KSABC
The Korean Society for Applied Biological Chemistry

ARTICLE Open Access

A comprehensive database of human and livestock fecal microbiome for community-wide microbial source tracking: a case study in South Korea

Hokyung Song[1] and Tatsuya Unno[1*]

Abstract

Fecal waste from livestock farms contains numerous pathogens, and improperly managed waste may flow into water bodies, causing water-borne diseases. Along with the popularization of high-throughput technologies, community-wide microbial source-tracking methods have been actively developed in recent years. This study aimed to construct a comprehensive fecal microbiome database for community-wide microbial source tracking and apply the database to identify contamination sources in the Miho River, South Korea. Total DNA was extracted from the samples, and the 16 S rRNA gene was amplified to characterize the microbial communities. The fecal microbiome database was validated by developing machine-learning models that predict host species based on microbial community structure. All machine learning models developed in this study showed high performance, where the area under the receiver operating characteristic curve was approximately 1. Community-wide microbial source tracking results showed a higher contribution of fecal sources to the contamination of the main streams after heavy rain. In contrast, the contribution of fecal sources remained comparatively stable in tributaries after rainfall. Considering that farms are more concentrated upstream of tributaries compared to the main streams, this result implies that the pathway for manure contaminants to reach the main streams could be groundwater rather than surface runoff. Systematic monitoring of the water quality, which encompasses river water and groundwater, should be conducted in the future. In addition, continuous efforts to identify and plug abandoned wells are necessary to prevent further water contamination.

Keywords Water contamination, Fecal microbiome, Livestock, 16S rRNA gene, Microbial source tracking (MST), Machine learning

*Correspondence:
Tatsuya Unno
tatsu@cbnu.ac.kr
[1]Department of Biological Sciences and Biotechnology, Chungbuk National University, Seowon-Gu, Cheongju 28644, Republic of Korea

Introduction

Global demand for meat has increased significantly, resulting in extensive livestock farming and increased manure generation [1]. A large amount of manure is stockpiled before it is used as fertilizer or temporarily stored before it is transported to livestock waste facilities for energy production and disposal [2–6]. Many farms, owing to a lack of indoor spaces, store livestock manure in open areas where it can leach into stream waters through surface runoff or into groundwater. Because fecal materials from livestock contain a myriad of pathogens that may cause waterborne diseases, it is important to properly manage livestock manure and track leaching events to prevent disease outbreaks [2, 7].

Microbial source tracking (MST) is a collection of methods used to discriminate fecal pollution sources in aquatic environments using microbes [8]. Traditional MST methods include genomic fingerprinting methods such as repetitive-element palindromic PCR (rep-PCR) [9, 10] and pulsed-field gel electrophoresis (PFGE) [11–13]. More recently, quantitative PCR (qPCR) has been widely used for MST because it is culture-independent. Numerous qPCR markers have been designed, including HF183 and BacHum, which detect human-specific *Bacteroides* [14, 15], Pig2Bac, which detects pig-specific *Bacteroidales* [16], and GFD, which detects bird-specific *Helicobacter* spp [17].

With the advent of high-throughput sequencing methods, community-wide approaches to MST have emerged. Knights et al. [18] introduced a bioinformatics tool for community-wide MST called SourceTracker. SourceTracker models the contributions of source communities to the contamination of sink communities. Staley et al. [19] reported the applicability of SourceTracker for MST through double-blind tests using samples spiked with one to five source libraries. Unlike previous MST methods, which have limited use in the detection of predetermined fecal indicator bacteria, community-wide approaches directly estimate source proportions with much higher resolution.

Community-wide MST methods are powerful, but can sometimes be resource-intensive because of the increased size of microbial community sequence data. Recently, other machine learning-based community-wide source-tracking bioinformatics tools, such as FEAST (fast expectation-maximization for microbial source tracking) [20] and STENSL (Microbial Source Tracking with Environment SeLection [21] have been developed to overcome a few of the drawbacks of SourceTracker. FEAST was developed based on a highly efficient expectation maximization-based method that enables community-wide MST on time. STENSL identifies true contributing sources and reduces the noise introduced by noncontributing sources by incorporating sparsity into the model.

In South Korea, 142,155 tons of livestock manure are generated daily [22]. To understand the contribution of fecal sources to water contamination, we conducted a case study in the Miho River, South Korea, where the average number of total coliforms in 2022 reached 16,870 CFU (colony forming unit)/100 mL [23] (at sampling point MH10 [Fig. 1]). We first developed a comprehensive fecal microbiome database and validated it using machine learning models that predict host species based on fecal microbial community structures. Based on the constructed database, we performed a community-wide MST to track the contamination sources of the Miho River in South Korea. We aimed to diagnose the current status of fecal pollution in detail, identify its causes, and suggest appropriate control methods.

Materials and methods

Sample collection and physicochemical measurements

In total, 633 fecal samples (125 human samples, 144 poultry samples, 116 swine samples, 42 horse samples, and 206 cow samples) were collected in Jeju and Gwangju, South Korea, between 2016 and 2020 (Supplementary Table S1) and stored at -20 °C before DNA extraction. River samples were collected from the mainstream (MH08, MH09, SY01, BR01, and MH10) and tributaries (BC02, JO02, and WH01) of the Miho River watershed, located across Cheongju (upstream) and Sejong (downstream) in South Korea, where the three major livestock species are cows, pigs, and poultry (chicken and duck) (Supplementary Table S2) (Fig. 1). Three replicate samples were collected at each sampling point before and after the heavy rain on June 26th, 2023 (daily rainfall of 34.1 mm) (Supplementary Fig. S1). Water temperature, pH, and electrical conductivity (EC) were measured using a multifunction meter CX-401 (Elmetron, Poland). Water samples were filtered using cellulose nitrate filters (pore size of 0.45 μm and diameter of 47 mm) (Whatman, UK) and stored at -20 °C before DNA extraction.

DNA extraction and high-throughput sequencing

Fecal DNA was extracted using either a PowerFecal Isolation Kit (MOBIO, Carlsbad, CA, USA) or a QIAamp PowerFecal DNA Kit (Qiagen, Germany). DNA was extracted from the filters using a DNeasy PowerWater Kit (Qiagen, Germany). The amplicon sequencing library targeting the V4 (or V3-V4) region of the bacterial 16 S rRNA gene was prepared according to the “16S Metagenomic Sequencing Library Preparation” guidelines provided by Illumina [24]. For the amplification of the V3–V4 region, we used the primer sets 341 F (5′-CCTACGGGNGGCWGCAG-3′) and 805R (5′-GACTACHVGGGTATCTAATCC-3′) and to amplify the V4 region, we used the primer sets 515 F (5′-GTGCCAGCMGCCGCGGTAA-3′) and 806R (5′-GGACTACHVGGGTWTCTAAT-3′) [25].

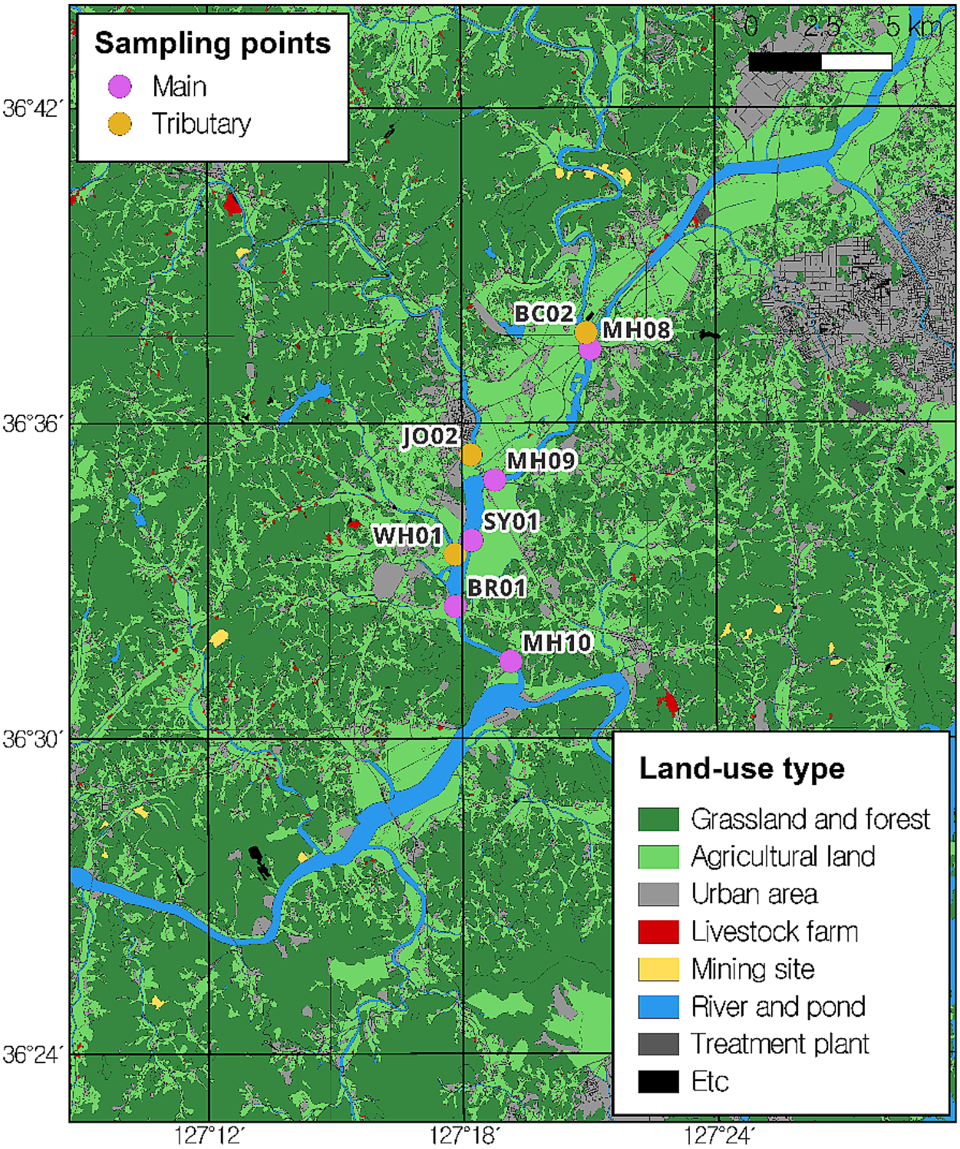

Fig. 1 A map showing land-use type of the study area and sampling points. Water flows from the north to the south. Land-use information was collected from the National Geographic Information Institute of Korea. Map image was generated through QGIS 3.34.3

The pooled library was sent to Macrogen Inc. (Seoul, South Korea) for sequencing. Fastq-formatted sequence data have been deposited in the Sequence Read Archive under project ID PRJNA1071275 for Miho River samples and PRJNA1071195 for fecal samples, except for 21 human samples that have already been published in another study [26] (Sequence Read Archive project ID of PRJNA544370).

Sequence processing

Sequences were processed using Mothur software [27] following MiSeq SOP (https://mothur.org/wiki/miseq_sop/). Sequences with ambiguous base pairs and homopolymers (>8 base pairs) were removed. Sequences of <250 bp or >550 bp were removed. Sequences were aligned against the Silva database v. 138 [28], and the "pcr.seqs" command (with a start option of 11,895 and an end option of 25,316, which covers the 515–805 bp region of the bacterial 16 S rRNA gene) was used for the consistency between V4 amplicons and V3–4 amplicons. The chimeric sequences were removed using the VSEARCH algorithm [29]. Sequences were classified based on RDP database v. 18 [30], and the sequences annotated as "Chloroplast," "Mitochondria," "unknown," and "Eukaryota" were removed. Sequences with a similarity greater than 97% were clustered into operational taxonomic units (OTUs) using the OptiClust algorithm [31].

Statistical analysis, machine learning, and microbial source tracking

To test if the microbial community structure varies significantly depending on their hosts, a pairwise permutational multivariate analysis of variance (PERMANOVA) test was performed using the "vegan" package [32] and the "ranacapa" package [33] in R. Before performing pairwise PERMANOVA, we subsampled 10,575 reads per sample and calculated the Bray–Curtis distance between samples based on the square-root transformed OTU abundance data. To visualize the distance between the samples, we generated a non-metric multidimensional (nMDS) plot using the "vegan" package in R.

We constructed machine learning classification models to predict hosts based on microbial communities. The relative abundances of the genera were used as features (independent variables) of the models, and five hosts (poultry, cow, horse, human, and pig) were used as traits (dependent variables). Unclassified genera were excluded when building machine learning models. We used five machine-learning algorithms: (1) random forest (RF) [34], (2) extreme gradient boosting (XGBoost) [35], (3) support vector machine (SVM) [36], (4) logistic regression (Logr), and (5) K-nearest neighbor (KNN) [37]. The Python "xgboost" module was used to construct the XGBoost model, and the "scikit-learn" module [38] was used to construct the other four models. Hyperparameters were tuned using the "GridSearchCV" function in "sklearn.model_selection," except for the XGBoost model, in which we used the default values due to resource limitations. The samples were randomly divided into training and test data 100 times, and the models were evaluated using 5-fold cross-validation. To find out important features in the random forest model, the "feature_importances_" attribute in the Python "scikit-learn" package was used.

Community-wide MST was performed at the OTU-level using the "FEAST" function in the R "FEAST" package [20]. Fecal microbiome data were used as sources, and Miho River data were used as sinks. The source contribution values of the fecal samples were summed for each host.

Results

Community composition of the fecal samples and the Miho River freshwater samples

At the phylum level, Firmicutes comprised over 60% of the gut microbiome in poultry samples and were dominant in other fecal sources (Fig. 2A). Bacteroidetes were dominant in fecal samples, except in poultry. The freshwater samples collected before heavy rain were dominated by Proteobacteria, Bacteroidetes, and Cyanobacteria. In contrast, freshwater samples collected after heavy rainfall were dominated by Proteobacteria, followed by Actinobacteria and Bacteroidetes.

At the genus level, *Prevotella* was dominant in human and pig samples but not in the other samples (Fig. 2B). In human samples, *Phocaeicola*, *Bacteroides*, *Faecalibacterium*, and *Bifidobacterium* were dominant. *Lactobacillus* was dominant in both pig and poultry samples. In poultry samples, *Romboutsia* and *Streptococcus* were dominant. Cow samples were dominated by *Phocaeicola* and *Alistipes*, whereas horse samples were dominated by *Treponema* and *Methanocorpusculum*. No overlap exists in the major (top five) genera between the fecal and freshwater samples.

The nMDS results showed strong clustering for each sample group (Fig. 2C). PERMANOVA results showed significant differences between the different sample groups (global $R^2 = 0.51637$, $p < 0.001$; pairwise test results are shown in Supplementary Table S3). The microbial community in freshwater samples shifted towards the fecal microbiomes after heavy rain (Fig. 2C, Supplementary Table S3). The horse samples were most distantly located in the freshwater samples on the nMDS plot.

Verification of the fecal microbiome database using machine learning

To evaluate the fecal microbiome database constructed in this study, we built machine-learning models that predicted hosts based on fecal microbial community composition. The hyperparameters used to construct the final models are listed in Supplementary Table S4. All five machine learning classification models performed well, with areas under the receiver operating characteristic curves of approximately 1 (Fig. 3A, Supplementary Fig. S2, and Table S5). The most important 20 features identified in the RF model included the major genera represented in Fig. 2B, including *Faecalibacterium*, *Alistipes*, *Methanobrevibacter*, *Treponema*, and *Romboutsia*, and minor genera such as *Paeniclostridum*, *Clostridioides*, *Paludibacter*, *Monoglobus*, and *Ihubacter* (Fig. 3B and C).

Community-wide microbial source tracking of the miho river

The water temperature was approximately 24–27 °C during the sampling period, both before and after the heavy rain (Table 1). The pH and EC decreased after rainfall at most sampling points. The MST results demonstrated that freshwater samples were contaminated with the fecal microbiome of humans, cows, pigs, and poultry to a minor extent before heavy rain (Fig. 4). However, the source contributions of human, cow, pig, and especially poultry samples increased after rain in the mainstream. In contrast, there were relatively smaller changes in the source contribution profiles of tributary samples, such as BC02 and WH01, after the rain. Overall, the downstream

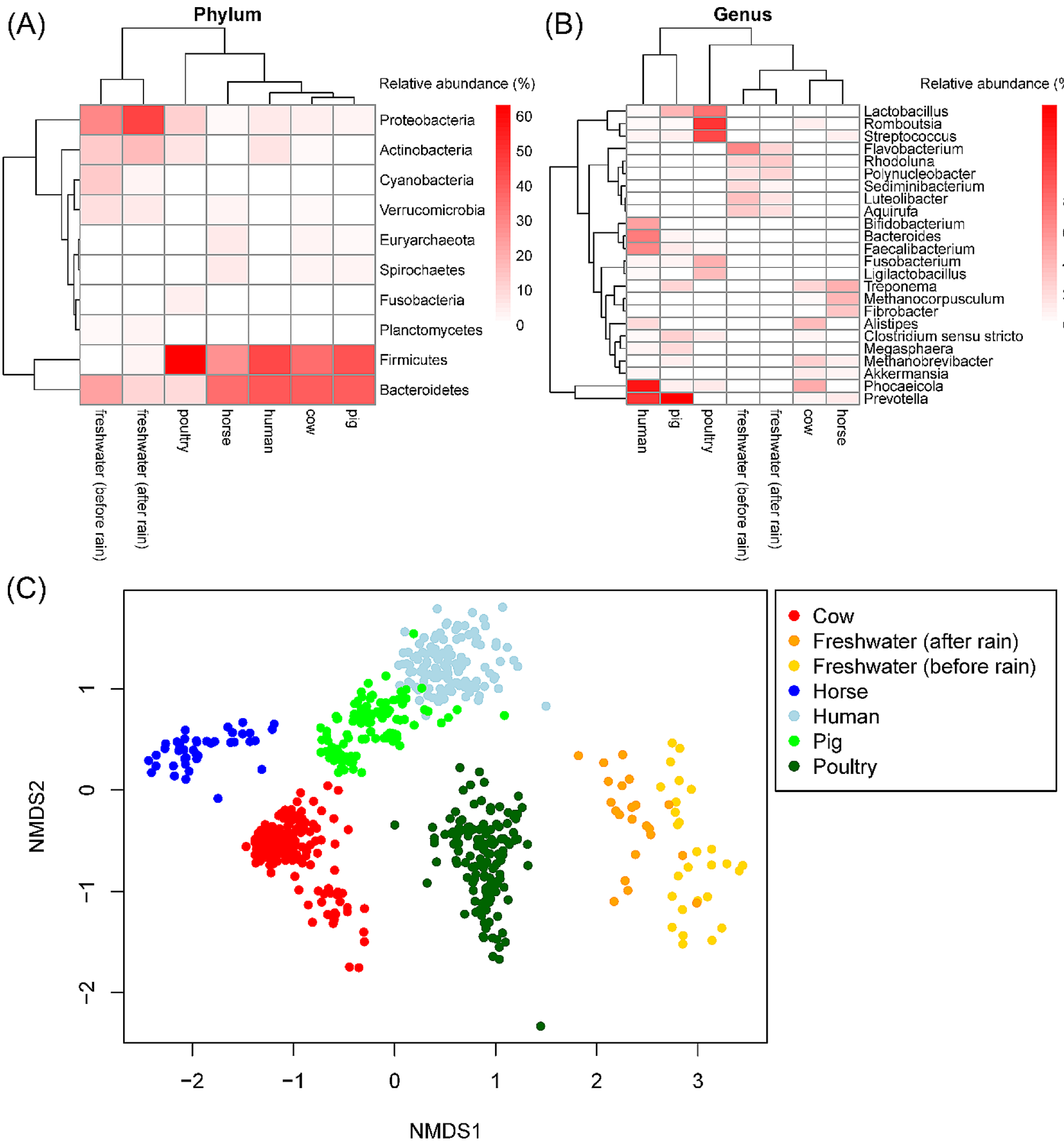

Fig. 2 (**A**) Phylum-level composition of the studied samples. Top 5 phyla (except unclassified) for each sample group were chosen. (**B**) Phylum-level composition of the studied samples. Top 5 genera (except unclassified) for each sample group were chosen. (**C**) An nMDS plot generated based on the OTU-level composition of the samples

samples (BR01 and MH10) showed a highly contaminated profile after rain compared to the other samples. Regardless of sampling time, the horse fecal microbiome had nearly zero contribution.

Discussion

In this study, we constructed a comprehensive fecal microbiome database based on 633 fecal samples collected from poultry, cows, horses, pigs, and humans in South Korea for community-wide MST. The database constructed in this study can save time and effort in collecting fecal samples and facilitate comparative studies.

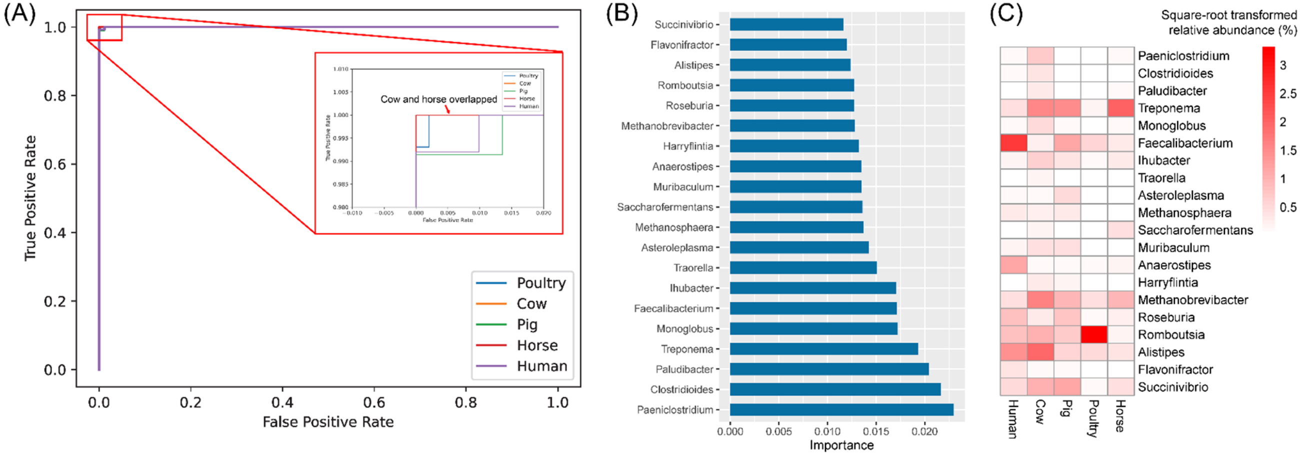

Fig. 3 (**A**) Receiver operating characteristic (ROC) curve of the random forest (RF) model. (**B**) Importance values of the 20 most important features of the constructed RF model. (**C**) Square-root transformed relative abundance of the important features

Table 1 Physicochemical conditions of the Miho River before and after heavy rain

Sample ID	Temperature (°C)		pH		Electrical Conductivity (μS/cm)	
	Before	After	Before	After	Before	After
BC02	26.7 ± 0.4	27.1 ± 0.1	8.5 ± 0	7.1 ± 0	379 ± 7.9	278.3 ± 0.6
MH08	25.1 ± 0	26.5 ± 0.1	7.9 ± 0	6.7 ± 0	548.2 ± 4.4	354.9 ± 11.1
MH09	26.4 ± 0.3	26.5 ± 0.1	7.8 ± 0.3	6.6 ± 0	491.3 ± 10.9	303.7 ± 5.5
JO2	26.1 + 0	24.8 + 0.9	7.3 + 0	6.8 + 0	390.9 + 8.4	498.1 + 6.4
SY01	24.7 ± 0.2	27 ± 0.1	8.2 ± 0	6.7 ± 0	554.5 ± 12.8	267.9 ± 9
WH01	24.7 ± 0.1	24.3 ± 0.1	6.8 ± 0.1	6.7 ± 0	223 ± 8.8	220.6 ± 3.4
BR01	25 ± 0.1	25.8 ± 0.1	8.2 ± 0	6.7 ± 0	547.8 ± 1.2	317.5 ± 6.5
MH10	26.4 ± 0.7	25.6 ± 0.2	7.8 ± 0	6.8 ± 0	582.6 ± 0.6	250.4 ± 2.7

Traditional MST methods, such as rep-PCR and PFGE, have a high possibility of producing false negatives (failure to identify a source when present), as the target indicator bacteria (*Escherichia coli*, *Enterococci*, etc.) constitute only a small proportion of the overall community. This issue can be resolved in community-wide MST, as it does not target a single bacterial species but instead considers multiple species collaboratively.

Traditional MST methods have limitations in distinguishing different host species from contamination sources. In this study, the fecal microbiomes were distinguishable by different host groups, as reported in many other studies [39, 40]. The machine learning models constructed in this study showed nearly 100% accuracy in predicting host groups based on fecal microbial community structures. The features that contributed most to accurate prediction in the RF model included the major genera and the minor ones, which have often been neglected. This indicates that these minor genera can function as important indicators and help enhance the resolution of community-wide MST.

The MST results for the Miho River revealed that humans, chickens, cows, and pigs were the main contributors to fecal contamination. The contribution of poultry samples was generally higher than that of other sources, particularly after rainfall. This could be due to the high number of poultry farms and poultry individuals in the study area (Supplementary Table S2). Horses contributed marginally to the contamination of the Miho River both before and after the rain. This corresponds to the low number of horse farms and horses in the study area (Supplementary Table S2).

Tributary samples, such as WH01 and BC02, were only minimally contaminated before and after the rain, even though a few livestock farms were located upstream of these tributaries (Fig. 1). In contrast, in the mainstream, especially downstream (BR01 and MH10), there was more severe contamination after rain, even though few livestock farms were located nearby. This suggests that the contamination of the mainstream may not originate from surface water flushing from tributaries or nearby surface water, but likely from groundwater. In South Korea, it has been estimated that there are more than one million abandoned tubular wells [41]. Abandoned wells function as direct channels for surface contaminants to pollute groundwater; therefore, an appropriate plugging method is generally required. In the study area, it has been reported that the quality of groundwater originating from abandoned tubular wells is lower than that of general groundwater [41]. Although the local governments of Cheongju and Sejong have made huge efforts to plug abandoned tubular wells annually, the results of this study suggest that there are still many unmanaged wells that contaminate groundwater and, subsequently, the mainstream of the Miho River. Further source tracking of groundwater could be helpful for assessing the contamination source locations in detail.

The results of this study show that the aquatic environment in farming areas can be contaminated by diverse fecal sources after rainfall. A strong demand exists for proper monitoring plans and surveillance systems to improve water quality. In addition, government support for proper livestock waste storage systems may be helpful. Further efforts are necessary to identify and plug the tubular wells.

In this study, we identified the possible causes of water contamination by applying community-wide microbial source tracking methods. In recent years, more sophisticated MST methods have been developed, such as SNV-FEAST, which uses single nucleotide variants for MST [42]. However, these metagenome-based methods are resource-intensive and require high-performance servers, which limits their range of applications. Currently, 16 S rRNA gene-based community-wide source tracking is a reliable and cost-effective MST method. The database construction and validation methods used here and the case study of the Miho River can be applied to other source-tracking studies and can aid policy decision-making processes. As the fecal microbiome can vary geographically [43, 44], region- or country-specific databases must be developed before performing community-wide MST.

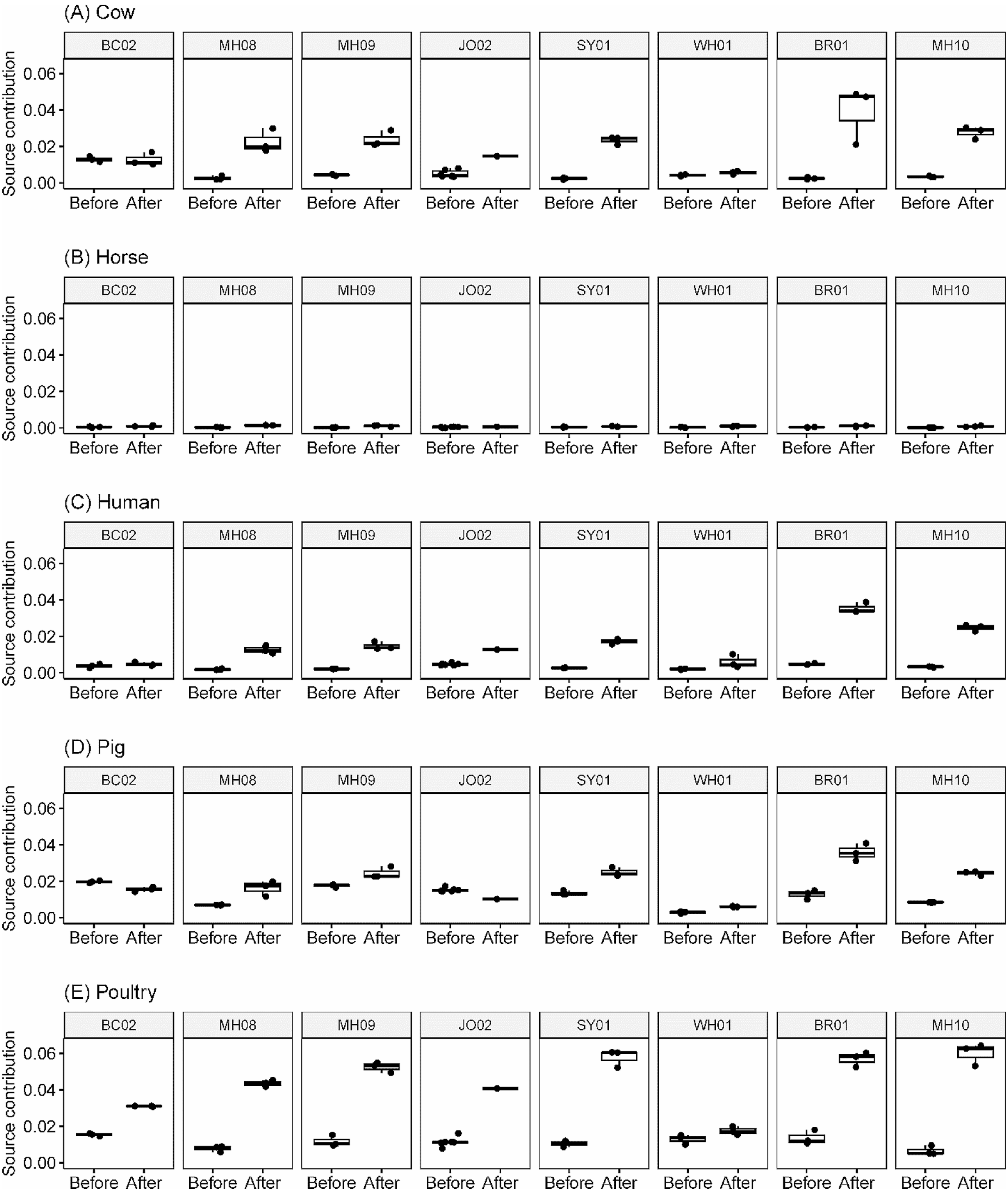

Fig. 4 Boxplots representing the source contributions of the fecal microbiome on Miho River samples

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13765-024-00915-5.

Supplementary Material 1

Supplementary Material 2

Acknowledgements

We would like to thank Editage [http://www.editage.com] for editing and reviewing this manuscript for English language.

Author contributions

Hokyung Song: Conceptualization, Methodology, Formal analysis, Investigation, Data curation,

Writing original draft. Tatsuya Unno: Conceptualization, Methodology, Writing - review & editing, Funding acquisition, Supervision.

Funding

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean Ministry of Science and ICT (NRF-2021R1A2C1094426). This work was supported by the research grant of the Chungbuk National University in 2023.

Data availability

Fastq-formatted sequence data have been deposited in the Sequence Read Archive under project ID PRJNA1071275 for Miho River samples and PRJNA1071195 for fecal samples, except for 21 human samples that have already been published in another study [26] (Sequence Read Archive project ID of PRJNA544370).

Declarations

Conflict of interest

The authors declare no conflict of interest. Tatsuya Unno is an Associate Editor of Applied Biological Chemistry. Associate Editor status has no bearing on editorial consideration.

Received: 20 May 2024 / Accepted: 16 June 2024

Published online: 24 June 2024

References

1. Varma VS, Parajuli R, Scott E, Canter T, Lim TT, Popp J, Thoma G (2021) Dairy and swine manure management – challenges and perspectives for sustainable treatment technology. Sci Total Environ 778:146319. https://doi.org/10.1016/j.scitotenv.2021.146319
2. Alegbeleye OO, Sant'Ana AS (2020) Manure-borne pathogens as an important source of water contamination: an update on the dynamics of pathogen survival/transport as well as practical risk mitigation strategies. Int J Hyg Environ Health 227:113524. https://doi.org/10.1016/j.ijheh.2020.113524
3. De Rosa D, Biala J, Nguyen TH, Mitchell E, Friedl J, Scheer C, Grace PR, Rowlings DW (2022) Environmental and economic trade-offs of using composted or stockpiled manure as partial substitute for synthetic fertilizer. J Environ Qual 51(4):589–601. https://doi.org/10.1002/jeq2.20255
4. Janni K, Cortus E (2020) Common Animal Production Systems and Manure Storage Methods. Animal Manure. pp. 27–43
5. Khoshnevisan B, Duan N, Tsapekos P, Awasthi MK, Liu Z, Mohammadi A, Angelidaki I, Tsang DCW, Zhang Z, Pan J, Ma L et al (2021) A critical review on livestock manure biorefinery technologies: sustainability, challenges, and future perspectives. Renew Sustain Energy Rev 135:110033. https://doi.org/10.1016/j.rser.2020.110033
6. Parihar S, Saini K, Lakhani G, Jain A, Roy B, Ghosh S, Aharwal B (2019) Livestock waste management: A review
7. Lee J-H, Yun S-T, Yu S, Yoo C-H, Jeong Y-S, Kim K-H, Kim H-R, Kim H (2022) Development of an integrated hydrochemical index for delineating livestock manure-derived groundwater plumes in agro-livestock farming areas. Ecol Indic 138:108838. https://doi.org/10.1016/j.ecolind.2022.108838
8. Rock C, Rivera B, Gerba CP (2015) Chap. 14 - Microbial Source Tracking. In: Pepper IL, Gerba CP, Gentry TJ, editors. Environmental Microbiology (Third Edition). San Diego: Academic Press. pp. 309 – 17
9. Labrador KL, Nacario MAG, Malajacan GT, Abello JJM, Galarion LH, Rensing C, Rivera WL (2019) Selecting rep-PCR markers to source track fecal contamination in Laguna Lake, Philippines. J Water Health 18(1):19–29. https://doi.org/10.2166/wh.2019.042
10. Lyautey E, Lu Z, Lapen DR, Berkers TE, Edge TA, Topp E (2010) Optimization and validation of rep-PCR genotypic libraries for microbial source tracking of environmental Escherichia coli isolates. 56(1):8–17. https://doi.org/10.1139/w09-113
11. Furukawa T, Suzuki Y (2013) A proposal for source tracking of fecal pollution in recreational waters by pulsed-field gel electrophoresis. Microbes Environ 28(4):444–449. https://doi.org/10.1264/jsme2.me13075
12. Furukawa T, Yoshida T, Suzuki Y (2011) Application of PFGE to source tracking of faecal pollution in coastal recreation area: a case study in Aoshima Beach. Japan 110(3):688–696. https://doi.org/10.1111/j.1365-2672.2010.04918.x
13. Prevost G, Jaulhac B, Piemont Y (1992) DNA fingerprinting by pulsed-field gel electrophoresis is more effective than ribotyping in distinguishing among methicillin-resistant Staphylococcus aureus isolates. J Clin Microbiol 30(4):967–973. https://doi.org/10.1128/jcm.30.4.967-973.1992
14. Paruch L, Paruch AM (2022) An Overview of Microbial Source Tracking Using Host-Specific Genetic Markers To Identify Origins of Fecal Contamination in different water environments. 14(11):1809
15. Zheng G, Shen ZJJFM, Safety H (2018) Host-specific genetic markers of fecal bacteria for fecal source tracking in food and water. 3(1):1–8
16. Mieszkin S, Furet J-P, Corthier G, Gourmelon M (2009) Estimation of Pig Fecal Contamination in a River Catchment by Real-Time PCR using two pig-specific Bacteroidales 16S rRNA genetic markers. Appl Environ Microbiol 75(10):3045–3054. https://doi.org/10.1128/AEM.02343-08
17. Green HC, Dick LK, Gilpin B, Samadpour M, Field KG (2012) Genetic markers for rapid PCR-based identification of gull, Canada goose, duck, and chicken fecal contamination in water. Appl Environ Microbiol 78(2):503–510. https://doi.org/10.1128/aem.05734-11
18. Knights D, Kuczynski J, Charlson ES, Zaneveld J, Mozer MC, Collman RG, Bushman FD, Knight R, Kelley ST (2011) Bayesian community-wide culture-independent microbial source tracking. Nat Methods 8(9):761–763. https://doi.org/10.1038/nmeth.1650
19. Staley C, Kaiser T, Lobos A, Ahmed W, Harwood VJ, Brown CM, Sadowsky MJ (2018) Application of SourceTracker for Accurate Identification of Fecal Pollution in recreational freshwater: a double-blinded study. Environ Sci Technol 52(7):4207–4217. https://doi.org/10.1021/acs.est.7b05401
20. Shenhav L, Thompson M, Joseph TA, Briscoe L, Furman O, Bogumil D, Mizrahi I, Pe'er I, Halperin E (2019) FEAST: fast expectation-maximization for microbial source tracking. Nat Methods 16(7):627–632. https://doi.org/10.1038/s41592-019-0431-x
21. An U, Shenhav L, Olson CA, Hsiao EY, Halperin E, Sankararaman S (2022) STENSL: Microbial Source Tracking with ENvironment SeLection. mSystems 7(5):e0099521. https://doi.org/10.1128/msystems.00995-21
22. Ministry of Environment Livestock manure treatment statistics https://www.me.go.kr/home/web/public_info/read.do;jsessionid=EwfO3skDzd5S7BtPml4yS52VcIUanbHEAnVZkEKqUgZyviWW6aVfl6HRY51HgImb.meweb2vhost_servlet_engine1?pagerOffset=30&maxPageItems=10&maxIndexPages=10&searchKey=&searchValue=&menuId=10357&orgCd=&condition.publicInfoMasterId=3&publicInfoId=88&menuId=10357. Accessed 1 January 2023
23. Water Environment Information System https://water.nier.go.kr/web. Accessed 1 January 2024
24. Illumina (2013) 16s metagenomic sequencing library preparation
25. Kozich JJ, Westcott SL, Baxter NT, Highlander SK, Schloss PD (2013) Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform. Appl Environ Microbiol 79(17):5112–5120. https://doi.org/10.1128/aem.01043-13
26. Ko G, Kim J-K, Jo S-W, Jeong D-Y, Unno T (2020) Effects of fermented coffee on human gut microbiota. J Appl Biol Chem 63(1):83–87
27. Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, Lesniewski RA, Oakley BB, Parks DH, Robinson CJ, Sahl JW et al (2009) Introducing mothur: Open-Source, Platform-Independent, community-supported Software for describing and comparing Microbial communities. 75(23):7537–7541. https://doi.org/10.1128/AEM.01541-09
28. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO (2013) The SILVA ribosomal RNA gene database project: improved data

processing and web-based tools. Nucleic Acids Res 41(Database issue):D590–D596. https://doi.org/10.1093/nar/gks1219

29. Rognes T, Flouri T, Nichols B, Quince C, Mahé F (2016) VSEARCH: a versatile open source tool for metagenomics. PeerJ 4:e2584. https://doi.org/10.7717/peerj.2584
30. Cole JR, Wang Q, Fish JA, Chai B, McGarrell DM, Sun Y, Brown CT, Porras-Alfaro A, Kuske CR, Tiedje JM (2014) Ribosomal database project: data and tools for high throughput rRNA analysis. Nucleic Acids Res 42(Database issue):D633–D642. https://doi.org/10.1093/nar/gkt1244
31. Westcott SL, Schloss PD (2017) OptiClust, an Improved Method for Assigning Amplicon-Based Sequence Data to Operational Taxonomic Units. mSphere 2(2). https://doi.org/10.1128/mSphereDirect.00073-17
32. Dixon PJJVS (2003) VEGAN, a package of R functions for community ecology. 14(6):927 – 30
33. Kandlikar G, Gold Z, Cowen M, Meyer R, Freise A, Kraft N, Moberg-Parker J, Sprague J, Kushner D, Curd E (2018) Ranacapa: an R package and Shiny web app to explore environmental DNA data with exploratory statistics and interactive visualizations [version 1; peer review: 1 approved, 2 approved with reservations]. 7(1734). https://doi.org/10.12688/f1000research.16680.1
34. Breiman L (2001) Random forests. Mach Learn 45(1):5–32. https://doi.org/10.1023/A:1010933404324
35. Chen T, Guestrin C (eds) (2016) Xgboost: A scalable tree boosting system. Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining
36. Cortes C, Vapnik V (1995) Support-vector networks. Mach Learn 20(3):273–297. https://doi.org/10.1007/BF00994018
37. Altman NS (1992) An introduction to Kernel and Nearest-Neighbor Nonparametric Regression. Am Stat 46(3):175–185. https://doi.org/10.1080/00031305.1992.10475879
38. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg VJJL (2011) Scikit-learn: machine learning in Python. 12:2825–2830
39. Liang H, Yu Z, Wang B, Ndayisenga F, Liu R, Zhang H, Wu G (2021) Synergistic application of molecular markers and community-based Microbial Source Tracking methods for Identification of Fecal Pollution in River Water during Dry and Wet Seasons. Front Microbiol 12. https://doi.org/10.3389/fmicb.2021.660368
40. Tamai S, Suzuki Y (2023) Diversity of Fecal Indicator Enterococci among different hosts: importance to Water Contamination Source Tracking. Microorganisms 11(12):2981
41. Chungcheongbuk-do Institute of Health and Environment (2008) A Research on Groundwater of Abandoned Tubular Wells in Chungcheongbuk-do Province
42. Briscoe L, Halperin E, Garud NR (2023) SNV-FEAST: microbial source tracking with single nucleotide variants. Genome Biol 24(1):101. https://doi.org/10.1186/s13059-023-02927-8
43. Gupta VK, Paul S, Dutta C (2017) Geography, ethnicity or subsistence-specific variations in Human Microbiome Composition and Diversity. Front Microbiol 8. https://doi.org/10.3389/fmicb.2017.01162
44. Henderson G, Cox F, Ganesh S, Jonker A, Young W, Abecia L, Angarita E, Aravena P, Nora Arenas G, Ariza C, Attwood GT et al (2015) Rumen microbial community composition varies with diet and host, but a core microbiome is found across a wide geographical range. Sci Rep 5(1):14567. https://doi.org/10.1038/srep14567

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.