occurred in the same animal group as the correct judgments. (Cat and dog were considered as domestic pets, deer and cow were considered as ruminant animals)
[c] Judgments with probability value $\geq$ 45.00% were defined as positive and presented in this table.

4. Discussion

4.1. The representative ASVs in specific bacterial orders provided sufficient microbial information for fecal identification

It has been well established that the tremendous diversity of intestinal bacterial composition is due to complicated interactions among factors of diet, geography, physiology, age (Bonder et al., 2016; Nishida and Ochman, 2018; Tigchelaar et al., 2016; Zhernakova et al., 2016), etc. The gut microbiome tends to adapt and assist in the metabolism processes of the host during long-term co-evolution. Nishida and Ochman (2018) pointed out that it was possible to differentiate hosts by their gut microbiomes even among recently diverged species with similar diet preferences. Since Jaccard dissimilarity was defined as the ratio of the intersection to the union of two sets (Hamers et al., 1989), the microbial structure presented in Fig. S1 could explain the clustering and separation of fecal samples showed in Fig. 2. For instance, chickens, geese and gulls tended to gather in similar positions (Fig. 2), which may be related to their shared order *Lactobacillales* and the species within, consistent with previously published work (Wei et al., 2013). Besides, the samples of domestic pets (dogs and cats) contained *Erysipelotrichales,* proved herein and in previous literature (Ahmed et al., 2015). This may result in a relatively close biological distance between these two groups, whereas the specific dominance of *Turicibacterales* in the feces of dogs may lead to the isolation of this category (Figs. S1 and Fig. 2).

Other than the orders mentioned above, the most abundant bacterial orders were *Clostridiales* and *Bacteroidales* in most fecal samples (Fig. 2). They were reported to be reliable groups that provided information for discriminating fecal sources (McLellan and Eren, 2014; Roguet et al., 2018). The strategy of focusing on a *narrow* taxonomy may have multiple practical advantages: 1) Although Staley et al. (2015) reported that the displayed taxonomic biases caused by different DNA extraction methods did not impact the overall biological conclusions drawn, this strategy can possibly minimize the impact of external factors (i.e. sample collection and sample processing (Knight et al., 2018)) and ensure the robustness of the model. 2) It could help to reduce unnecessary and redundant information produced by rarely appeared ASVs. 3) It may eliminate the impacts of bacterial assemblages that are prevalent in the environment in the identification of contamination source.

However, the problem is how to choose a suitable bacterial assemblage to perfectly balance the coverage of the bacterial library and the advantages mentioned above. According to the results of taxonomic annotation, the relative abundances of sequences belong to *Clostridiales* and *Bacteroidales* in the samples obtained from gulls were less than 5%,

Table 2
Prediction results for 29 freshwater or artificial spiked water using SVM classifier.

Sample id	Type of contamination	Proportion of contamination material % vol/vol	Prediction results	Source with max probability
Lake water1	None/Unknown[d]	—	—	Poultry (36.67%)
Lake water2	None/Unknown[d]	—	—	Wastewater (35.54%)
Lake water3	None/Unknown[d]	—	—	Gull (38.48%)
Lake water4	None/Unknown[d]	—	—	Poultry (31.35%)
Lake water5	None/Unknown[d]	—	—	Poultry (38.50%)
Lake water6	None/Unknown[d]	—	—	Gull (34.05%)
Spiked mesocosm1	WWTP[e]	30.0	Wastewater[a]	Wastewater (52.02%)
Spiked mesocosm2	WWTP[e]	30.0	—	Wastewater (37.86%)
Spiked mesocosm3	WWTP[e]	30.0	Wastewater[c]	Wastewater (42.94)
Spiked mesocosm4	WWTP[e]	30.0	—	Gull (37.58%)
Spiked mesocosm5	Cow	10.0	Cow[a]	Cow (59.51%)
Spiked mesocosm6	Cow	10.0	Cow[a] -Deer[a]	Cow (59.76%)
Spiked water10	Dog	1.0	Dog[c]	Dog (40.79%)
Spiked water08	WWTP[f]	10.0	Wastewater[a]	Wastewater (55.51%)
Spiked water06	WWTP[f]	0.1	Wastewater[a]	Wastewater (52.12%)
Spiked water21	Cat	0.1	Cat[a]	Cat (56.20%)
Spiked water22	Dog + WWTP[f]	0.5 + 0.5	Dog[b]	Dog (45.89%)
Spiked water19	Cat	10.0	Cat[a]	Cat (52.35%)
Spiked water17	Cow + WWTP[f]	0.5 + 0.5	Cow[a] -WWTP[c]	Cow (50.76%)
Spiked water15	Dog + WWTP[f]	0.05 + 0.05	Dog[a]	Dog (52.50%)
Spiked water16	Cat	1.0	Cat[a]	Cat (52.87%)
Spiked water33	Cow	0.1	Cow[a]	Cow (51.91%)
Spiked water13	Dog + WWTP[f]	5.0 + 5.0	—	Wastewater (27.45%)
Spiked water29	WWTP	1.0	Wastewater[a]	Wastewater (53.07)
Spiked water31	Cow	10.0	Cow[a]	Cow (58.99%)
Spiked water25	Dog	10.0	—	Dog (38.84%)
Spiked water28	Cat + WWTP[f]	0.5 + 0.5	Cat[a]	Cat (55.33%)
Spiked water04	Cow	1.0	Cow[a] -Deer[c]	Cow (58.54%)
Spiked water02	None/Unknown[d]	—	—	Poultry (31.48%)

[a] Representing the probability value generated by SVM model higher than 50%.
[b] Representing the probability value generated by SVM model between 45% and 50%.
[c] Representing the probability value generated by SVM model between 40% and 45%.
[d] Representing the samples that were obtained from natural water body, they were not contaminated or the contamination source was unknown.
[e] Representing untreated secondary wastewater effluent
[f] Representing primary-treated influent from wastewater treatment plants

but the abundances of *Lactobacillales* were up to 85%, which agreed with a previous study (Ahmed et al., 2015). Therefore, using only *Clostridiales* or *Bacteroidales* to build the classifiers was inappropriate, because the abundances of these bacteria could be greatly affected by external conditions. Comparing the accuracies on classifying local samples using RF model in this work and previous work (Roguet et al., 2018), selecting ASVs from *Clostridiales*, *Bacteroidales* and *Lactobacillales* assemblages could be a relatively proper choice. Despite the limited research results of present and a previous study (Roguet et al., 2018), it is necessary to further investigate the status of representative bacterial assemblages for MST tasks in future research.

According to the decreasing curve of importance indexes generated by the ExtraTrees Classifier (Fig. S2), 1.51–12.64% of the unique sequences provided 70% of the information in discriminating one source from others, i.e. a quite small number of bacteria species or strains represented most of the characteristics of a certain category. Through the study of these ASVs, a major proportion of them were found to be host-specific instead of host-preferred (Fig. 4c), and the host-specific ASVs were believed to offer more accurate classification due to their exclusive presence pattern. Besides, it is worth noting that the dominant status of these host-specific species in quantity happened to ensure the independent judgment of each base classifier to a certain extent, which helped the overall model to handle water samples with multiple pollution sources. The important status of selected ASVs in the fecal microbiome could be suggested by the fact that up to 50% of the absolute abundances of the total bacterial community were comprised of these sequences (Fig. 4b). The effectiveness of classification modeling based on the information entropy theory was reflected not only by the low error rate of the training samples but also the distribution pattern of the samples visualized by PCA. According to Fig. 3, the variance obtained by only the first two principal components contributed sufficient information to make it linearly separable for two categories, proving that these representative ASVs were important and reliable indicators of a certain pollution source.

4.2. Comparison of SVM model and sourcetracker toolbox on the application of identifying non-local samples

The average similarities of the bacterial community among the same species of which fecal samples were collected from different geographical locations were evaluated to be less than 20% by SourceTracker (Staley et al., 2018), indicating that this software showed unsatisfactory ability to characterize fecal samples without a geographically associated sample library. It has been well documented that geographic region has a great impact on the gut microbiome composition of the same animal host (Lozupone et al., 2012; Yatsunenko et al., 2012), resulting in the difficulties of source identification for samples from another location. As a matter of fact, we observed significant differences ($p < 0.01$) between Australian and U.S. samples as well by conducting pairwise Adonis tests (Table S14). However, the statistical parameter indicated that the microbiome differences between the same host from two locations (e.g. Australian dogs/U.S. dogs) were relatively smaller than different hosts (e.g. Australian dogs/U.S. cats) as shown by the relatively smaller F value (Table S14). There existed the probability that the distinctions in diet and intestinal factors between the same species in different locations were less than those between different species, which might explain this phenomenon. This fact suggested that it was possible to identify the source of fecal contamination for foreign samples with a statistical method. In this study, the SVM model successfully classified most of the fecal samples into the correct category or a closely related category, showing the powerful capability of generalization of this algorithm. As documented in previously published article (Staley et al.,

2018), highly abundant bacteria were identified as potentially discriminatory species and they were crucial markers contributing to the predictions of unknown samples as determined by SourceTracker. However, less abundant species may largely account for the particularity of one source (Holmes et al., 2012). Therefore, due to the characteristic of highly specific but less sensitive source identification, the Bayesian-based SourceTracker tool may have a lower generalization ability that could not enable it to accurately identify the source with a different microbiome (i.e. geographically non-local situation).

However, the limitations and uncertainties of the machine learning models involved in this study should be presented: 1) The prediction results of models were obviously prone to classifiers with higher coverage of ASVs. As shown in Tables S12 and S13, the RF and the Adaboost model were prone to output false-positive results of poultry due to the higher number of features selected by the poultry classifier. Balancing the number of ASVs and the contribution represented by these bacteria should be considered in the future. 2) Limited by the principle of algorithm, compared with the SourceTracker based on the Bayesian probability estimation theory, the SVM model in this article was difficult to provide the contribution rate of each source to the overall microbial pollution. Although Roguet et al. (2018) have tried to use the proportion of bacterial abundance of source samples contained by the sink sample to represent the contribution of pollution, the statistical significance of such characterization method still needs to be further explored. 3) Due to the limited types of available samples in this study, other configurations of samples with multiple pollution sources have not been convincingly verified in this study (Table 2), which should be further investigated in the future.

4.3. Maximize the recall of data to track the fecal contamination source using SVM algorithm on samples in hand

The continuing development and reduction of cost of the next-generation sequencing (NGS) prompted the researchers in the field of MST to pursue new biomarkers and methodologies (Holcomb and Stewart, 2020). The ability of NGS-MST to distinguish fine differences between pollution sources has been proved by several studies (Bauza et al., 2019; Staley et al., 2018).

In this study, the SVM algorithm showed its superior performance on the MST task using the community-based method. In the theory of statistical learning, the recall rate and the precision rate are negatively related in general cases (Mehta et al., 2019). In other words, the decrease in false-positive judgments often brings about the increase in false-negative classifications. It should be recognized that in the particular task of fecal source tracking, the cost for false-negative judgments (i.e. missing true contamination source) can be much greater than false-positive judgments. Therefore, we need to pursue the maximization of recall and accuracy rate in the application of MST based on machine learning algorithms. RF classifier was proved to be a robust model that could precisely locate a fecal sample, but it was also poor to capture the weak sign of feces in an unknown sample (Fig. 5 and Table S12). The performance of RF on the non-local dataset also supported this hypothesis (Table 1). In contrast, despite a few false-positive judgments occurred in closely related categories, the SVM model achieved the recall rate of 100% on predicting local samples, proving that it was a suitable choice among commonly-used algorithms.

It should be noted that all three models performed poorly in identifying non-local wastewater samples and the wastewater signatures in spiked water samples (Tables 1 and 2). In addition to the microbiome discrepancy due to the geographical inconsistency, different sampling section of the wastewater treatment plant (WWTP) may be an important reason (Table S4), due to the significant difference of bacterial community in each process (Cai et al., 2014). Therefore, validation by qPCR to assist the detection of human fecal signatures could be necessary. For example, human mitochondrial DNA was reported to be a promising target for fecal source tracking due to the high sensitivity and specificity (Holcomb and Stewart, 2020). Simply using this kind of markers as a means of verification for human fecal contamination or combining with the observation matrices of other sources for machine learning classification may significantly raise the detection rate of human source pollution (O'Dea et al., 2019).

As an economic bioinformatics application of sequencing the V5-V6 region of the 16S rRNA gene, the limited interpretation ability of the amplicon data may not represent all of the underlying biological significances (Zhang et al., 2018). Although the SVM model algorithm can be used for data mining tasks to reach the touchable ceiling of accuracy and recall rate, increasing the resolution and quality of the information hidden in the microbial complexity may dramatically improve the performance of the model. Therefore, taking the advantages of the enhanced ability to detect the diversity of bacterial composition and the functional profiles offered by Illumina metagenomics sequencing (Colston and Jackson, 2016) and more portable long-read sequencing platforms to characterize fecal pollution (Hu et al., 2018) could be an interesting future direction in community-based microbial source tracking.

5. Conclusion

This study demonstrated a community-based method for fecal pollution source tracking in water body using support vector machine algorithm. On the premise of high similarity and exclusivity of the microbiome in the same host, the SVM classifier can be an effective tool for source identification with the extension for assessment of large-scale high-throughput sequencing data. The usage of tree-based algorithm provided a group of highly discriminatory biomarkers for source identification. It eliminated much of the noise in the initial ASV matrix and promoted the robustness of the model. The SVM model was proved to be effective to identify fecal samples collected in Australia using training dataset that was comprised of U.S. samples, but the combination of qPCR analyses of marker genes and the community-based method is still necessary for the confirmation of human-sourced fecal pollution. For the need of rapid computational methods with low-resource demands, SVM classification method could serve as a useful tool for assessment of pathogen risk in the environmental water body, source forensics, emergency treatment, and medium/long-term pollution monitoring and control.

CRediT authorship contribution statement

Tan Qiaowen: Writing - original draft, Methodology, Software, Visualization, Formal analysis. **Li Weiying:** Conceptualization, Supervision, Resources, Validation. **Chen Xiao:** Visualization, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

We are grateful for the cooperation and participation of the utilities that were involved, which are supported by the Natural Science Foundation of China (Project No. 51979194) and 2019 Academician Strategic Consulting Research Project of Chongqing Academy of Engineering & Technology Development Strategy of China (Project No. 2019-CQ-ZD-1).

Appendix A. Supporting information

Supplementary data associated with this article can be found in the

online version at doi:10.1016/j.jhazmat.2020.124821.

References

Ahmed, W., Staley, C., Sadowsky, M.J., Gyawali, P., Sidhu, J.P.S., Palmer, A., Beale, D.J., Toze, S., 2015. Toolbox approaches using molecular markers and 16S rRNA gene amplicon data sets for identification of fecal pollution in surface water. Appl. Environ. Microbiol. 81 (20), 7067–7077.

Anderson, M.J., 2001. A new method for non-parametric multivariate analysis of variance. Austral Ecol. 26 (1), 32–46.

Bauza, V., Madadi, V., Ocharo, R.M., Nguyen, T.H., Guest, J.S., 2019. Microbial source tracking using 16S rRNA amplicon sequencing identifies evidence of widespread contamination from young children's feces in an urban slum of Nairobi, Kenya. Environ. Sci. Technol. 53 (14), 8271–8281.

Bokulich, N.A., Kaehler, B.D., Rideout, J.R., Dillon, M., Bolyen, E., Knight, R., Huttley, G. A., Caporaso, J.G., 2018. Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2′s q2-feature-classifier plugin. Microbiome 6, 90.

Bokulich, N.A., Subramanian, S., Faith, J.J., Gevers, D., Gordon, J.I., Knight, R., Mills, D. A., Caporaso, J.G., 2013. Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. Nat. Methods 10 (1), 57–59.

Bolyen, E., Rideout, J.R., Dillon, M.R., Bokulich, N., Abnet, C.C., Al-Ghalith, G.A., Alexander, H., Alm, E.J., Arumugam, M., Asnicar, F., Bai, Y., Bisanz, J.E., Bittinger, K., Brejnrod, A., Brislawn, C.J., Brown, C.T., Callahan, B.J., Caraballo-Rodriguez, A.M., Chase, J., Cope, E.K., Da Silva, R., Diener, C., Dorrestein, P.C., Douglas, G.M., Durall, D.M., Duvallet, C., Edwardson, C.F., Ernst, M., Estaki, M., Fouquier, J., Gauglitz, J.M., Gibbons, S.M., Gibson, D.L., Gonzalez, A., Gorlick, K., Guo, J., Hillmann, B., Holmes, S., Holste, H., Huttenhower, C., Huttley, G.A., Janssen, S., Jarmusch, A.K., Jiang, L., Kaehler, B.D., Bin Kang, K., Keefe, C.R., Keim, P., Kelley, S.T., Knights, D., Koester, I., Kosciolek, T., Kreps, J., Langille, M.G. I., Lee, J., Ley, R., Liu, Y.-X., Loftfield, E., Lozupone, C., Maher, M., Marotz, C., Martin, B.D., McDonald, D., McIver, L.J., Melnik, A.V., Metcalf, J.L., Morgan, S.C., Morton, J.T., Naimey, A.T., Navas-Molina, J.A., Nothias, L.F., Orchanian, S.B., Pearson, T., Peoples, S.L., Petras, D., Preuss, M.L., Pruesse, E., Rasmussen, L.B., Rivers, A., Robeson, M.S., Rosenthal II, P., Segata, N., Shaffer, M., Shiffer, A., Sinha, R., Song, S.J., Spear, J.R., Swafford, A.D., Thompson, L.R., Torres, P.J., Trinh, P., Tripathi, A., Turnbaugh, P.J., Ul-Hasan, S., vander Hooft, J.J.J., Vargas, F., Vazquez-Baeza, Y., Vogtmann, E., von Hippel, M., Walters, W., Wan, Y., Wang, M., Warren, J., Weber, K.C., Williamson, C.H.D., Willis, A.D., Xu, Z.Z., Zaneveld, J.R., Zhang, Y., Zhu, Q., Knight, R., Caporaso, J.G., 2019. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. In: Nat. Biotechnol., 37, pp. 852–857.

Bonder, M.J., Kurilshikov, A., Tigchelaar, E.F., Mujagic, Z., Imhann, F., Vila, A.V., Deelen, P., Vatanen, T., Schirmer, M., Smeekens, S.P., Zhernakova, D.V., Jankipersadsing, S.A., Jaeger, M., Oosting, M., Cenit, M.C., Masclee, A.A.M., Swertz, M.A., Li, Y., Kumar, V., Joosten, L., Harmsen, H., Weersma, R.K., Franke, L., Hofker, M.H., Xavier, R.J., Jonkers, D., Netea, M.G., Wijmenga, C., Fu, J., Zhernakova, A., 2016. The effect of host genetics on the gut microbiome. Nat. Genet. 48 (11), 1407–1412.

Boser, B.E., Guyon, I.M., Vapnik, V.N., 1992. A training algorithm for optimal margin classifiers, pp. 144–152.

Brown, C.M., Mathai, P.P., Loesekann, T., Staley, C., Sadowsky, M.J., 2019. Influence of library composition on source tracker predictions for community-based microbial source tracking. Environ. Sci. Technol. 53 (1), 60–68.

Brown, C.M., Staley, C., Wang, P., Dalzell, B., Chun, C.L., Sadowsky, M.J., 2017. A high-throughput DNA-sequencing approach for determining sources of fecal bacteria in a lake superior estuary. Environ. Sci. Technol. 51 (15), 8263–8271.

Byvatov, E., Schneider, G.J.Ab, 2003. Support vector machine applications in bioinformatics. Appl. Bioinform. 2 (2), 67–77.

Cai, L., Ju, F., Zhang, T., 2014. Tracking human sewage microbiome in a municipal wastewater treatment plant. Appl. Microbiol. Biotechnol. 98 (7), 3317–3326.

Callahan, B.J., McMurdie, P.J., Rosen, M.J., Han, A.W., Johnson, A.J.A., Holmes, S.P., 2016. DADA2: high-resolution sample inference from Illumina amplicon data. Nat. Methods 13 (7), 581–583.

Cao, Y., Van De Werfhorst, L.C., Dubinsky, E.A., Badgley, B.D., Sadowsky, M.J., Andersen, G.L., Griffith, J.F., Holden, P.A., 2013. Evaluation of molecular community analysis methods for discerning fecal sources and human waste. Water Res. 47 (18), 6862–6872.

Colston, T.J., Jackson, C.R., 2016. Microbiome evolution along divergent branches of the vertebrate tree of life: what is known and unknown. Mol. Ecol. 25 (16), 3776–3800.

Dorff, K.C., Chambwe, N., Srdanovic, M., Campagne, F.J.B., 2010. BDVal: reproducible large-scale predictive model development and validation in high-throughput datasets. Bioinformatics 26 (19), 2472–2473.

Eren, A.M., Morrison, H.G., Lescault, P.J., Reveillaud, J., Vineis, J.H., Sogin, M.L., 2015. Minimum entropy decomposition: unsupervised oligotyping for sensitive partitioning of high-throughput marker gene sequences. ISME J. 9 (4), 968–979.

Freund, Y., Schapire, R.E., 1997. A decision-theoretic generalization of on-line learning and an application to boosting. J. Comput. Syst. Sci. 55 (1), 119–139.

Geurts, P., Ernst, D., Wehenkel, L., 2006. Extremely randomized trees. Mach. Learn. 63 (1), 3–42.

Grauman, K., Darrell, T., 2005. The pyramid match kernel: discriminative classification with sets of image features, IEEE, pp. 1458–1465.

Guan, W.-j, Ni, Z.-y, Hu, Y., Liang, W.-h, Ou, C.-q, He, J.-x, Liu, L., Shan, H., Lei, C.-l, Hui, D.S.C., Du, B., Li, L.-j, Zeng, G., Yuen, K.-Y., Chen, R.-c, Tang, C.-l, Wang, T., Chen, P.-y, Xiang, J., Li, S.-y, Wang, J.-l, Liang, Z.-j, Peng, Y.-x, Wei, L., Liu, Y., Hu, Y.-h, Peng, P., Wang, J.-m, Liu, J.-y, Chen, Z., Li, G., Zheng, Z.-j, Qiu, S.-q, Luo, J., Ye, C.-j, Zhu, S.-y, Zhong, N.-s, 2020. Clinical characteristics of coronavirus disease 2019 in China. N. Engl. J. Med. 382 (18), 1708–1720.

Hamers, L., Hemeryck, Y., Herweyers, G., Janssen, M., Keters, H., Rousseau, R., Vanhoutte, A., 1989. Similarity measures in scientometric research - the Jaccard index versus Salton cosine formula. Inf. Process. Manag. 25 (3), 315–318.

Holcomb, D.A., Stewart, J.R., 2020. Microbial indicators of fecal pollution: recent progress and challenges in assessing water quality. Curr. Environ. Health Rep. 7 (3), 311–324.

Holmes, I., Harris, K., Quince, C., 2012. Dirichlet multinomial mixtures: generative models for microbial metagenomics. PLoS One 7 (2), e30126.

Hu, Y.O.O., Ndegwa, N., Alneberg, J., Johansson, S., Logue, J.B., Huss, M., Kaller, M., Lundeberg, J., Fagerberg, J., Andersson, A.F., 2018. Stationary and portable sequencing-based approaches for tracing wastewater contamination in urban stormwater systems. Sci. Rep. 8, 11907.

Iacucci, M., Cannatelli, R., Labarile, N., Mao, R., Panaccione, R., Danese, S., Kochhar, G. S., Ghosh, S., Shen, B., 2020. Endoscopy in inflammatory bowel diseases during the COVID-19 pandemic and post-pandemic period. Lancet Gastroenterol. Hepatol. 5 (6), 598–606.

Katoh, K., Misawa, K., Kuma, K., Miyata, T., 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 30 (14), 3059–3066.

Knights, D., Kuczynski, J., Charlson, E.S., Zaneveld, J., Mozer, M.C., Collman, R.G., Bushman, F.D., Knight, R., Kelley, S.T., 2011. Bayesian community-wide culture-independent microbial source tracking. Nat. Methods 8 (9), 761–763.

Knight, R., Vrbanac, A., Taylor, B.C., Aksenov, A., Callewaert, C., Debelius, J., Gonzalez, A., Kosciolek, T., McCall, L.-I., McDonald, D., Melnik, A.V., Morton, J.T., Navas, J., Quinn, R.A., Sanders, J.G., Swafford, A.D., Thompson, L.R., Tripathi, A., Xu, Z.Z., Zaneveld, J.R., Zhu, Q., Caporaso, J.G., Dorrestein, P.C., 2018. Best practices for analysing microbiomes. Nat. Rev. Microbiol. 16 (7), 410–422.

Lozupone, C.A., Stombaugh, J.I., Gordon, J.I., Jansson, J.K., Knight, R., 2012. Diversity, stability and resilience of the human gut microbiota. Nature 489 (7415), 220–230.

Mangen, M.-J.J., Bouwknegt, M., Friesema, I.H.M., Haagsma, J.A., Kortbeek, L.M., Tariq, L., Wilson, M., van Pelt, W., Havelaar, A.H., 2015. Cost-of-illness and disease burden of food-related pathogens in the Netherlands, 2011. Int. J. Food Microbiol. 196, 84–93.

Martin, M., 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet. J. 17 (1), 10–12.

McDonald, D., Price, M.N., Goodrich, J., Nawrocki, E.P., DeSantis, T.Z., Probst, A., Andersen, G.L., Knight, R., Hugenholtz, P., 2012. An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. ISME J. 6 (3), 610–618.

McLellan, S.L., Eren, A.M., 2014. Discovering new indicators of fecal pollution. Trends Microbiol. 22 (12), 697–706.

Mehta, P., Bukov, M., Wang, C.-H., Day, A.G.R., Richardson, C., Fisher, C.K., Schwab, D. J., 2019. A high-bias, low-variance introduction to machine Learning for physicists. Phys. Rep. Rev. Sect. Phys. Lett. 810, 1–124.

Mulder, A.C., Franz, E., de Rijk, S., Versluis, M.A.J., Coipan, C., Buij, R., Muskens, G., Koene, M., Pijnacker, R., Duim, B., Bloois, LvdG.-v, Veldman, K., Wagenaar, J.A., Zomer, A.L., Schets, F.M., Blaak, H., Mughini-Gras, L., 2020. Tracing the animal sources of surface water contamination with Campylobacter jejuni and Campylobacter coli. Water Res. 187, 116421.

Neave, M., Luter, H., Padovan, A., Townsend, S., Schobben, X., Gibb, K., 2014. Multiple approaches to microbial source tracking in tropical northern Australia. Microbiologyopen 3 (6), 860–874.

Nishida, A.H., Ochman, H., 2018. Rates of gut microbiome divergence in mammals. Mol. Ecol. 27 (8), 1884–1897.

O'Dea, C., Zhang, Q., Staley, C., Masters, N., Kuballa, A., Fisher, P., Veal, C., Stratton, H., Sadowsky, M.J., Ahmed, W., Katouli, M., 2019. Compositional and temporal stability of fecal taxon libraries for use with SourceTracker in sub-tropical catchments. Water Res. 165, 114967.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830.

Price, M.N., Dehal, P.S., Arkin, A.P., 2010. FastTree 2 - approximately maximum-likelihood trees for large alignments. PLoS One 5 (3), e9490.

Roguet, A., Eren, A.M., Newton, R.J., McLellan, S.L., 2018. Fecal source identification using random forest. Microbiome 6, 185.

Soller, J.A., Schoen, M.E., Bartrand, T., Ravenscroft, J.E., Ashbolt, N.J., 2010. Estimated human health risks from exposure to recreational waters impacted by human and non-human sources of faecal contamination. Water Res. 44 (16), 4674–4691.

Staley, C., Gould, T.J., Wang, P., Phillips, J., Cotner, J.B., Sadowsky, M.J., 2015. Evaluation of water sampling methodologies for amplicon-based characterization of bacterial community structure. J. Microbiol. Methods 114, 43–50.

Staley, C., Kaiser, T., Lobos, A., Ahmed, W., Harwood, V.J., Brown, C.M., Sadowsky, M. J., 2018. Application of SourceTracker for accurate identification of fecal pollution in recreational freshwater: a double-blinded study. Environ. Sci. Technol. 52 (7), 4207–4217.

Tigchelaar, E.F., Bonder, M.J., Jankipersadsing, A., Fu, J., Wijmenga, C., Zhernakova, A., 2016. Gut microbiota composition associated with stool consistency. Gut 65 (3), 540–542.

Unno, T., Di, D.Y.W., Jang, J., Suh, Y.S., Sadowsky, M.J., Hur, H.-G., 2012. Integrated online system for a pyrosequencing-based microbial source tracking method that targets bacteroidetes 165 rDNA. Environ. Sci. Technol. 46 (1), 93–98.

Vapnik, V., 2013. The Nature of Statistical Learning Theory. Springer Science & Business Media.

Wang, Y., Wang, S., Wu, C., Chen, X., Duan, Z., Xu, Q., Jiang, W., Xu, L., Wang, T., Su, L., 2019. Oral microbiome alterations associated with early childhood caries highlight the importance of carbohydrate metabolic activities. MSystems 4 (6) e00450–00419.

Wei, S., Morrison, M., Yu, Z., 2013. Bacterial census of poultry intestinal microbiome. Poult. Sci. 92 (3), 671–683.

Yatsunenko, T., Rey, F.E., Manary, M.J., Trehan, I., Dominguez-Bello, M.G., Contreras, M., Magris, M., Hidalgo, G., Baldassano, R.N., Anokhin, A.P., Heath, A.C., Warner, B., Reeder, J., Kuczynski, J., Caporaso, J.G., Lozupone, C.A., Lauber, C., Clemente, J.C., Knights, D., Knight, R., Gordon, J.I., 2012. Human gut microbiome viewed across age and geography. Nature 486 (7402), 222–227.

Ye, Z., Yang, J., Zhong, N., Tu, X., Jia, J., Wang, J., 2020. Tackling environmental challenges in pollution controls using artificial intelligence: a review. Sci. Total Environ. 699, 134279.

Zendehboudi, A., Baseer, M.A., Saidur, R., 2018. Application of support vector machine models for forecasting solar and wind energy resources: a review. J. Clean. Prod. 199, 272–285.

Zhang, J., Ding, X., Guan, R., Zhu, C., Xu, C., Zhu, B., Zhang, H., Xiong, Z., Xue, Y., Tu, J., Lu, Z., 2018. Evaluation of different 16S rRNA gene V regions for exploring bacterial diversity in a eutrophic freshwater lake. Sci. Total Environ. 618, 1254–1267.

Zhernakova, A., Kurilshikov, A., Bonder, M.J., Tigchelaar, E.F., Schirmer, M., Vatanen, T., Mujagic, Z., Vila, A.V., Falony, G., Vieira-Silva, S., Wang, J., Imhann, F., Brandsma, E., Jankipersadsing, S.A., Joossens, M., Cenit, M.C., Deelen, P., Swertz, M. A., Weersma, R.K., Feskens, E.J.M., Netea, M.G., Gevers, D., Jonkers, D., Franke, L., Aulchenko, Y.S., Huttenhower, C., Raes, J., Hofker, M.H., Xavier, R.J., Wijmenga, C., Fu, J., LifeLines Cohort, S., 2016. Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity. Science 352 (6285), 565–569.