Novel *Escherichia coli* genetic markers for tracking domestic wastewater contamination in environmental waters

Ryota Gomi\textsuperscript{a,*}, Eiji Haramoto\textsuperscript{b}, Hiroyuki Wada\textsuperscript{c}, Yoshinori Sugie\textsuperscript{c}, Chih-Yu Ma\textsuperscript{c}, Sunayana Raya\textsuperscript{d}, Bikash Malla\textsuperscript{b}, Fumitake Nishimura\textsuperscript{c}, Hiroaki Tanaka\textsuperscript{c}, Masaru Ihara\textsuperscript{c,e,*}

\textsuperscript{a}Department of Environmental Engineering, Graduate School of Engineering, Kyoto University, Katsura, Nishikyo-ku, 615-8540, Kyoto, Japan

\textsuperscript{b}Interdisciplinary Center for River Basin Environment, University of Yamanashi, 4-3-11 Takeda, Kofu, 400-8511, Yamanashi, Japan

\textsuperscript{c}Research Center for Environmental Quality Management, Graduate School of Engineering, Kyoto University, 1-2 Yumihama, Otsu, 520-0811, Shiga, Japan

\textsuperscript{d}Department of Engineering, University of Yamanashi, Kofu, 400-8511, Yamanashi, Japan

\textsuperscript{e}Faculty of Agriculture and Marine Science, Kochi University, Nankoku, 783-8502, Kochi, Japan

\textsuperscript{*Corresponding authors

E-mail address: gomi.ryota.34v@kyoto-u.jp; Tel: +81-75-383-3354; Fax: +81-75-383-3358

E-mail address: ihara.masaru@kochi-u.ac.jp; Tel: +81-88-864-5163
**ABSTRACT**

*Escherichia coli* has been used as an indicator of fecal pollution in environmental waters. However, its presence in environmental waters does not provide information on the source of water pollution. Identifying the source of water pollution is paramount to be able to effectively prevent contamination. The present study aimed to identify *E. coli* microbial source tracking (MST) markers that can be used to identify domestic wastewater contamination in environmental waters. We first analyzed wastewater *E. coli* genomes sequenced in this study (n = 50) and RefSeq animal *E. coli* genomes of fecal origin (n = 82), and identified 144 candidate wastewater-specific marker genes. The sensitivity and specificity of the candidate marker genes were then assessed by screening the genes in 335 RefSeq wastewater *E. coli* genomes and 3,318 RefSeq animal *E. coli* genomes. We finally identified two MST markers, namely W_nqrC and W_clsA_2, which could be used for tracking wastewater contamination. These two markers showed higher performance than the previously developed human wastewater-associated *E. coli* markers H8 and H12. PCR assays to detect W_nqrC and W_clsA_2 were also developed and validated. The developed PCR assays are potentially useful for detecting *E. coli* isolates of wastewater origin in environmental waters. However, further studies are needed to assess the applicability of the developed markers to a culture-independent approach.
1. **Introduction**

Surface waters contaminated by wastewater can increase human health risks because untreated and insufficiently-treated wastewater contains various pathogens (Castro-Hermida et al., 2008; Naidoo and Olaniran, 2013; Zhi et al., 2020). Humans can be exposed to such pathogens through activities such as swimming and the consumption of foods irrigated with poor-quality water (Boehm et al., 2018; Steele and Odumeru, 2004). Because it is impractical to test for all possible pathogens in surface waters, fecal indicator bacteria (FIB), such as total coliforms, fecal coliforms, and *Escherichia coli*, are monitored as proxies for pathogens (Field and Samadpour, 2007; Naidoo and Olaniran, 2013). However, the presence of FIB does not necessarily indicate that the water is contaminated by human pathogens for several reasons (e.g., the presence of naturalized FIB (Ishii et al., 2006)). The major limitation of FIB is that, due to their cosmopolitan nature, their presence does not provide information on the possible origin of contamination (Warish et al., 2015). FIB can originate from many sources, such as community sewage (containing human-specific pathogens and zoonotic pathogens) and diffuse pollution from animal farms and wildlife (containing zoonotic pathogens). It is known that human health risks can vary according to the source of fecal pollution, and human health risks from recreational waters impacted by human sources are generally considered to be higher than those impacted by non-human sources (Soller et al., 2010). Identifying the origin of water
pollution is paramount in order to assess the associated health risks and determine the actions needed to prevent contamination (Scott et al., 2002).

Microbial source tracking (MST) is an approach used to determine the sources of fecal contamination in environmental waters (Harwood et al., 2014). MST methods can be divided into culture-dependent methods and culture-independent methods. Culture-dependent methods require the growing of environmental isolates from water, while in culture-independent methods genetic markers are assayed directly from a water sample or from DNA extracted from a water sample (Field and Samadpour, 2007). Markers such as the 16S rRNA gene of Bacteroides and pepper mild mottle virus (PMMoV) have been used to track the sources of fecal contamination in MST in a culture-independent manner (Gonzalez-Fernandez et al., 2021; Mathai et al., 2020). However, routine water quality assessment has focused primarily on traditional FIB, and alternative indicators such as Bacteroides are not usually monitored for this purpose (Senkbeil et al., 2019; Teixeira et al., 2020). Because E. coli is still a widely-used FIB, MST markers targeting E. coli would be useful if included in the course of routine monitoring. Even a culture-dependent approach would be useful for E. coli, because users do not have to collect isolates exclusively for MST purposes but can use E. coli colonies obtained in routine water quality assessments.

Although little success has been achieved in developing MST markers targeting E. coli, we have previously identified a set of human wastewater, cattle, pig, and chicken feces-associated
*E. coli* MST markers (Gomi et al., 2014). The following studies reported two of the proposed markers, namely H8 and H12, showed relatively high sensitivity and specificity and could be useful as human wastewater-associated markers (Nopprapun et al., 2020; Senkbeil et al., 2019; Warish et al., 2015). However, the original genetic markers were developed by analyzing a small number of *E. coli* genomes (n = 22). Since then, a growing number of *E. coli* genomes have become available in public databases. These *E. coli* genomes include those from various isolation sources and geographically diverse hosts, which might allow the development of reliable MST markers that can be used globally. Here, we analyzed *E. coli* genomes in the database of the National Center for Biotechnology Information (NCBI) as well as genomes sequenced in the present study to identify more reliable *E. coli* genetic markers for tracking domestic wastewater contamination in environmental waters. Our aim in the present study was to identify culture-dependent MST markers, but applicability to a culture-independent approach was also discussed. The sensitivity and specificity of the previously developed human wastewater-associated markers H8 and H12 were also re-evaluated.

2. Materials and methods

2.1. *E. coli* isolation from wastewater and genome sequencing.
The flowchart of the study design is shown in Figure 1. We obtained wastewater *E. coli* isolates from municipal wastewater treatment plants (WWTPs) in Shiga Prefecture, Japan. Briefly, biologically treated wastewater samples before/after chlorination were collected from three different WWTPs (WWTP A, WWTP B, and WWTP C) between December 2016 and December 2017. The chlorinated samples were treated with sodium thiosulfate immediately after collection to neutralize the chlorine. Samples were collected using sterile sampling bags or sterile sampling bottles, stored at 4 °C, and processed within 24 hours. *E. coli* were cultured using the pour plate method with XM-G agar (Nissui, Tokyo, Japan) at 37 °C for 18h, and colonies were randomly picked from the plates. Colonies suspected to be mixed were subcultured with fresh XM-G agar plates to obtain pure isolates. A total of 50 isolates (29 isolates from WWTP A, nine isolates from WWTP B, and 12 isolates from WWTP C) were obtained for genome sequencing (see Table S1 for information on the collected *E. coli*).

DNA was extracted from each isolate using a DNeasy blood and tissue kit (Qiagen, Hilden, Germany) following the manufacturer’s instructions, and the concentration of extracted DNA was measured using a Qubit fluorometer (Thermo Fisher Scientific, Waltham, MA, United States). Extracted DNA was stored at -30°C for up to a week until sample preparation for genome sequencing. The extracted DNA was processed using a Nextera XT DNA sample preparation kit (Illumina, San Diego, CA) and sequenced on an Illumina MiSeq instrument for 600 cycles (300-bp paired-end sequencing).
Figure 1. Overview of the strategy to identify and test the wastewater-specific *E. coli* genetic markers. RefSeq *E. coli* genomes used for blastn analysis (335 wastewater *E. coli* genomes and 3,318 animal *E. coli* genomes) and *in silico* PCR analysis (91 wastewater *E. coli* genomes, 824 animal *E. coli* genomes, and 303 environmental *E. coli* genomes) included both complete and
draft genome assemblies. There was no overlap between the genomes used for blastn analysis and in silico PCR analysis.

2.2. Retrieving complete E. coli genome sequences from the NCBI database.

The Reference Sequence (RefSeq) E. coli genomes with an assembly level of “complete genome” (n = 1,278) were downloaded in May 2021. Metadata information, such as strain name, host, and isolation source, was extracted from the GenBank files using in-house Python scripts. E. coli genomes of non-primate animals and of fecal/intestinal origin (n = 82) were kept for further analysis (E. coli genomes from birds were also included here. See Table S2 for information on these animal E. coli genomes).

2.3. Genome assembly and annotation.

Raw sequence reads from wastewater E. coli were trimmed using fastp (v0.20.0) (Chen et al., 2018), and the trimmed reads were assembled using Unicycler (v0.4.8) with the --no_correct option (Wick et al., 2017). For the complete RefSeq genomes, we generated simulated paired-end reads using wgsim (v1.12) (https://github.com/lh3/wgsim). Simulated reads were trimmed using fastp and assembled using Unicycler with the same parameters as those used for assembling the wastewater E. coli genomes. We did not use the complete RefSeq genomes but used these draft assemblies in the following analysis. This was to remove any biases introduced
by differences in assembly status (i.e., complete assemblies vs. draft assemblies) in the pan-
genome-wide association study (pan-GWAS). Assembled genomes were annotated using Prokka (v1.14.6) (Seemann, 2014).

2.4. Identification of candidate wastewater-specific marker genes using a pan-GWAS approach.

The pan-genome of 132 *E. coli* genomes (50 wastewater *E. coli* genomes + 82 animal *E. coli* genomes) was constructed by running Roary (v3.13.0) on GFF3 files generated by Prokka (Page et al., 2015). The -s option was used to turn off paralog splitting. The pan-GWAS approach implemented by Scoary was used to find genes that were enriched in wastewater *E. coli* genomes (Brynildsrud et al., 2016). Scoary (v 1.6.16) was run on Roary’s gene presence/absence file. The --collapse option was used to collapse genes that had identical distribution patterns among the genomes. Because our aim was to identify genes that were enriched in wastewater *E. coli* genomes and not to infer causal association, the --no_pairwise option was used as recommended by the developers (https://github.com/AdmiralenOla/Scoary).

Genes with Benjamini-Hochberg adjusted p-values of lower than 0.05 and with a specificity exceeding 95% were defined as candidate wastewater-specific marker genes. We set the specificity threshold so as to reduce the number of false positives, considering the future application to DNA directly extracted from environmental samples (this type of DNA can
contain DNA from multiple *E. coli* isolates and thus is sensitive to false positives). When a set
of collapsed genes met the p-value and specificity criteria, one gene was randomly selected as
a candidate marker.

2.5. Retrieving genome sequences from the NCBI database for sensitivity and specificity
analysis.

The RefSeq *E. coli* genomes (n = 20,237), including both complete and draft assemblies, were
downloaded in June 2021. Metadata information was extracted from the GenBank files using
in-house Python scripts. Genomes of wastewater origin (n = 335) and genomes of non-primate
animals and of fecal/intestinal origin (n = 3,318) were kept for sensitivity and specificity
analysis (see Table S3 for information on the wastewater *E. coli* genomes and Table S4 for
information on the animal *E. coli* genomes).

2.6. Evaluating the sensitivity and specificity of candidate marker genes.

The candidate marker genes were detected in 3,653 genome assemblies (335 RefSeq
wastewater *E. coli* genomes + 3,318 RefSeq animal *E. coli* genomes) using the blastn program
with the default parameters. Thresholds of a minimum percent identity of 80% and a minimum
DNA coverage of 50% were used to define the presence of a gene. For each candidate marker
gene, the following metrics were calculated: true positive (TP) (the number of wastewater *E.
coli genomes positive for the candidate marker), true negative (TN) (the number of animal E. coli genomes negative for the candidate marker), false positive (FP) (the number of animal E. coli genomes positive for the candidate marker), and false negative (FN) (the number of wastewater E. coli genomes negative for the candidate marker). Sensitivity [TP/(TP + FN)] and specificity [TN/(TN + FP)] were then calculated for each candidate marker gene (note that sensitivity and specificity were reported as %). The above blastn analysis and calculation were also performed for H8 and H12. Note that H8 corresponded to one of the candidate marker genes.

2.7. Development of PCR assays.

Primers that yield desired product sizes were designed for the identified marker genes as shown in Table 1. For each target gene, the validity of the primers was checked with Primer 3 software using the default settings (Untergasser et al., 2012). Moreover, using Primer-BLAST with the default settings, we confirmed that an amplicon of the desired size was produced from each wastewater genome with the target gene that had been sequenced in 2.1 (Ye et al., 2012).

To further validate the developed PCR primers, we performed PCR on E. coli isolates collected from the following sources: wastewater influent (n = 78), wastewater effluent (after biological treatment and chlorination) (n = 49), deer (n = 30), and cattle (n = 30). Wastewater influent isolates (n = 48) and effluent isolates (n = 49) were newly collected in 2021 from WWTP A in
Shiga Prefecture, Japan. Other wastewater influent isolates (n = 30) were collected in 2018 from a WWTP (WWTP D) in Yamanashi Prefecture, Japan. We used the same sampling protocol and *E. coli* isolation procedure as described in 2.1, except that we used Chromocult coliform agar (Merck, Darmstadt, Germany) or CHROMagar ECC (CHROMagar, Paris, France) and incubated the plates for ~24 hours at 37 °C to isolate *E. coli* from the influent samples collected from WWTP D. Animal *E. coli* isolates were obtained from animal feces. Briefly, eight deer fecal samples were collected in Yamanashi Prefecture, and two deer fecal samples and 10 cattle fecal samples were collected in Shiga Prefecture. Deer fecal samples were collected using sterile sampling bags, stored at 4 °C, and processed within 24-48 hours. The deer fecal samples collected in Yamanashi Prefecture were stored at 4 °C and transported to the laboratory in Shiga Prefecture, where the samples were processed. Cattle fecal samples were collected using sterile centrifuge tubes, stored at 4 °C, and processed within 24 hours. Animal feces were suspended and diluted with PBS and processed using the pour plate method with XM-G agar, except for two deer fecal samples for which CHROMagar ECC was used. The plates were incubated for 18 hours at 37 °C. Sub-culturing was performed to obtain pure colonies. Up to three isolates were obtained from a single individual (Dombek et al., 2000). DNA was extracted from *E. coli* obtained from WWTP A and animal feces using a DNeasy blood and tissue kit. The concentration of extracted DNA was measured using a Qubit fluorometer. The extracted DNA was stored at -30°C for up to one year until the PCR
experiments. For *E. coli* obtained from WWTP D, we used cell suspensions as the inputs in the PCR experiments. Cell suspensions were prepared by either of the following two methods: (i) a single *E. coli* colony grown on the XM-G plate was picked up using a clean tip, suspended in 500 μL of ultrapure water, and then vortexed; or (ii) a glycerol stock (*E. coli* stored in 25% glycerol) was diluted 10 times using ultrapure water. PCR reactions were performed on the extracted DNA or cell suspension using QuantiFast SYBR Green PCR Kit (Qiagen, Hilden, Germany) with PCR conditions as described below. PCR mixture (25 μL) was composed of 12.5 μL of 2× QuantiFast SYBR Green PCR Master Mix, 2.5 μL each of forward and reverse primers (5 μM), 5 μL of template DNA or cell suspension, and 2.5 μL of PCR grade water or ultrapure water. The reactions were carried out by incubation at 95 °C for 5 min, followed by 25 to 30 cycles (depending on the template concentrations) consisting of 95 °C (denaturation) for 10 s and 60 °C (combined annealing/extension) for 30 s. A melting curve analysis was performed after the amplification. In addition to the primers for W_nqrC and W_clsA_2, we also used previously published primers for amplifying H8, H12, and also *uidA* (primers UAL and UAR): we included *uidA* primers to confirm the *E. coli* identification (Gomi et al., 2014; Maheux et al., 2009). All PCR reactions were performed on a Thermal Cycler Dice Real Time System III (TaKaRa Bio, Inc.). Previously sequenced *E. coli* isolates, namely KOr026 (positive: W_clsA_2 and H8, negative: W_nqrC and H12, BioSample accession: SAMD00053091) and KMi027 (positive: W_nqrC and H12, negative: W_clsA_2 and H8,
BioSample accession: SAMD00053165), were used as positive/negative controls (Gomi et al., 2017b). We also included no template controls. The correct PCR products were confirmed by comparison with the melting curves of the positive controls, which showed peaks at melting temperatures of 82.5 °C, 89.0 °C, 91.6 °C, 85.7 °C, and 84.4 °C for W_nqrC, W_clsA_2, H8, H12, and uidA, respectively. We performed replicate reactions for a subset of samples (50 samples, including 42 for extracted DNA and 8 for cell suspensions) and confirmed the results to be consistent between all technical replicates.

We also performed in silico PCR using the developed primers on RefSeq E. coli genomes that were made public after June 2021 (i.e., genomes that were not used for blastn analysis). This is to test the developed PCR primers using E. coli genomes of diverse origins and from various geographical locations. In silico PCR was previously shown to be highly concordant with in vitro PCR when used for gene detection in E. coli (Beghain et al., 2018; Lindsey et al., 2017).

Briefly, RefSeq E. coli genomes, including both complete and draft assemblies, were downloaded in August 2022. Among the downloaded genomes, those used for blastn analysis were removed, leaving 6,399 genomes. Metadata information was extracted from the GenBank files, and the genomes of wastewater origin (n = 91) and those of non-primate animals and of fecal/intestinal origin (n = 824) were retrieved. In silico PCR was performed on the Primer-BLAST website using primer sequences of W_nqrC, W_clsA_2, H8, and H12 with the default parameters, except that we used 500 bp as max target amplicon size (Ye et al., 2012).
Table 1. Primers designed in this study.

| Target   | Sequence (5′-3′)                      | Product size (bp) |
|----------|--------------------------------------|-------------------|
| W_nqrC   | F: CATGCGTTTCTTGCTCTGGG              | 105               |
|          | R: GTTGGCGTCTTTCCACCCTAG             |                   |
| W_clsA_2 | F: ACCGATGATTGGTGCGGATACT            | 165               |
|          | R: GCTAAACGCGCCCATGAAG               |                   |

2.8. Detection of MST markers in *E. coli* isolates from environmental waters.

*E. coli* isolates were obtained in January 2022 from a river in Kyoto Prefecture in Japan. The sampling points in the river were downstream of two wastewater treatment plants and known to be heavily affected by treated wastewater (Figure S1). A total of eight water samples were collected from the river using sterile sampling bags, stored at 4 °C, and processed within 24 hours. Ten *E. coli* colonies were obtained from each sample using the membrane filter method with XM-G agar plates. The plates were incubated for 18 hours at 37 °C, and colonies were sub-cultured with fresh XM-G agar plates to obtain pure isolates. PCR reactions were performed as described above to test for the presence of W_nqrC and W_clsA_2 in the obtained *E. coli* isolates.

We also performed *in silico* PCR on RefSeq *E. coli* genomes of environmental water origin. Environmental *E. coli* genomes were extracted from the RefSeq *E. coli* genomes downloaded in August 2022 based on the metadata information. *In silico* PCR was performed using the retrieved genomes (n = 303) on the Primer-BLAST website as described above.
2.9. **Sequence Data Accession Number(s).**

The raw sequencing data obtained in the present study were deposited in the NCBI SRA under BioProject accession number PRJNA765288.

3. **Results and discussion**

3.1. **Basic characteristics of wastewater *E. coli* genomes sequenced in this study.**

The 50 wastewater *E. coli* genomes sequenced in this study were highly diverse, comprising 44 different sequence types (STs). These draft assemblies had a median of 140 contigs (range 49 contigs to 346 contigs), a median genome size of 4,875,061 bp (range 4,576,995 bp to 5,375,884 bp), and a median N50 of 179,046 bp (range 63,435 bp to 737,274 bp) after removing contigs shorter than 100 bp (Table S1). All 50 assemblies met the quality control criteria of (i) contig numbers: $\leq 800$; (ii) genome size: 3.7 Mbp to 6.4 Mbp; and (iii) N50: $>20$ kb, as recommended by EnteroBase (Zhou et al., 2020).

3.2. **Identification of candidate wastewater-specific marker genes.**

A pan-genome was constructed with Roary using 50 wastewater *E. coli* genomes sequenced in this study and 82 RefSeq animal *E. coli* genomes. Roary identified a total of 25,343 genes,
including 2,925 core genes (genes shared among ≥ 99% of strains) and 22,418 accessory genes (genes observed among < 99% of strains). The pan-GWAS approach was employed to identify candidate wastewater-specific genes among the accessory genes. A total of 144 genes met the criteria of Benjamini-Hochberg adjusted p-values of less than 0.05 and specificity higher than 95%, and thus were defined as candidate marker genes (see Table S5 for details of the identified genes). These 144 genes were subjected to sensitivity and specificity analysis by blastn using a larger number of RefSeq E. coli genomes.

3.3. Sensitivity and specificity of candidate marker genes determined by blastn analysis.

The sensitivity and specificity of 144 candidate marker genes were assessed by screening these genes in 335 RefSeq wastewater E. coli genomes and 3,318 RefSeq animal E. coli genomes. Figure 2 shows the distribution of sensitivity and specificity of each candidate marker gene. Most candidate markers achieved high specificity, which is congruent with the fact that we used a threshold of 95% specificity for identifying candidate markers in the Scoary analysis. However, there were no genes that simultaneously showed high sensitivity and high specificity, which may be partially due to the presence of E. coli isolates that can colonize multiple host species (Johnson and Clabots, 2006; Stenske et al., 2009). In fact, among the genes showing >70% specificity, no genes showed >40% sensitivity. It should be noted that there were genes with almost 100% sensitivity and 0% specificity (i.e., detected in almost all the tested isolates).
even though we selected genes with specificity higher than 95% in the Scoary analysis. This can be because we used blastp with a minimum percent identity of 95% for clustering sequences in the Roary analysis but we used blastn with a minimum percent identity of 80% for gene screening. In other words, genes assigned to different clusters by Roary can share >80% nucleotide sequence identity and thus can be detected in the same blastn analysis.

Because the species composition of the hosts of RefSeq animal *E. coli* genomes was biased (some host species were more prevalent than others), the specificity of candidate markers was separately calculated for different host groups, namely, “Cattle”, “Swine”, “Chicken”, “Wild boar (*Sus scrofa*)”, “Canine”, “Sheep and goat”, and “Others” (**Figure S2**). We also considered the specificity calculated for these host groups when selecting MST markers from the candidates.
Figure 2. Sensitivity and specificity of candidate wastewater-specific marker genes as determined by blastn analysis. The distribution of sensitivity and specificity of 144 candidate marker genes and previously identified marker genes, H8 and H12, are plotted (H8 in fact corresponded to one of the 144 candidate marker genes). The positions of markers mentioned in the main text (W_scrK_2, W_cra_2, and W_g_3442) are also indicated.

3.4. Selection of W_nqrC and W_clsA_2.
As noted in the Materials and Methods section, selecting a marker with high specificity is required to reduce the number of false positives, especially when the marker is applied to DNA directly extracted from environmental samples. We therefore selected W_nqrC from the 144 candidates, as this marker showed high specificity (99.1%) and relatively high sensitivity (16.4%) (Figure 2). Moreover, W_nqrC showed the highest sensitivity among the candidate markers which showed >95% specificity for all host groups (Figure S2). We then tried to select another marker that was prevalent among the RefSeq wastewater E. coli genomes without W_nqrC. This was to increase the sensitivity when both markers are used in combination.

Among the candidate markers showing >99% specificity, W_cra_2, W_scrK_2, W_clsA_2, and W_g_3442 showed different marker distribution patterns (the red bars in Figure S3) from W_nqrC and also showed relatively high sensitivities (Figure S3). Although W_cra_2 and W_scrK_2 showed high specificity (>99.7%) and relatively high sensitivity (12.2% and 11.0%, respectively), similar sequences with 60%–75% nucleotide sequence identity were prevalent among animal E. coli isolates. We therefore selected W_clsA_2 as the second marker because there were no similar sequences prevalent among animal E. coli, and this marker showed higher specificity than W_g_3442 (99.8% vs. 99.4%). Moreover, the specificity of W_clsA_2 was 100% for all host groups except “Others” (only six E. coli genomes from silver gulls were positive for W_clsA_2) (Figure S2). The sequences of W_nqrC and W_clsA_2 are shown in Figure S4.
Prokka annotation indicated \textit{W\textsubscript{nqrC}} encodes \textit{Na\textsuperscript{+}}-translocating NADH-quinone reductase subunit C, and \textit{W\textsubscript{clsA\_2}} encodes major cardiolipin synthase ClsA. The BLASTN searches against the nucleotide collection (nr/nt) database revealed that \textit{W\textsubscript{nqrC}} is mostly present on plasmids, while \textit{W\textsubscript{clsA\_2}} is equally present on chromosomes and plasmids. Among the RefSeq wastewater \textit{E. coli} genomes positive for \textit{W\textsubscript{nqrC}} (\(n = 55\)), almost half of the genomes (\(n = 27\)) were assigned to ST131, which was previously reported to be present among wastewater \textit{E. coli} (Dolejska et al., 2011; Gomi et al., 2017a; Sghaier et al., 2019) and also is known for its association with extraintestinal infections (Nicolas-Chanoine et al., 2014; Riley, 2014; Stoesser et al., 2016). The remaining \textit{W\textsubscript{nqrC}}-positive genomes were assigned to various STs, including ST69 (\(n = 4\)), ST95 (\(n = 4\)), and ST10 (\(n = 3\)). The \textit{W\textsubscript{nqrC}}-positive wastewater \textit{E. coli} genomes sequenced in this study (\(n = 9\)) belonged to diverse STs, including ST73 (\(n = 2\)), ST95 (\(n = 2\)), and ST357 (\(n = 2\)) (Table S6). Among the RefSeq wastewater \textit{E. coli} genomes positive for \textit{W\textsubscript{clsA\_2}} (\(n = 31\)), more than half of the genomes (\(n = 17\)) were assigned to ST635, which was previously reported to be prevalent among wastewater \textit{E. coli} isolates in Canada and hospital sink \textit{E. coli} isolates in the UK (Constantinides et al., 2020; Zhi et al., 2019). The remaining \textit{W\textsubscript{clsA\_2}}-positive genomes belonged to various STs, including ST399 (\(n = 3\)), ST401 (\(n = 2\)), ST607 (\(n = 2\)), and ST3168 (\(n = 2\)). The \textit{W\textsubscript{clsA\_2}}-positive wastewater \textit{E. coli} genomes sequenced in this study (\(n = 5\)) belonged to five different STs (ST399, ST472, ST607, ST635, and ST5295) (Table S7).
We note that W_nqrC/W_clsA_2-positive RefSeq wastewater *E. coli* genomes originated from various countries, including the Czech Republic, Germany, Canada, the USA, and South Africa (Table S6, Table S7). In fact, 43 (78%) of W_nqrC-positive RefSeq wastewater *E. coli* genomes (n = 55) and 31 (100%) of W_clsA_2-positive RefSeq wastewater *E. coli* genomes (n = 31) originated from countries other than Japan, which indicates that these markers could be used globally.

### 3.5. Sensitivity and specificity of H8 and H12.

We also assessed the sensitivity and specificity of the two previously developed markers, H8 and H12, which were calculated based on the results of blastn analysis (Figure 2). Notably, H8 corresponded to one of the 144 candidate marker genes identified by Scoary. Figure 2 shows that both H8 and H12 achieved high specificity (97.9% and 99.8%, respectively), which is consistent with previous studies (Senkbeil et al., 2019; Warish et al., 2015). However, the markers proposed in the present study showed better performances. W_nqrC showed higher specificity and sensitivity than H8 (99.1% vs. 97.9% and 16.4% vs. 16.1%, respectively), and W_clsA_2 showed almost the same specificity (~99.8%) but higher sensitivity than H12 (9.3% vs. 8.4%). Moreover, when used in combination, a combination of W_nqrC and W_clsA_2 showed higher specificity and sensitivity than a combination of H8 and H12 (98.9% vs. 97.8% and 25.7% vs. 24.2%, respectively). These results suggest that although H8 and H12 could be
useful for tracking wastewater contamination, \textit{W\textsubscript{nqrC}} and \textit{W\textsubscript{clsA\_2}} may yield better performances.

3.6. Development of PCR assays targeting \textit{W\textsubscript{nqrC}} and \textit{W\textsubscript{clsA\_2}}.

We designed PCR primers for detection of \textit{W\textsubscript{nqrC}} and \textit{W\textsubscript{clsA\_2}}. PCR assays on positive controls yielded the expected DNA bands for both \textit{W\textsubscript{nqrC}} (105 bp) and \textit{W\textsubscript{clsA\_2}} (165 bp), whereas no bands were observed for the negative controls (Figure S5). The developed PCR assays were applied to detect the MST markers in \textit{E. coli} isolates obtained from wastewater and animal feces (deer and cattle feces) (Table 2). \textit{W\textsubscript{nqrC}} or \textit{W\textsubscript{clsA\_2}} was detected in 21 (26.9\%) and 9 (18.4\%) of wastewater influent and effluent isolates, respectively (Table 2). Importantly, there was no overlap between the isolates positive for \textit{W\textsubscript{nqrC}} and \textit{W\textsubscript{clsA\_2}}, which suggests that the use of \textit{W\textsubscript{nqrC}} and \textit{W\textsubscript{clsA\_2}} in combination could increase the sensitivity. On the other hand, no isolates from animal feces were positive for \textit{W\textsubscript{nqrC}} or \textit{W\textsubscript{clsA\_2}}. We also performed PCR to detect H8 and H12 using the same set of \textit{E. coli} isolates. H8 or H12 was detected in 11 (14.1\%) and 7 (14.3\%) of wastewater influent and effluent isolates, respectively (Table 2). This indicates that a combination of \textit{W\textsubscript{nqrC}} and \textit{W\textsubscript{clsA\_2}} is more sensitive for detecting wastewater \textit{E. coli} than a combination of H8 and H12 and supports the findings of the above blastn analysis. As with \textit{W\textsubscript{nqrC}} and \textit{W\textsubscript{clsA\_2}}, no isolates from animal feces were positive for H8 or H12.
We also performed *in silico* PCR analysis using RefSeq *E. coli* genomes that were not used for blastn analysis. Among the 91 wastewater *E. coli* genomes, 29 genomes (31.9%) were positive for W_nqrC or W_clsA_2 (*Table 2, Table S8*). On the other hand, only 16 genomes (1.9%) among the 824 animal *E. coli* genomes were positive for W_nqrC or W_clsA_2 (*Table 2, Table S9*). These results highlight the high specificity of the developed MST markers. *In silico* PCR analysis of H8 and H12 revealed 25 wastewater *E. coli* genomes (27.5%) to be positive for H8 or H12, while 27 animal *E. coli* genomes (3.3%) were also positive for H8 or H12 (*Table 2, Table S8, Table S9*). This again highlights that a combination of W_nqrC and W_clsA_2 can yield better performance than a combination of H8 and H12.

**Table 2.** Prevalence of W_nqrC, W_clsA_2, H8, and H12 in *E. coli* determined by (*in silico*) PCR.

| Source                  | No. of *E. coli* isolates/ genomes tested | No. of *E. coli* isolates/ genomes positive for W_nqrC (%) | No. of *E. coli* isolates/ genomes positive for W_clsA_2 (%) | No. of *E. coli* isolates/ genomes positive for H8 (%) | No. of *E. coli* isolates/ genomes positive for H12 (%) | No. of *E. coli* isolates/ genomes positive for W_nqrC or W_clsA_2 (%) |
|-------------------------|------------------------------------------|-------------------------------------------------------------|-------------------------------------------------------------|--------------------------------------------------------|--------------------------------------------------------|------------------------------------------------------------------|
| Wastewater influent     | 78                                       | 15 (19)                                                     | 6 (8)                                                       | 8 (10)                                                 | 5 (6)                                                  | 21 (27)                                                          | 11 (14)*                                                       |
| Sample Type          | N  | 5%  | 4%  | 6%  | 1%  | 9%  | 7%  |
|----------------------|----|-----|-----|-----|-----|-----|-----|
| Wastewater effluent  | 49 | 5 (10)| 4 (8) | 6 (12) | 1 (2) | 9 (18) | 7 (14) |
| Deer                | 30 | 0 (0) | 0 (0) | 0 (0) | 0 (0) | 0 (0) | 0 (0) |
| Cattle              | 30 | 0 (0) | 0 (0) | 0 (0) | 0 (0) | 0 (0) | 0 (0) |
| River water         | 80 | 8 (10) | 6 (8) | 7 (4) | 0 (0) | 14 (18) | 7 (4) |

**In silico PCR**

| Sample Type          | N  | 5%  | 4%  | 6%  | 1%  | 9%  | 7%  |
|----------------------|----|-----|-----|-----|-----|-----|-----|
| Wastewater effluent  | 91 | 26 (29) | 3 (3) | 5 (5) | 20 (22) | 29 (32) | 25 (27) |
| Pig                  | 237 | 3 (1) | 0 (0) | 11 (5) | 0 (0) | 3 (1) | 11 (5) |
| (swine, Sus scrofa domestic a) |
| Bird                 | 199 | 4 (2) | 0 (0) | 3 (2) | 0 (0) | 4 (2) | 3 (2) |
| (other than chicken) |
| Cattle               | 178 | 6 (3) | 1 (1) | 7 (4) | 0 (0) | 7 (4) | 7 (4) |
| (bovine, cow, Bos taurus) |
| Chicken              | 100 | 0 (0) | 0 (0) | 0 (0) | 0 (0) | 0 (0) | 0 (0) |
| (Gallus gallus)      |
| Goat/Sheep           | 19  | 0 (0) | 0 (0) | 2 (11) | 0 (0) | 0 (0) | 2 (11) |
| Dog                  | 12  | 2 (17) | 0 (0) | 0 (0) | 2 (17) | 2 (17) | 2 (17) |
| (Canis lupus familiaris) |
| Red deer             | 12  | 0 (0) | 0 (0) | 0 (0) | 0 (0) | 0 (0) | 0 (0) |
| Wild boar            | 11  | 0 (0) | 0 (0) | 0 (0) | 0 (0) | 0 (0) | 0 (0) |
| (Sus scrofa) | Other animals<sup>d</sup> | Environmental waters |
|-------------|--------------------------|----------------------|
|             | 56  | 0 (0) | 0 (0) | 2 (4) | 0 (0) | 0 (0) | 2 (4) |
|             | 303 | 21 (7) | 1 (0.3) | 12 (4) | 8 (3) | 22 (7) | 20 (7) |

<sup>a</sup>Two isolates carried both H8 and H12.

<sup>b</sup>We could not determine the prevalence of H8 and H12 due to the loss of glycerol stocks.

<sup>c</sup>Influent and effluent were not distinguished for many of the RefSeq wastewater *E. coli* genomes.

<sup>d</sup>Animal hosts with fewer than ten *E. coli* genomes were combined into this category.

### 3.7. Prevalence of MST markers in *E. coli* isolates from environmental waters.

We performed PCR to determine the prevalence of W_nqrC and W_clsA_2 in *E. coli* isolates in river water. Of the 80 isolates tested, eight (10.0%) and six (7.5%) were positive for W_nqrC and W_clsA_2, respectively (<strong>Table 2</strong>). The prevalence of W_nqrC and W_clsA_2 was lower than but almost the same as that in *E. coli* isolates obtained from wastewater effluents (10.2% and 8.2%), which is congruent with the fact that the sampling points were strongly affected by wastewater.

We also performed <em>in silico</em> PCR on RefSeq environmental *E. coli* genomes. Of the 303 genomes, 21 genomes (6.9%), one genome (0.3%), 12 genomes (4.0%), and eight genomes...
(2.6%) were positive for W_nqrC, W_clsA_2, H8, and H12, respectively (Table 2, Table S10). The prevalence of MST markers was lower than in the RefSeq wastewater E. coli genomes: this is consistent with the fact that not all environmental E. coli originate from wastewater. We note that the two markers developed in the present study, especially W_nqrC, are commonly found on plasmids and could be subjected to horizontal gene transfer (HGT) in the environment. Previous studies suggested that HGT could occur, for example, in biofilms, although HGT by conjugation rarely occurs between motile planktonic cells (Abe et al., 2020). Though we assume the contribution of HGT to the spread of the developed markers to be low in environmental waters, where water is constantly flowing, this should be kept in mind.

3.8. Study limitations.

This study has some limitations. First, when used in combination, the developed markers showed quite high specificity but showed relatively low sensitivity, indicating that a relatively large number of E. coli isolates would be needed in MST for each environmental sample. This was due to the tradeoff between the sensitivity and specificity of genetic markers (Figure 2). We assume this limitation to stem from using E. coli as an MST target organism rather than the methodologies employed in our study. Actually, not a few E. coli strains appear to be able to colonize multiple host species (Gomi et al., 2014; Tenaillon et al., 2010), which might have hampered the development of high-sensitivity and high-specificity MST markers. However,
we note that the sensitivity calculated in this study is at the isolate level, and sample-level
sensitivity (i.e., sensitivity calculated for a wastewater sample, which can contain multiple E.
coli strains) should be higher (Senkbeil et al., 2019).

Second, the developed markers were tested at the isolate level (i.e., in a culture-dependent
manner) but not tested for DNA directly extracted from samples (i.e., in a culture-independent
manner). It should be noted that in silico PCR analysis using the nr database on the Primer-
BLAST website, which also includes the genomes of non-E. coli species, returned PCR
products from non-E. coli species for both W_nqrC (e.g., Klebsiella pneumoniae and
Salmonella enterica) and W_clsA_2 (e.g., Pseudomonas aeruginosa and K. pneumoniae). In
the case of W_clsA_2, only 29 (18%) of 157 products were from E. coli. However, in the case
of W_nqrC, 517 (93%) of 556 products were from E. coli, and only 39 (7%) were from other
species. Moreover, the hosts of those 39 genomes from other species were found to be mainly
from humans (Homo sapiens), while only two were from animals. (Note that the hosts of some
genomes could not be identified due to the lack of metadata information in the GenBank
database.) Interestingly, three of the 39 products were from plasmids of uncultured bacteria
(accession numbers: AJ851089.1, JX127248.1, and MF554641.1), all of which were from
WWTPs. These results indicate that W_nqrC could be used in a culture-independent method
(note that this does not preclude the applicability of W_clsA_2 to a culture-independent
method). We also observed H8 to be prevalent among non-E. coli bacteria (Gomi et al., 2014),
but it was proven in a subsequent study to be applicable in a culture-independent manner (Senkbeil et al., 2019). These indicate that further studies are needed to test the sensitivity and specificity of W_nqrC and W_clsA_2 using DNA directly extracted from wastewater and feces from animal hosts to verify their applicability to culture-independent methods. Future studies should also assess the applicability of these markers to DNA directly extracted from environmental samples. Nonetheless, we believe the two markers to be useful, since culture-dependent methods are used to detect *E. coli* in routine water quality assessments, and there is a need to identify the sources of the detected *E. coli* colonies. Moreover, cell suspensions prepared from *E. coli* collected in routine water quality assessments can be used directly for PCR reactions without DNA extraction. We also found it possible to use colored *E. coli* colonies on selective media, in our case XM-G agar, for PCR reactions.

4. Conclusions

In the present study we identified two *E. coli* MST markers, namely W_nqrC and W_clsA_2, which can be used for tracking wastewater contamination in environmental waters. These two markers showed better performance than the previously developed H8 and H12 markers at the isolate level. Although further studies are needed to test the applicability of the developed markers in a culture-independent manner, this study showed that the developed MST markers
could be useful for identifying the sources of *E. coli* (whether they originated from domestic wastewater or not) at least at the isolate level. We believe that the PCR assays developed in this study can be readily applied to identify the sources of *E. coli* isolates collected during routine water quality assessments.

**Acknowledgements**

This work was supported by the Environment Research and Technology Development Fund (JPMEERF20205006) of the Ministry of the Environment Japan, and Keihanshin Consortium for Fostering the Next Generation of Global Leaders in Research (K-CONNEX), established by Human Resource Development Program for Science and Technology, MEXT. Computations were partially performed on the NIG supercomputer at ROIS National Institute of Genetics.

**Supporting information**

Supplementary information associated with this article can be found, in the online version, at _____.
Abe, K., Nomura, N. and Suzuki, S.  2020.  Biofilms: hot spots of horizontal gene transfer (HGT) in aquatic environments, with a focus on a new HGT mechanism. FEMS Microbiol Ecol 96(5).

Beghain, J., Bridier-Nahmias, A., Le Nagard, H., Denamur, E. and Clermont, O.  2018.  ClermonTyping: an easy-to-use and accurate in silico method for Escherichia genus strain phylotyping. Microb Genom 4(7).

Boehm, A.B., Graham, K.E. and Jennings, W.C.  2018.  Can We Swim Yet? Systematic Review, Meta-Analysis, and Risk Assessment of Aging Sewage in Surface Waters. Environ Sci Technol 52(17), 9634-9645.

Brynildsrud, O., Bohlin, J., Scheffer, L. and Eldholm, V.  2016.  Rapid scoring of genes in microbial pan-genome-wide association studies with Scoary. Genome Biol 17(1), 238.

Castro-Hermida, J.A., Garcia-Presedo, I., Almeida, A., Gonzalez-Warleta, M., Correia Da Costa, J.M. and Mezo, M.  2008.  Contribution of treated wastewater to the contamination of recreational river areas with Cryptosporidium spp. and Giardia duodenalis. Water Res 42(13), 3528-3538.

Chen, S., Zhou, Y., Chen, Y. and Gu, J.  2018.  fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics 34(17), i884-i890.

Constantinides, B., Chau, K.K., Quan, T.P., Rodger, G., Andersson, M.I., Jeffery, K., Lipworth, S., Gweon, H.S., Peniket, A., Pike, G., Millo, J., Byukusenge, M., Holdaway, M., Gibbons, C., Mathers, A.J., Crook, D.W., Peto, T.E.A., Walker, A.S. and Stoesser, N.  2020.  Genomic surveillance of Escherichia coli and Klebsiella spp. in hospital sink drains and patients. Microb Genom 6(7).

Dolejska, M., Frolkova, P., Florek, M., Jamborova, I., Purgertova, M., Kutilova, I., Cizek, A., Guenther, S. and Literak, I.  2011.  CTX-M-15-producing Escherichia coli clone B2-O25b-ST131 and Klebsiella spp. isolates in municipal wastewater treatment plant effluents. J Antimicrob Chemother 66(12), 2784-2790.

Dombek, P.E., Johnson, L.K., Zimmerley, S.T. and Sadowsky, M.J.  2000.  Use of repetitive DNA sequences and the PCR To differentiate Escherichia coli isolates from human and animal sources. Appl Environ Microbiol 66(6), 2572-2577.

Field, K.G. and Samadpour, M.  2007.  Fecal source tracking, the indicator paradigm, and managing water quality. Water Res 41(16), 3517-3538.

Gomi, R., Matsuda, T., Matsui, Y. and Yoneda, M.  2014.  Fecal source tracking in water by next-generation sequencing technologies using host-specific Escherichia coli genetic markers. Environ Sci Technol 48(16), 9616-9623.

Gomi, R., Matsuda, T., Matsumura, Y., Yamamoto, M., Tanaka, M., Ichiyama, S. and Yoneda, M.  2017a.  Occurrence of Clinically Important Lineages, Including the Sequence Type 131 C1-M27 Subclone, among Extended-Spectrum-beta-Lactamase-Producing Escherichia coli in Wastewater. Antimicrob Agents Chemother 61(9), e00564-00517.

Gomi, R., Matsuda, T., Matsumura, Y., Yamamoto, M., Tanaka, M., Ichiyama, S. and Yoneda, M.  2017b.  Whole-Genome Analysis of Antimicrobial-Resistant and Extraintestinal Pathogenic Escherichia coli in River Water. Appl Environ Microbiol 83(5), e02703-02716.
Gonzalez-Fernandez, A., Symonds, E.M., Gallard-Gongora, J.F., Mull, B., Lukasik, J.O., Rivera Navarro, P., Badilla-Aguilar, A., Peraud, J., Brown, M.L., Mora Alvarado, D., Breitbart, M., Cairns, M.R. and Harwood, V.J. 2021. Relationships among microbial indicators of fecal pollution, microbial source tracking markers, and pathogens in Costa Rican coastal waters. Water Res 188, 116507.

Harwood, V.J., Staley, C., Badgley, B.D., Borges, K. and Korajkic, A. 2014. Microbial source tracking markers for detection of fecal contamination in environmental waters: relationships between pathogens and human health outcomes. FEMS Microbiol Rev 38(1), 1-40.

Ishii, S., Ksoll, W.B., Hicks, R.E. and Sadowsky, M.J. 2006. Presence and growth of naturalized *Escherichia coli* in temperate soils from Lake Superior watersheds. Appl Environ Microbiol 72(1), 612-621.

Johnson, J.R. and Clabots, C. 2006. Sharing of virulent *Escherichia coli* clones among household members of a woman with acute cystitis. Clin Infect Dis 43(10), e101-108.

Lindsey, R.L., Garcia-Toledo, L., Fasulo, D., Gladney, L.M. and Strockbine, N. 2017. Multiplex polymerase chain reaction for identification of *Escherichia coli*, *Escherichia albertii* and *Escherichia fergusonii*. J Microbiol Methods 140, 1-4.

Maheux, A.F., Picard, F.J., Boissinot, M., Bissonnette, L., Paradis, S. and Bergeron, M.G. 2009. Analytical comparison of nine PCR primer sets designed to detect the presence of *Escherichia coli/Shigella* in water samples. Water Res 43(12), 3019-3028.

Mathai, P.P., Staley, C. and Sadowsky, M.J. 2020. Sequence-enabled community-based microbial source tracking in surface waters using machine learning classification: A review. J Microbiol Methods 177, 106050.

Naidoo, S. and Olaniran, A.O. 2013. Treated wastewater effluent as a source of microbial pollution of surface water resources. Int J Environ Res Public Health 11(1), 249-270.

Nicolas-Chanoine, M.H., Bertrand, X. and Madec, J.Y. 2014. *Escherichia coli* ST131, an intriguing clonal group. Clin Microbiol Rev 27(3), 543-574.

Nopprapun, P., Boonthanon, S.K., Harada, H., Surinkul, N. and Fuji, S. 2020. Evaluation of a human-associated genetic marker for *Escherichia coli* (H8) for fecal source tracking in Thailand. Water Sci Technol 82(12), 2929-2936.

Page, A.J., Cummins, C.A., Hunt, M., Wong, V.K., Reuter, S., Holden, M.T., Fookes, M., Falush, D., Keane, J.A. and Parkhill, J. 2015. Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics 31(22), 3691-3693.

Riley, L.W. 2014. Pandemic lineages of extraintestinal pathogenic *Escherichia coli*. Clin Microbiol Infect 20(5), 380-390.

Scott, T.M., Rose, J.B., Jenkins, T.M., Farrah, S.R. and Lukasik, J. 2002. Microbial source tracking: current methodology and future directions. Appl Environ Microbiol 68(12), 5796-5803.

Seemann, T. 2014. Prokka: rapid prokaryotic genome annotation. Bioinformatics 30(14), 2068-2069.

Senkbeil, J.K., Ahmed, W., Conrad, J. and Harwood, V.J. 2019. Use of *Escherichia coli* genes associated with human sewage to track fecal contamination source in subtropical waters. Sci Total Environ 686, 1069-1075.
Sghaier, S., Abbassi, M.S., Pascual, A., Serrano, L., Diaz-De-Alba, P., Said, M.B., Hassen, B., Ibrahim, C., Hassen, A. and Lopez-Cerero, L. 2019. Extended-spectrum beta-lactamase-producing Enterobacteriaceae from animal origin and wastewater in Tunisia: first detection of O25b-B23-CTX-M-27-ST131 Escherichia coli and CTX-M-15/OXA-204-producing Citrobacter freundii from wastewater. J Glob Antimicrob Resist 17, 189-194.

Soller, J.A., Schoen, M.E., Bartrand, T., Ravenscroft, J.E. and 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.

Steele, M. and Odumeru, J. 2004. Irrigation water as source of foodborne pathogens on fruit and vegetables. J Food Prot 67(12), 2839-2849.

Stenske, K.A., Bemis, D.A., Gillespie, B.E., D’Souza, D.H., Oliver, S.P., Draughon, F.A., Matteson, K.J. and Bartges, J.W. 2009. Comparison of clonal relatedness and antimicrobial susceptibility of fecal Escherichia coli from healthy dogs and their owners. Am J Vet Res 70(9), 1108-1116.

Stoesser, N., Sheppard, A.E., Parkhurst, L., De Maio, N., Moore, C.E., Sebra, R., Turner, P., Anson, L.W., Kasarskis, A., Batty, E.M., Kos, V., Wilson, D.J., Phetsouvanh, R., Wylie, D., Sokurenko, E., Manges, A.R., Johnson, T.J., Price, L.B., Peto, T.E., Johnson, J.R., Didelot, X., Walker, A.S., Crook, D.W. and Modernizing Medical Microbiology Informatics, G. 2016. Evolutionary History of the Global Emergence of the Escherichia coli Epidemic Clone ST131. mBio 7(2), e02162.

Teixeira, P., Dias, D., Costa, S., Brown, B., Silva, S. and Valerio, E. 2020. Bacteroides spp. and traditional fecal indicator bacteria in water quality assessment - An integrated approach for hydric resources management in urban centers. J Environ Manage 271, 110989.

Tenaillon, O., Skurnik, D., Picard, B. and Denamur, E. 2010. The population genetics of commensal Escherichia coli. Nat Rev Microbiol 8(3), 207-217.

Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B.C., Remm, M. and Rozen, S.G. 2012. Primer3-new capabilities and interfaces. Nucleic Acids Res 40(15), e115.

Warish, A., Triplett, C., Gomi, R., Gyawali, P., Hodgers, L. and Toze, S. 2015. Assessment of Genetic Markers for Tracking the Sources of Human Wastewater Associated Escherichia coli in Environmental Waters. Environ Sci Technol 49(15), 9341-9346.

Wick, R.R., Judd, L.M., Gorrie, C.L. and Holt, K.E. 2017. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. PLoS Comput Biol 13(6), e1005595.

Ye, J., Coulouris, G., Zaretskaya, I., Cutcutache, I., Rozen, S. and Madden, T.L. 2012. Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction. BMC Bioinformatics 13, 134.

Zhi, S., Banting, G., Stothard, P., Ashbolt, N.J., Checkley, S., Meyer, K., Otto, S. and Neumann, N.F. 2019. Evidence for the evolution, clonal expansion and global dissemination of water treatment-resistant naturalized strains of Escherichia coli in wastewater. Water Res 156, 208-222.

Zhi, S., Stothard, P., Banting, G., Scott, C., Huntley, K., Ryu, K., Otto, S., Ashbolt, N., Checkley, S., Dong, T., Ruecker, N.J. and Neumann, N.F. 2020. Characterization of water treatment-resistant and multidrug-resistant urinary pathogenic Escherichia coli in treated wastewater. Water Res 182, 115827.
Zhou, Z., Alikhan, N.F., Mohamed, K., Fan, Y., Agama Study, G. and Achtman, M. 2020. The EnteroBase user's guide, with case studies on Salmonella transmissions, Yersinia pestis phylogeny, and Escherichia core genomic diversity. Genome Res 30(1), 138-152.