Development of two microbial source tracking markers for detection of wastewater-associated *Escherichia coli* isolates

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

[a]Department of Environmental Engineering, Graduate School of Engineering, Kyoto University, Katsura, Nishikyo-ku, 615-8540, Kyoto, Japan

[b]Interdisciplinary Center for River Basin Environment, University of Yamanashi, 4-3-11 Takeda, Kofu, 400-8511, Yamanashi, Japan

[c]Research Center for Environmental Quality Management, Graduate School of Engineering, Kyoto University, 1-2 Yumihama, Otsu, 520-0811, Shiga, Japan

[d]Department of Engineering, University of Yamanashi, Kofu, 400-8511, Yamanashi, Japan

[e]Faculty of Agriculture and Marine Science, Kochi University, Nankoku, 783-8502, Kochi, Japan

*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

18 **ABSTRACT**

19 *Escherichia coli* has been used as an indicator of fecal pollution in environmental waters.
20 However, its presence in environmental waters does not provide information on the source of
21 water pollution. Identifying the source of water pollution is paramount to be able to effectively
22 reduce contamination. The present study aimed to identify *E. coli* microbial source tracking
23 (MST) markers that can be used to identify domestic wastewater contamination in
24 environmental waters. We first analyzed wastewater *E. coli* genomes sequenced by us (n = 50)
25 and RefSeq animal *E. coli* genomes of fecal origin (n = 82), and identified 144 candidate
26 wastewater-associated marker genes. The sensitivity and specificity of the candidate marker
27 genes were then assessed by screening the genes in 335 RefSeq wastewater *E. coli* genomes
28 and 3,318 RefSeq animal *E. coli* genomes. We finally identified two MST markers, namely
29 W_nqrC and W_clsA_2, which could be used for detection of wastewater-associated *E. coli*
30 isolates. These two markers showed higher performance than the previously developed human
31 wastewater-associated *E. coli* markers H8 and H12. When used in combination, W_nqrC and
32 W_clsA_2 showed specificity of 98.9% and sensitivity of 25.7%. PCR assays to detect
33 W_nqrC and W_clsA_2 were also developed and validated. The developed PCR assays are
34 potentially useful for detecting *E. coli* isolates of wastewater origin in environmental waters,
35 though users should keep in mind that the sensitivity of these markers is not high. Further

36 studies are needed to assess the applicability of the developed markers to a culture-independent

37 approach.

38

39 **KEYWORDS**: *Escherichia coli*, fecal indicator bacteria, microbial source tracking, whole

40 genome sequencing, genomic analysis, PCR

41

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, it is a standard practice to monitor fecal indicator bacteria (FIB), such as total coliforms, fecal coliforms, and *Escherichia coli* (Field and Samadpour, 2007; Naidoo and Olaniran, 2013). Although FIB are not normally pathogenic to humans, their presence in water indicates the possible presence of pathogens (e.g., *Campylobacter* spp., *Salmonella* spp., and *Cryptosporidium parvum*), which also live in human and animal digestive systems (Ahmed et al., 2019b). 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). 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). Identifying the origin of water pollution is paramount in order to assess the associated health risks and determine the actions needed to reduce contamination (Scott et al., 2002).

60 Microbial source tracking (MST) is an approach used to determine the sources of fecal
61 contamination in environmental waters (Harwood et al., 2014). Because culture-independent
62 methods have many advantages, such as being fast and not labor-intensive, the majority of
63 MST methods in use today are culture-independent. For example, markers such as the 16S
64 rRNA gene of *Bacteroides* spp. and pepper mild mottle virus (PMMoV) have been used to
65 track the sources of fecal contamination in MST in a culture-independent manner (Gonzalez-
66 Fernandez et al., 2021; Mathai et al., 2020). However, routine water quality assessment has
67 focused primarily on traditional FIB, and alternative indicators such as *Bacteroides* spp. are
68 not usually monitored for this purpose (Senkbeil et al., 2019; Teixeira et al., 2020). Because *E.*
69 *coli* is still a widely-used FIB, MST markers targeting *E. coli* would be useful if included in
70 the course of routine monitoring. Even a culture-dependent approach would be useful for *E.*
71 *coli*, because users do not have to collect isolates exclusively for MST purposes but can use *E.*
72 *coli* colonies obtained in routine water quality assessments.
73 *E. coli* has been used and studied as an indicator of fecal contamination for many years. Some
74 previous studies have focused on host-specific traits of this organism to identify potential
75 sources of fecal contamination. Earlier studies employed library-dependent techniques such as
76 rep-PCR, ribotyping, and denaturing gradient gel electrophoresis (DGGE) to differentiate *E.*
77 *coli* from different sources (Buchan et al., 2001; Carson et al., 2001; Dombek et al., 2000).
78 Later, a few studies identified library-independent *E. coli* markers such as cattle and swine-

79 specific *E. coli* toxin genes (Khatib et al., 2002; 2003), though none seem to be widely used
80 today. While *E. coli* has been used in both library-dependent and -independent MST studies, it
81 has been suggested that *E. coli* is not an ideal target for MST studies because it may not be
82 distinct enough to be separated into host-specific groups (Ishii and Sadowsky, 2008). The
83 sharing/transmission of *E. coli* clones among different host species has also been reported
84 (Johnson and Clabots, 2006; Li et al., 2019; Stenske et al., 2009). However, a number of studies,
85 for example those based on genome analysis, also report certain host-associated traits of *E. coli*
86 (Lupolova et al., 2017; Yu et al., 2021), indicating there to be some level of host-specificity in
87 this species. Although little success has been achieved in developing MST markers targeting
88 *E. coli*, we have previously identified a set of human wastewater, cattle, pig, and chicken feces-
89 associated *E. coli* MST markers (Gomi et al., 2014). The following studies reported two of the
90 proposed markers, namely H8 and H12, showed relatively high sensitivity and specificity and
91 could be useful as human wastewater-associated markers (Nopprapun et al., 2020; Senkbeil et
92 al., 2019; Warish et al., 2015). However, the original genetic markers were developed by
93 analyzing a small number of *E. coli* genomes (n = 22). Since then, a growing number of *E. coli*
94 genomes have become available in public databases. These *E. coli* genomes include those from
95 various isolation sources and geographically diverse hosts, which might allow the development
96 of reliable MST markers that can be used globally. Here, we analyzed *E. coli* genomes in the
97 database of the National Center for Biotechnology Information (NCBI) as well as genomes

98 sequenced by us to identify more reliable *E. coli* genetic markers for tracking domestic
99 wastewater contamination in environmental waters. Our aim in the present study was to
100 identify culture-dependent MST markers, but applicability to wastewater samples was also
101 tested in a culture-independent manner. The sensitivity and specificity of the previously
102 developed human wastewater-associated markers H8 and H12 were also re-evaluated.
103
104
105 **2. Materials and methods**
106 **2.1. *E. coli* isolation from wastewater and genome sequencing.**
107 The flowchart of the study design is shown in **Figure 1**. We obtained wastewater *E. coli* isolates
108 from three municipal wastewater treatment plants (WWTPs), namely WWTP A, WWTP B,
109 and WWTP C, in Shiga Prefecture, Japan. The amount of sewage treated by each WWTP per
110 day is 80,000 m^3 for WWTP A, 60,000 m^3 for WWTP B, and 250,000 m^3 for WWTP C. Almost
111 all the areas served by these WWTPs have separate sewer systems, and the WWTPs mainly
112 receive domestic wastewater. Biologically treated wastewater samples before/after
113 chlorination were collected from the WWTPs between December 2016 and December 2017.
114 The chlorinated samples were treated with sodium thiosulfate immediately after collection to
115 neutralize the chlorine. Samples were collected using sterile sampling bags or sterile sampling
116 bottles, stored at 4 °C, and processed within 24 hours. Samples were serially diluted with PBS

117 (ten-fold serial dilution), and 1 mL of undiluted/diluted samples were processed using the pour
118 plate method with XM-G agar (Nissui, Tokyo, Japan). Plates were incubated at 37 °C for 18h,
119 and colonies were randomly picked from the plates. Colonies suspected to be mixed were sub-
120 cultured with fresh XM-G agar plates to obtain pure isolates. A total of 50 isolates (29 isolates
121 from WWTP A, nine isolates from WWTP B, and 12 isolates from WWTP C) were obtained
122 for genome sequencing (see **Table S1** for information, such as sample type and collection date,
123 on the collected *E. coli*).
124 Colonies were resuspended in PBS and centrifuged to prepare a cell pellet. DNA was extracted
125 from the pellet using a DNeasy blood and tissue kit (Qiagen, Hilden, Germany) following the
126 manufacturer's instructions, and the concentration of extracted DNA was measured using a
127 Qubit fluorometer with dsDNA High sensitivity kit (Thermo Fisher Scientific, Waltham, MA,
128 United States). Extracted DNA was stored at -30°C for up to a week until sample preparation
129 for genome sequencing. The extracted DNA was processed using a Nextera XT DNA sample
130 preparation kit (Illumina, San Diego, CA) and sequenced on an Illumina MiSeq instrument for
131 600 cycles (300-bp paired-end sequencing) as described previously (Ma et al., 2022).
132

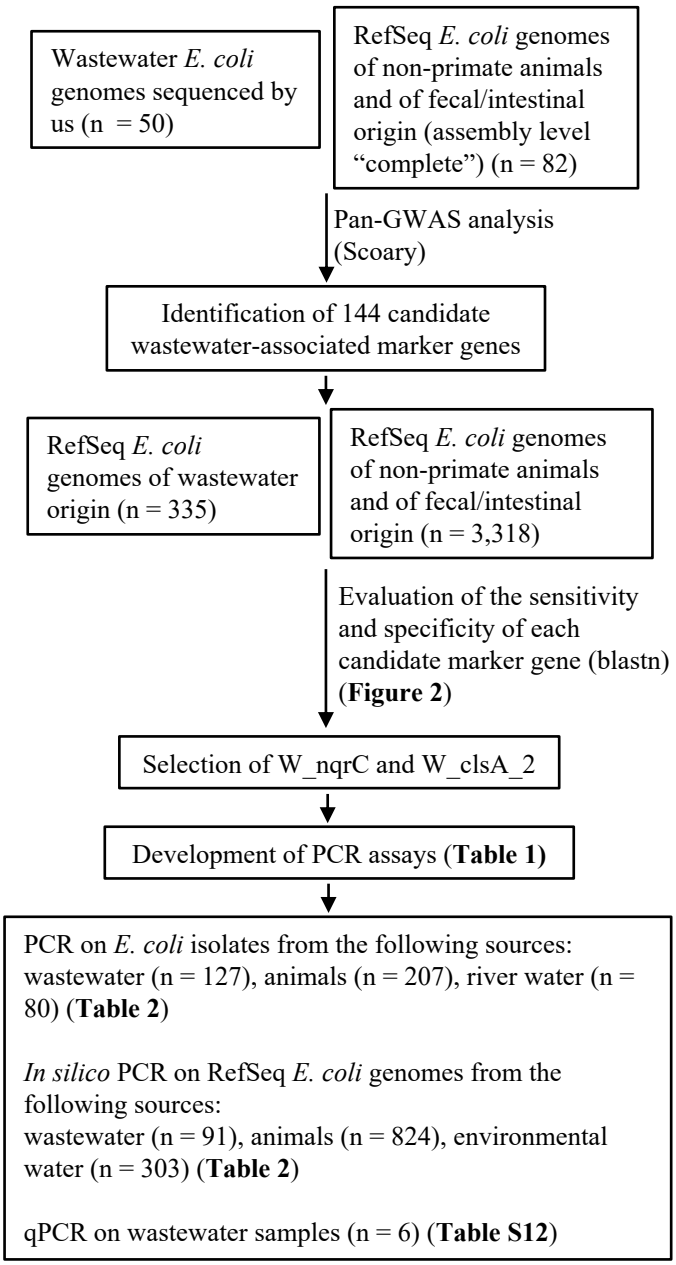


Figure 1. Overview of the strategy to identify and test the wastewater-associated *E. coli* genetic markers. RefSeq *E. coli* genomes used for blastn analysis (335 wastewater *E. coli* genomes and 3,318 animal *E. coli* genomes) and *in silico* PCR analysis (91 wastewater *E. coli* genomes, 824 animal *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.

140

141 **2.2. Retrieving complete *E. coli* genome sequences from the NCBI database.**

142 The Reference Sequence (RefSeq) *E. coli* genomes with an assembly level of "complete
143 genome" (n = 1,278) were downloaded in May 2021. Metadata information, such as strain
144 name, host, and isolation source, was extracted from the GenBank files using in-house Python
145 scripts, which uses tools such as Biopython Bio.SeqIO module (Cock et al., 2009). *E. coli*
146 genomes of non-primate animals and of fecal/intestinal origin (n = 82) were kept for further
147 analysis (*E. coli* genomes from birds were also included here. See **Table S2** for information on
148 these animal *E. coli* genomes).

149

150 **2.3. Genome assembly and annotation.**

151 Raw sequence reads from wastewater *E. coli* were trimmed using fastp (v0.20.0) (Chen et al.,
152 2018), and the trimmed reads were assembled using Unicycler (v0.4.8) with the --no_correct
153 option (Wick et al., 2017). For the complete RefSeq genomes, we generated simulated paired-
154 end reads using wgsim (v1.12) (https://github.com/lh3/wgsim). Simulated reads were trimmed
155 using fastp and assembled using Unicycler with the same parameters as those used for
156 assembling the wastewater *E. coli* genomes. We did not use the complete RefSeq genomes but
157 used these draft assemblies in the following analysis. This was to remove any biases introduced
158 by differences in assembly status (i.e., complete assemblies vs. draft assemblies) in the pan-

159 genome-wide association study (pan-GWAS). Assembled genomes were annotated using
160 Prokka (v1.14.6) (Seemann, 2014).
161
162 **2.4. Identification of candidate wastewater-associated marker genes using a pan-GWAS**
163 **approach.**
164 The pan-genome of 132 *E. coli* genomes (50 wastewater *E. coli* genomes + 82 animal *E. coli*
165 genomes) was constructed by running Roary (v3.13.0) on GFF3 files generated by Prokka
166 (Page et al., 2015). The -s option was used to turn off paralog splitting. The pan-GWAS
167 approach implemented by Scoary was used to find genes that were enriched in wastewater *E.*
168 *coli* genomes (Brynildsrud et al., 2016). Scoary (v 1.6.16) was run on Roary's gene
169 presence/absence file. The --collapse option was used to collapse genes that had identical
170 distribution patterns among the genomes. Because our aim was to identify genes that were
171 enriched in wastewater *E. coli* genomes and not to infer causal association, the --no_pairwise
172 option was used as recommended by the developers (https://github.com/AdmiralenOla/Scoary).
173 Genes with Benjamini-Hochberg adjusted p-values of lower than 0.05 and with a specificity
174 exceeding 95% were defined as candidate wastewater-associated marker genes. We set the
175 specificity threshold so as to reduce the number of false positives, considering the future
176 application to DNA directly extracted from environmental samples (this type of DNA can
177 contain DNA from multiple *E. coli* isolates and thus is sensitive to false positives). When a set

178 of collapsed genes met the p-value and specificity criteria, one gene was randomly selected as

179 a candidate marker.

180

181 **2.5. Retrieving genome sequences from the NCBI database for sensitivity and specificity**

182 **analysis.**

183 The RefSeq *E. coli* genomes (n = 20,237), including both complete and draft assemblies, were

184 downloaded in June 2021. Metadata information was extracted from the GenBank files using

185 in-house Python scripts. Genomes of wastewater origin (n = 335) and genomes of non-primate

186 animals and of fecal/intestinal origin (n = 3,318) were kept for sensitivity and specificity

187 analysis (see **Table S3** for information on the wastewater *E. coli* genomes and **Table S4** for

188 information on the animal *E. coli* genomes).

189

190 **2.6. Evaluating the sensitivity and specificity of candidate marker genes.**

191 The candidate marker genes were detected in 3,653 genome assemblies (335 RefSeq

192 wastewater *E. coli* genomes + 3,318 RefSeq animal *E. coli* genomes) using the blastn program

193 with the default parameters. Thresholds of a minimum percent identity of 80% and a minimum

194 DNA coverage of 50% were used to define the presence of a gene. For each candidate marker

195 gene, the following metrics were calculated: true positive (TP) (the number of wastewater *E.*

196 *coli* genomes positive for the candidate marker), true negative (TN) (the number of animal *E.*

197 *coli* genomes negative for the candidate marker), false positive (FP) (the number of animal *E.*
198 *coli* genomes positive for the candidate marker), and false negative (FN) (the number of
199 wastewater *E. coli* genomes negative for the candidate marker). Sensitivity [TP/(TP + FN)] and
200 specificity [TN/(TN + FP)] were then calculated for each candidate marker gene (note that
201 sensitivity and specificity were reported as %). The above blastn analysis and calculation were
202 also performed for H8 and H12. Note that H8 corresponded to one of the candidate marker
203 genes.
204
205 **2.7. Development of PCR assays.**
206 Primers that yield desired product sizes were designed for the identified marker genes as shown
207 in **Table 1**. For each target gene, the validity of the primers was checked with Primer 3 software
208 using the default settings (Untergasser et al., 2012). Moreover, using Primer-BLAST with the
209 default settings, we confirmed that an amplicon of the desired size was produced from each
210 wastewater genome with the target gene that had been sequenced in **2.1** (Ye et al., 2012).
211 To further validate the developed PCR primers, we performed PCR on *E. coli* isolates collected
212 from the following sources: wastewater influent (n = 78), wastewater effluent (after biological
213 treatment and chlorination) (n = 49), cattle (n = 39), chickens (n = 30), deer (n = 30), pigs (n =
214 30), waterfowl (n = 30), cats (n = 24), and dogs (n = 24). Detailed information on these isolates
215 is available in **Table S5**. Briefly, wastewater influent isolates (n = 48) and effluent isolates (n

216 = 49) were newly collected in 2021 from WWTP A in Shiga Prefecture, Japan, and other
217 wastewater influent isolates (n = 30) were collected in 2018 from a WWTP (WWTP D) in
218 Yamanashi Prefecture, Japan. WWTP D treats 110,000 m^3 of sewage per day and the area
219 served by this WWTP has separate sewer systems. We used the same sampling protocol and *E.*
220 *coli* isolation procedure as described in **2.1**, except that we used Chromocult coliform agar
221 (Merck, Darmstadt, Germany) or CHROMagar ECC (CHROMagar, Paris, France) and
222 incubated the plates for ~24 hours at 37 °C to isolate *E. coli* from the influent samples collected
223 from WWTP D. Animal *E. coli* isolates were obtained from animal feces collected in Kyoto
224 Prefecture, Nara Prefecture, Shiga Prefecture, or Yamanashi Prefecture between 2011 and 2022,
225 as described in **Table S5**. All animal feces were collected using sterile sampling bags or sterile
226 centrifuge tubes, stored at 4 °C, and processed within 24-48 hours. Animal feces were directly
227 streaked onto XM-G agar using sterile swabs/inoculating needles or suspended and diluted
228 with PBS and processed using the pour plate method with XM-G agar (for two deer fecal
229 samples, the pour plate method with CHROMagar ECC was used). The plates were incubated
230 for 18-24 hours at 37 °C. Sub-culturing was performed to obtain pure colonies. Up to three
231 isolates were obtained from a single individual (Dombek et al., 2000).
232 DNA was extracted from *E. coli* obtained from WWTP A and cattle and deer feces using a
233 DNeasy blood and tissue kit. The concentration of extracted DNA was measured using a Qubit
234 fluorometer with dsDNA High sensitivity kit. The extracted DNA was stored at -30°C for up

235 to one year until the PCR experiments. For *E. coli* obtained from WWTP D and other animal
236 feces, we used cell suspensions as the inputs in the PCR experiments. Cell suspensions were
237 prepared by either of the following two methods: (i) a single *E. coli* colony grown on the XM-
238 G plate was picked up using a clean tip, suspended in 500 μL of ultrapure water, and then
239 vortexed; or (ii) a glycerol stock (*E. coli* stored in 25% glycerol) was diluted 10 times using
240 ultrapure water. PCR reactions were performed on the extracted DNA or cell suspension using
241 QuantiFast SYBR Green PCR Kit (Qiagen, Hilden, Germany) or QuantiNova SYBR Green
242 PCR Kit (Qiagen) with PCR conditions as described below. For QuantiFast SYBR Green PCR
243 Kit, PCR mixture (25 μL) was composed of 12.5 μL of 2× QuantiFast SYBR Green PCR
244 Master Mix, 2.5 μL each of forward and reverse primers (5 μM), 5 μL of template DNA or cell
245 suspension, and 2.5 μL of PCR grade water or ultrapure water. The reactions were carried out
246 by incubation at 95 °C for 5 min, followed by 25 to 30 cycles (depending on the template
247 concentrations) consisting of 95 °C (denaturation) for 10 s and 60 °C (combined
248 annealing/extension) for 30 s. A melting curve analysis was performed after the amplification.
249 For QuantiNova SYBR Green PCR Kit, PCR mixture (20 μL) was composed of 10 μL of 2×
250 QuantiNova SYBR Green PCR Master Mix, 2.8 μL each of forward and reverse primers (5
251 μM), 2 μL of template DNA or cell suspension, and 2.4 μL of PCR grade water or ultrapure
252 water. The reactions were carried out by incubation at 95 °C for 2 min, followed by 30 cycles
253 consisting of 95 °C (denaturation) for 5 s and 60 °C (combined annealing/extension) for 10 s.

254 A melting curve analysis was performed after the amplification. In addition to the primers for
255 W_nqrC and W_clsA_2, we also used previously published primers for amplifying H8, H12,
256 and also *uidA* (primers UAL and UAR): we included *uidA* primers to confirm the *E. coli*
257 identification (Gomi et al., 2014; Maheux et al., 2009). All PCR reactions were performed on
258 a Thermal Cycler Dice Real Time System III (Takara Bio, Inc.). Previously sequenced *E. coli*
259 isolates, namely KOr026 (positive: W_clsA_2 and H8, negative: W_nqrC and H12, BioSample
260 accession: SAMD00053091) and KMi027 (positive: W_nqrC and H12, negative: W_clsA_2
261 and H8, BioSample accession: SAMD00053165), were used as positive/negative controls for
262 both extracted DNA and cell suspensions (Gomi et al., 2017b). We also included no template
263 controls. The correct PCR products were confirmed by comparison with the melting curves of
264 the positive controls, which showed peaks at melting temperatures of 82.5 °C, 89.0 °C, 91.6 °C,
265 85.7 °C, and 84.4 °C for W_nqrC, W_clsA_2, H8, H12, and *uidA*, respectively. We performed
266 replicate reactions (duplicate reactions) for a subset of samples (50 samples, including 42 for
267 extracted DNA and 8 for cell suspensions) and confirmed the results to be consistent between
268 all technical replicates.
269 We also performed *in silico* PCR using the developed primers on RefSeq *E. coli* genomes that
270 were made public after June 2021 (i.e., genomes that were not used for blastn analysis). This
271 was to test the developed PCR primers using *E. coli* genomes of diverse origins and from
272 various geographical locations. *In silico* PCR was previously shown to be highly concordant

273 with *in vitro* PCR when used for gene detection in *E. coli* (Beghain et al., 2018; Lindsey et al.,

274 2017). Briefly, RefSeq *E. coli* genomes, including both complete and draft assemblies, were

275 downloaded in August 2022. Among the downloaded genomes, those made public after June

276 2021 were retained (n = 6,399). Metadata information was extracted from the retained

277 GenBank files, and the genomes of wastewater origin (n = 91) and those of non-primate

278 animals and of fecal/intestinal origin (n = 824) were retrieved. *In silico* PCR was performed on

279 the Primer-BLAST website using primer sequences of W_nqrC, W_clsA_2, H8, and H12 with

280 the default parameters, except that we used 500 bp as max target amplicon size (Ye et al., 2012).

281

282 **Table 1**. Primers designed in this study.

Target	Sequence (5′-3′)	Product size (bp)
W_nqrC	F: CATGCGTTTCTTGCTCTGGG	105
	R: GTTGGCGTCTTCCACCCTAG	
W_clsA_2	F: ACCGATGATTGGTGCGATAC	165
	R: GCTCAAAGCGCCCATGAAG	

283

284 **2.8. Detection of MST markers in *E. coli* isolates from environmental waters.**

285 *E. coli* isolates were obtained in January 2022 from a river in Kyoto Prefecture in Japan. The

286 sampling points in the river were downstream of two wastewater treatment plants and known

287 to be heavily affected by treated wastewater (**Figure S1**). A total of eight water samples were

288 collected from the river using sterile sampling bags, stored at 4 °C, and processed within 24

289 hours. Ten *E. coli* colonies were obtained from each sample using the membrane filter method

290 with XM-G agar plates. The plates were incubated for 18 hours at 37 °C, and colonies were
291 sub-cultured with fresh XM-G agar plates to obtain pure isolates. PCR reactions were
292 performed using cell suspensions as described above to test for the presence of W_nqrC and
293 W_clsA_2 in the obtained *E. coli* isolates.
294 We also performed *in silico* PCR on RefSeq *E. coli* genomes of environmental water origin.
295 Environmental *E. coli* genomes were extracted from the RefSeq *E. coli* genomes downloaded
296 in August 2022 based on the metadata information. *In silico* PCR was performed using the
297 retrieved genomes (n = 303) on the Primer-BLAST website as described above.
298
299 **2.9. Quantitative PCR (qPCR) assays on DNA extracts from wastewater samples.**
300 qPCR assays were performed to determine the concentrations of the developed markers in total
301 DNA extracted from wastewater samples. Three primary effluent samples and three
302 biologically treated wastewater samples before chlorination were taken from WWTP A on
303 different occasions in October and November 2022. Primary effluent samples (40–80 mL) and
304 treated wastewater samples (1000 mL) were filtered through 0.2 μm pore size polycarbonate
305 membranes (Advantec, Tokyo, Japan). Multiple membranes were used when membranes
306 became clogged during filtration. Filters were stored at -30°C until DNA extraction. DNA was
307 extracted from each filter using a FastDNA SPIN Kit for Soil (MP Biomedicals, Irvine, CA),
308 and PCR inhibitors were removed from the DNA extracts using CHROMA SPIN+TE-400

309 Columns (Takara Bio, Inc.) and Amicon Ultra-0.5 (100K) Centrifugal Filter Devices (Millipore,
310 Darmstadt, Germany). qPCR was performed using a QuantiNova SYBR Green PCR Kit under
311 the same conditions as PCR performed on *E. coli* isolates, except that 35 cycles of reactions
312 were performed. The qPCR reactions were performed in duplicate. The correct PCR products
313 were confirmed by comparison with the melting curves of the positive controls. qPCR
314 standards were prepared using 10-fold serial dilutions of genomic DNA extracted from pure
315 cultures of KOr026 and KMi027. The copy numbers (T) of genomic DNA used for standard
316 curve construction were estimated using the equation $T = [D/(5 \times 10^6 \times 660)] \times 6.022 \times 10^{23}$,
317 where D (g/μl) is the DNA concentration.
318
319 **2.10 Sequence Data Accession Number(s).**
320 The raw sequencing data for the wastewater *E. coli* isolates were deposited in the NCBI SRA
321 under BioProject accession number PRJNA765288.
322
323
324 **3. Results and discussion**
325 **3.1. Basic characteristics of wastewater *E. coli* genomes sequenced by us.**
326 The 50 wastewater *E. coli* genomes sequenced by us were highly diverse, comprising 44
327 different sequence types (STs). These draft assemblies had a median of 140 contigs (range 49

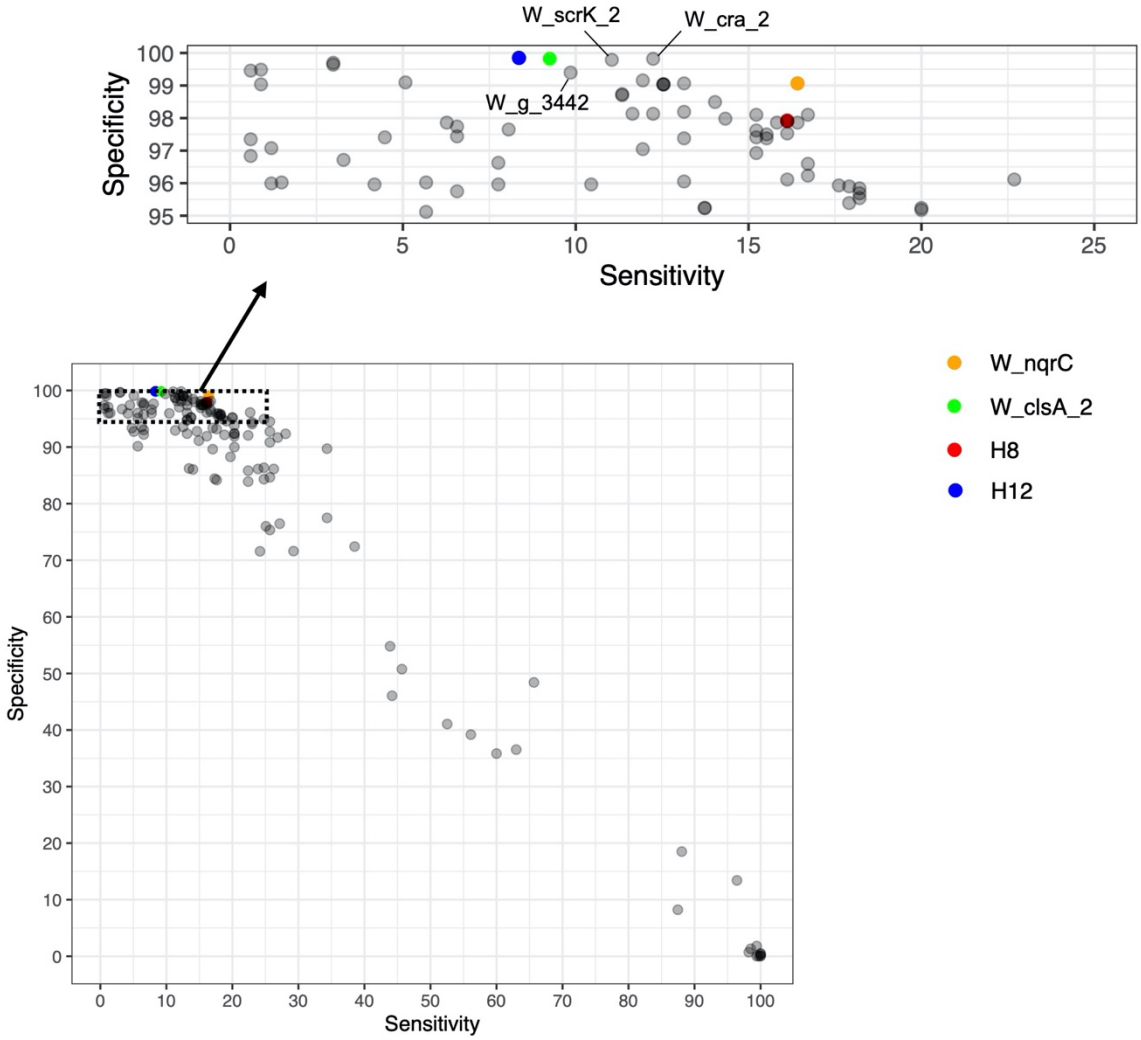
328 contigs to 346 contigs), a median genome size of 4,875,061 bp (range 4,576,995 bp to
329 5,375,884 bp), and a median N50 of 179,046 bp (range 63,435 bp to 737,274 bp) after removing
330 contigs shorter than 100 bp (**Table S1**). All 50 assemblies met the quality control criteria of (i)
331 contig numbers: ≤ 800; (ii) genome size: 3.7 Mbp to 6.4 Mbp; and (iii) N50: >20 kb, as
332 recommended by EnteroBase (Zhou et al., 2020).
333
334 **3.2. Identification of candidate wastewater-associated marker genes.**
335 A pan-genome was constructed with Roary using 50 wastewater *E. coli* genomes sequenced by
336 us and 82 RefSeq animal *E. coli* genomes. Roary identified a total of 25,343 genes, including
337 2,925 core genes (genes shared among ≥ 99% of strains) and 22,418 accessory genes (genes
338 observed among < 99% of strains). The pan-GWAS approach was employed to identify
339 candidate wastewater-associated genes among the accessory genes. A total of 144 genes met
340 the criteria of Benjamini-Hochberg adjusted p-values of less than 0.05 and specificity higher
341 than 95%, and thus were defined as candidate marker genes (see **Table S6** for details of the
342 identified genes). These 144 genes were subjected to sensitivity and specificity analysis by
343 blastn using a larger number of RefSeq *E. coli* genomes.
344
345 **3.3. Sensitivity and specificity of candidate marker genes determined by blastn analysis.**

346 The sensitivity and specificity of 144 candidate marker genes were assessed by screening these
347 genes in 335 RefSeq wastewater *E. coli* genomes and 3,318 RefSeq animal *E. coli* genomes.
348 **Figure 2** shows the distribution of sensitivity and specificity of each candidate marker gene.
349 Most candidate markers achieved high specificity, which is congruent with the fact that we
350 used a threshold of 95% specificity for identifying candidate markers in the Scoary analysis.
351 However, there were no genes that simultaneously showed high sensitivity and high specificity,
352 which may be partially due to the presence of *E. coli* isolates that can colonize multiple host
353 species (Johnson and Clabots, 2006; Stenske et al., 2009). In fact, among the genes showing
354 >70% specificity, no genes showed >40% sensitivity. It should be noted that there were genes
355 with almost 100% sensitivity and 0% specificity (i.e., detected in almost all the tested isolates)
356 even though we selected genes with specificity higher than 95% in the Scoary analysis. This
357 can be because we used blastp with a minimum percent identity of 95% for clustering
358 sequences in the Roary analysis but we used blastn with a minimum percent identify of 80%
359 for gene screening. In other words, genes assigned to different clusters by Roary can share
360 >80% nucleotide sequence identity and thus can be detected in the same blastn analysis.
361 Because the species composition of the hosts of RefSeq animal *E. coli* genomes was biased
362 (some host species were more prevalent than others), the specificity of candidate markers was
363 separately calculated for different host groups, namely, "Cattle", "Swine", "Chicken", "Wild
364 boar (*Sus scrofa*)", "Canine", "Sheep and goat", and "Others" (**Figure S2**). We also considered

365 the specificity calculated for these host groups when selecting MST markers from the

366 candidates.

367



368

369 **Figure 2**. Sensitivity and specificity of candidate wastewater-associated marker genes as

370 determined by blastn analysis. The distribution of sensitivity and specificity of 144 candidate

371 marker genes and previously identified marker genes, H8 and H12, are plotted (H8 in fact

372 corresponded to one of the 144 candidate marker genes). The positions of markers mentioned

373 in the main text (W_scrK_2, W_cra_2, and W_g_3442) are also indicated.

374

375 **3.4. Selection of W_nqrC and W_clsA_2.**

376 As noted in the Materials and Methods section, selecting a marker with high specificity is
377 required to reduce the number of false positives, especially when the marker is applied to DNA
378 directly extracted from environmental samples. We therefore selected W_nqrC from the 144
379 candidates, as this marker showed high specificity (99.1%) and the highest sensitivity (16.4%)
380 among the markers with specificity higher than 99% (**Figure 2**). Moreover, W_nqrC showed
381 the highest sensitivity among the candidate markers which showed >95% specificity for all
382 host groups (**Figure S2**). We then tried to select another marker that was prevalent among the
383 RefSeq wastewater *E. coli* genomes without W_nqrC. This was to increase the sensitivity when
384 both markers are used in combination. Among the candidate markers showing >99% specificity,
385 W_cra_2, W_scrK_2, W_clsA_2, and W_g_3442 showed different marker distribution
386 patterns (the red bars in **Figure S3**) from W_nqrC and also showed relatively high sensitivities
387 (**Figure S3**). Although W_cra_2 and W_scrK_2 showed high specificity (>99.7%), similar
388 sequences with 60%–75% nucleotide sequence identity were prevalent among animal *E. coli*
389 isolates. We therefore selected W_clsA_2 as the second marker because there were no similar
390 sequences prevalent among animal *E. coli*, and this marker showed higher specificity than
391 W_g_3442 (99.8% vs. 99.4%). Moreover, the specificity of W_clsA_2 was 100% for all host

392 groups except "Others" (only six *E. coli* genomes from silver gulls were positive for
393 W_clsA_2) (**Figure S2**). The sequences of W_nqrC and W_clsA_2 are shown in **Figure S4**.
394 Prokka annotation indicated W_nqrC encodes Na(+)-translocating NADH-quinone reductase
395 subunit C, and W_clsA_2 encodes major cardiolipin synthase ClsA. The BLASTN searches
396 against the nucleotide collection (nr/nt) database revealed that W_nqrC is mostly present on
397 plasmids, while W_clsA_2 is equally present on chromosomes and plasmids. As mentioned in
398 **2.4.**, the pan-GWAS approach used in this study simply identifies genes enriched in wastewater
399 *E. coli*, but does not infer causal association. Therefore, W_nqrC and W_clsA_2 may not
400 directly confer adaptive benefits to wastewater *E. coli*. However, these two markers are likely
401 plasmid-associated, which may partially explain the enrichment of these markers in wastewater
402 *E. coli*. This is because genes that confer resistance to antibiotics and heavy metals are also
403 often located on plasmids, which can favor the selection of bacteria with these genes in
404 wastewater (Di Cesare et al., 2016; Karkman et al., 2018; Li et al., 2015). Among the RefSeq
405 wastewater *E. coli* genomes positive for W_nqrC (n = 55), almost half of the genomes (n = 27)
406 were assigned to ST131, which was previously reported to be present among wastewater *E.*
407 *coli* (Dolejska et al., 2011; Gomi et al., 2017a; Sghaier et al., 2019) and also is known for its
408 association with extraintestinal infections (Nicolas-Chanoine et al., 2014; Riley, 2014; Stoesser
409 et al., 2016). The remaining W_nqrC-positive genomes were assigned to various STs, including
410 ST69 (n = 4), ST95 (n = 4), and ST10 (n = 3). The W_nqrC-positive wastewater *E. coli*

411 genomes sequenced by us (n = 9) belonged to diverse STs, including ST73 (n = 2), ST95 (n =
412 2), and ST357 (n = 2) (**Table S7**). Among the RefSeq wastewater *E. coli* genomes positive for
413 W_clsA_2 (n = 31), more than half of the genomes (n = 17) were assigned to ST635, which
414 was previously reported to be prevalent among wastewater *E. coli* isolates in Canada and
415 hospital sink *E. coli* isolates in the UK (Constantinides et al., 2020; Zhi et al., 2019). The
416 remaining W_clsA_2-positive genomes belonged to various STs, including ST399 (n = 3),
417 ST401 (n = 2), ST607 (n = 2), and ST3168 (n = 2). The W_clsA_2-positive wastewater *E. coli*
418 genomes sequenced by us (n = 5) belonged to five different STs (ST399, ST472, ST607, ST635,
419 and ST5295) (**Table S8**).
420 We note that W_nqrC/W_clsA_2-positive RefSeq wastewater *E. coli* genomes originated from
421 various countries, including the Czech Republic, Germany, Canada, the USA, and South Africa
422 (**Table S7**, **Table S8**). In fact, 43 (78%) of W_nqrC-positive RefSeq wastewater *E. coli*
423 genomes (n = 55) and 31 (100%) of W_clsA_2-positive RefSeq wastewater *E. coli* genomes
424 (n = 31) originated from countries other than Japan, which indicates that these markers may be
425 applicable to regions outside of our study site.
426
427 **3.5. Sensitivity and specificity of H8 and H12.**
428 We also assessed the sensitivity and specificity of the two previously developed markers, H8
429 and H12, which were calculated based on the results of blastn analysis (**Figure 2**). Notably, H8

430 corresponded to one of the 144 candidate marker genes identified by Scoary. **Figure 2** shows
431 that both H8 and H12 achieved high specificity (97.9% and 99.8%, respectively), which is
432 consistent with previous studies (Senkbeil et al., 2019; Warish et al., 2015). However, the
433 markers proposed in the present study showed better performances. W_nqrC showed higher
434 specificity and sensitivity than H8 (99.1% vs. 97.9% and 16.4% vs. 16.1%, respectively), and
435 W_clsA_2 showed almost the same specificity (~99.8%) but higher sensitivity than H12 (9.3%
436 vs. 8.4%). Moreover, when used in combination, a combination of W_nqrC and W_clsA_2
437 showed higher specificity and sensitivity than a combination of H8 and H12 (98.9% vs. 97.8%
438 and 25.7% vs. 24.2%, respectively). These results suggest that although H8 and H12 could be
439 useful for tracking wastewater contamination, W_nqrC and W_clsA_2 may yield better
440 performances.
441
442 **3.6. Development of PCR assays targeting W_nqrC and W_clsA_2.**
443 We designed PCR primers for detection of W_nqrC and W_clsA_2. PCR assays on positive
444 controls yielded the expected DNA bands for both W_nqrC (105 bp) and W_clsA_2 (165 bp),
445 whereas no bands were observed for the negative controls (**Figure S5**). The developed PCR
446 assays were applied to detect the MST markers in *E. coli* isolates obtained from wastewater
447 and animal feces (**Table 2**). W_nqrC or W_clsA_2 was detected in 21 (26.9%) and 9 (18.4%)
448 of wastewater influent and effluent isolates, respectively (**Table 2**). Importantly, there was

449 no overlap between the isolates positive for W_nqrC and W_clsA_2, which suggests that the
450 use of W_nqrC and W_clsA_2 in combination could increase the sensitivity. On the other hand,
451 only seven animal *E. coli* isolates (3.4%, six from dogs and one from waterfowl) were positive
452 for W_nqrC or W_clsA_2. We also performed PCR to detect H8 and H12 using the same set
453 of *E. coli* isolates. H8 or H12 was detected in 11 (14.1%) and 7 (14.3%) of wastewater influent
454 and effluent isolates, respectively (**Table 2**). Only two animal *E. coli* isolates (1.0%, one from
455 pigs and one from waterfowl) were positive for H8 or H12. In our PCR experiments, the
456 sensitivities of the W_nqrC and W_clsA_2 combination were higher than those of the H8 and
457 H12 combination for wastewater influent (26.9% vs 14.1%) and wastewater effluent (18.4%
458 vs 14.3%), while the specificity of the W_nqrC and W_clsA_2 combination was lower than
459 that of the H8 and H12 combination (96.6% vs 99.0%). The lower specificity of the W_nqrC
460 and W_clsA_2 combination was attributed to six dog isolates positive for W_nqrC (we note
461 that these six isolates were from two fecal samples, namely three from one and three from the
462 other). The blastn analysis also showed relatively lower specificity of W_nqrC for canine *E.*
463 *coli* isolates (**Figure S2E**). One possible reason for this is within-household transmission of *E.*
464 *coli* between dogs and humans (Johnson and Clabots, 2006; Stenske et al., 2009).
465 We also performed *in silico* PCR analysis using RefSeq *E. coli* genomes that were not used for
466 blastn analysis. Among the 91 wastewater *E. coli* genomes, 29 genomes (31.9%) were positive
467 for W_nqrC or W_clsA_2 (**Table 2**, **Table S9**). On the other hand, only 16 genomes (1.9%,

468 seven from cattle, four from birds, three from pigs, and two from dogs) among the 824 animal

469 *E. coli* genomes were positive for W_nqrC or W_clsA_2 (**Table 2**, **Table S10**). *In silico* PCR

470 analysis of H8 and H12 revealed 25 wastewater *E. coli* genomes (27.5%) to be positive for H8

471 or H12, while 27 animal *E. coli* genomes (3.3%, 11 from pigs, seven from cattle, three from

472 birds, two from goats, two from dogs, one from capybaras, and one from a South American

473 tapir) were also positive for H8 or H12 (**Table 2**, **Table S9**, **Table S10**). In the *in silico* PCR

474 analysis, the sensitivity and specificity of the W_nqrC and W_clsA_2 combination were higher

475 than those of the H8 and H12 combination (31.9% vs 27.5% and 98.1% vs 96.7%).

476

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

478 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 (%)	No. of *E. coli* isolates/ genomes positive for H8 or H12 (%)
PCR							
Wastewater influent	78	15 (19)	6 (8)	8 (10)	5 (6)	21 (27)	11 (14)[a]

Wastewater effluent	49	5 (10)	4 (8)	6 (12)	1 (2)	9 (18)	7 (14)
Cattle	39	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Chicken	30	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Deer	30	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Pig	30	0 (0)	0 (0)	1 (3)	0 (0)	0 (0)	1 (3)
Waterfowl	30	1 (3)	0 (0)	0 (0)	1 (3)	1 (3)	1 (3)
Cat	24	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Dog	24	6 (25)	0 (0)	0 (0)	0 (0)	6 (25)	0 (0)
River water	80	8 (10)	6 (8)	-[b]	-[b]	14 (18)	-[b]
***In silico* PCR**							
Wastewater[c]	91	26 (29)	3 (3)	5 (5)	20 (22)	29 (32)	25 (27)
Pig (swine, *Sus scrofa domestica*)	237	3 (1)	0 (0)	11 (5)	0 (0)	3 (1)	11 (5)
Bird (other than chicken)	199	4 (2)	0 (0)	3 (2)	0 (0)	4 (2)	3 (2)
Cattle (bovine, cow, *Bos taurus*)	178	6 (3)	1 (1)	7 (4)	0 (0)	7 (4)	7 (4)
Chicken (*Gallus gallus*)	100	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Goat/sheep	19	0 (0)	0 (0)	2 (11)	0 (0)	0 (0)	2 (11)
Dog (*Canis*	12	2 (17)	0 (0)	0 (0)	2 (17)	2 (17)	2 (17)

lupus familiaris)							
Red deer	12	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Wild boar (*Sus scrofa*)	11	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Other animals[d]	56	0 (0)	0 (0)	2 (4)	0 (0)	0 (0)	2 (4)
Environmental waters	303	21 (7)	1 (0.3)	12 (4)	8 (3)	22 (7)	20 (7)

479 [a]Two isolates carried both H8 and H12.

480 [b]We could not determine the prevalence of H8 and H12 due to the loss of glycerol stocks.

481 [c]Influent and effluent were not distinguished for many of the RefSeq wastewater *E. coli*

482 genomes.

483 [d]Animal hosts with fewer than ten *E. coli* genomes were combined into this category.

484

485

486 **3.7. Prevalence of MST markers in *E. coli* isolates from environmental waters.**

487 We performed PCR to determine the prevalence of W_nqrC and W_clsA_2 in *E. coli* isolates

488 in river water. Of the 80 isolates tested, eight (10.0%) and six (7.5%) were positive for W_nqrC

489 and W_clsA_2, respectively (**Table 2**). The prevalence of W_nqrC and W_clsA_2 was lower

490 than but almost the same as that in *E. coli* isolates obtained from wastewater effluents (10.2%

491 and 8.2%), which is congruent with the fact that the sampling points were strongly affected by

492 wastewater.

493 We also performed *in silico* PCR on RefSeq environmental *E. coli* genomes. Of the 303

494 genomes, 21 genomes (6.9%), one genome (0.3%), 12 genomes (4.0%), and eight genomes

495 (2.6%) were positive for W_nqrC, W_clsA_2, H8, and H12, respectively (**Table 2**, **Table S11**).

496 The prevalence of MST markers was lower than in the RefSeq wastewater *E. coli* genomes:

497 this is consistent with the fact that not all environmental *E. coli* originate from wastewater.

498 We note that the two markers developed in the present study, especially W_nqrC, are

499 commonly found on plasmids and could be subjected to horizontal gene transfer (HGT) in the

500 environment. Previous studies suggested that HGT could occur, for example, in biofilms,

501 although HGT by conjugation rarely occurs between motile planktonic cells (Abe et al., 2020).

502 Though we assume the contribution of HGT to the spread of the developed markers to be low

503 in environmental waters, where water is constantly flowing, this should be kept in mind.

504

505 **3.8. Concentrations of MST markers in wastewater samples.**

506 Concentrations of W_nqrC, W_clsA_2, H8, H12, and *uidA* in wastewater samples were

507 determined by qPCR experiments (**Table S12**). The amplification efficiencies of qPCR assays

508 were between 80%–100%, and the r^2 values ranged from 0.988–0.999. The concentrations of

509 markers and *uidA* in three primary effluent samples were: 7.13–7.24 $\log_{10}$ gene copies (GC)/L

510 for W_nqrC; 8.39–8.54 $\log_{10}$ GC/L for W_clsA_2; 6.26–6.47 $\log_{10}$ GC/L for H8; 6.03–6.41

511 $\log_{10}$ GC/L for H12; and 7.21–7.32 $\log_{10}$ GC/L for *uidA*. The concentrations in the three

512 biologically treated wastewater samples were: 4.43–4.76 $\log_{10}$ GC/L for W_nqrC; 6.12–6.18

513 $\log_{10}$ GC/L for W_clsA_2; 4.01–4.17 $\log_{10}$ GC/L for H8; 3.93–4.16 $\log_{10}$ GC/L for H12; and

514 4.50–4.80 $\log_{10}$ GC/L for *uidA*. Previously, *Bacteroides* HF183 and crAssphage markers were

515 suggested to be two of the most promising sewage-associated MST markers (Ahmed et al.,

516 2019a; Ahmed et al., 2019c). The mean concentrations of HF183 were previously reported to

517 be 9.23 $\log_{10}$ GC/L for raw sewage and 7.32 $\log_{10}$ GC/L for treated effluent (Ahmed et al.,

518 2016). The concentrations of crAssphage were previously reported to be 7.4–10.0 $\log_{10}$ GC/L

519 for untreated wastewater, and the mean concentration was reported to be 6.5 $\log_{10}$ GC/L for

520 treated effluent in high-income countries (Sabar et al., 2022). Although marker concentrations

521 in wastewater vary by region and treatment system, *E. coli* MST markers appear to be less

522 abundant in wastewater samples than markers such as HF183 and crAssphage.

523 The concentrations of W_nqrC were almost the same as *uidA* in all six samples, and the

524 concentrations of W_clsA_2 were more than ten times higher than *uidA* in all the samples. This

525 indicates that qPCR detected these marker genes not only from *E. coli* but also from non-*E.*

526 *coli* species. *In silico* PCR analysis using the nr database on the Primer-BLAST website, which

527 also includes the genomes of non-*E. coli* species, returned PCR products from non-*E. coli*

528 species for both W_nqrC (e.g., *Klebsiella pneumoniae* and *Salmonella enterica*) and W_clsA_2

529 (e.g., *Pseudomonas aeruginosa* and *K. pneumoniae*). In the case of W_nqrC, 517 (93%) of 556
530 products were from *E. coli*, and 39 (7%) were from other species. In the case of W_clsA_2,
531 only 29 (18%) of 157 products were from *E. coli*. Although detection of these marker genes in
532 non-*E. coli* species does not preclude the applicability of W_nqrC and W_clsA_2 to a culture-
533 independent method, further studies are needed to assess the applicability of these markers to
534 culture-independent testing. We note that we also observed H8 to be prevalent among non-*E.*
535 *coli* bacteria (Gomi et al., 2014), but it was proven in a subsequent study to be applicable in a
536 culture-independent manner (Senkbeil et al., 2019).
537
538 **3.9. Study limitations.**
539 This study has some limitations. First, when used in combination, the developed markers
540 showed quite high specificity but showed relatively low sensitivity, indicating that a relatively
541 large number of *E. coli* isolates would be needed in MST for each environmental sample. This
542 was due to the tradeoff between the sensitivity and specificity of genetic markers (**Figure 2**).
543 We assume this limitation to stem from using *E. coli* as an MST target organism rather than
544 the methodologies employed in our study. A previous study attempted to classify *E. coli*
545 genomes into different sources based on pan-genome information (Lupolova et al., 2017). The
546 study analyzed a pan-genome of *E. coli* from different sources (avian, bovine, canine,
547 environmental, human, and swine) and identified > 90,000 genes, three times more than the

548 total genes identified in the present study. Although *E. coli* genomes could in many cases be
549 classified into the correct sources, the classification was based on whole-genome sequence
550 information, making it not ideal for MST applications. They also found a less clear association
551 by host for *E. coli* when they compared the accessory genome trees of *E. coli* and another
552 bacterial species, *S. enterica*. Moreover, there were subsets of animal isolates sharing
553 significant genetic content with human isolates. The presence of the *E. coli* isolates that can
554 colonize multiple host species is also reported in other studies (Johnson and Clabots, 2006; Li
555 et al., 2019; Stenske et al., 2009). These factors indicate that *E. coli* is not the most suitable
556 organism for MST studies. This is reflected in the fact that the present study found no MST
557 markers with both high sensitivity and high specificity. However, we note that the sensitivity
558 calculated in this study is at the isolate level, and sample-level sensitivity (i.e., sensitivity
559 calculated for a wastewater sample, which can contain multiple *E. coli* strains) should be higher
560 (Senkbeil et al., 2019). Moreover, the development of a multiplex PCR assay for detecting
561 W_nqrC and W_clsA_2 in the same reaction will halve the time and labor needed to screen a
562 large number of isolates.
563 Second, an in-depth analysis of the sensitivity and specificity of the developed markers was
564 performed at the isolate/genome level but not in a culture-independent manner. Although we
565 determined the concentrations of the markers in wastewater samples by qPCR, further studies
566 are needed to assess the applicability (especially the specificity) of these markers in a culture-

567 independent manner. This can be done by performing qPCR on DNA directly extracted from
568 feces from different sources or by analyzing fecal metagenomes. In the latter case, one can
569 estimate the abundance of the markers in each metagenome, for example, by read mapping.
570 Analysis of metagenomes will remove the regional bias if metagenomes from geographically
571 diverse hosts are used. Although further studies are needed to be able to use these markers in a
572 culture-independent manner, we believe the two markers to be useful, since culture-dependent
573 methods are used to detect *E. coli* in routine water quality assessments, and there is a need to
574 identify the sources of the detected *E. coli* colonies. Moreover, cell suspensions prepared from
575 *E. coli* collected in routine water quality assessments can be used directly for PCR reactions
576 without DNA extraction. We also found it possible to use colored *E. coli* colonies on selective
577 media, in our case XM-G agar, for PCR reactions.
578
579
580 **4. Conclusions**
581 In the present study we identified two *E. coli* MST markers, namely W_nqrC and W_clsA_2,
582 which can be used for tracking wastewater contamination in environmental waters. These two
583 markers showed better performance than the previously developed H8 and H12 markers at the
584 isolate level, but users should keep in mind that the sensitivity remained low, at around 20%–
585 30%, when used in combination. Although further studies are needed to test the applicability

586 of the developed markers in a culture-independent manner, this study showed that the
587 developed MST markers could be useful for identifying the sources of *E. coli* (whether they
588 originated from domestic wastewater or not) at least at the isolate level. We believe that the
589 PCR assays developed in this study can be readily applied to identify the sources of *E. coli*
590 isolates collected during routine water quality assessments.
591
592
593 **Acknowledgements**
594 This work was supported by the Environment Research and Technology Development Fund
595 (JPMEERF20205006) of the Ministry of the Environment Japan, and Keihanshin Consortium
596 for Fostering the Next Generation of Global Leaders in Research (K-CONNEX), established
597 by Human Resource Development Program for Science and Technology, MEXT.
598 Computations were partially performed on the NIG supercomputer at ROIS National Institute
599 of Genetics. We thank those who provided animal fecal samples for this study.
600
601 **Supporting information**
602 Supplementary information associated with this article can be found, in the online version, at
603 _____.
604

605
606
607 **References**

608 Abe, K., Nomura, N. and Suzuki, S. 2020. Biofilms: hot spots of horizontal gene transfer (HGT) in aquatic
609 environments, with a focus on a new HGT mechanism. FEMS Microbiol Ecol 96(5).
610 Ahmed, W., Gyawali, P., Feng, S. and McLellan, S.L. 2019a. Host Specificity and Sensitivity of Established and
611 Novel Sewage-Associated Marker Genes in Human and Nonhuman Fecal Samples. Appl Environ
612 Microbiol 85(14).
613 Ahmed, W., Hughes, B. and Harwood, V.J. 2016. Current status of marker genes of *Bacteroides* and related
614 taxa for identifying sewage pollution in environmental waters. Water 8(6), 231.
615 Ahmed, W., O'Dea, C., Masters, N., Kuballa, A., Marinoni, O. and Katouli, M. 2019b. Marker genes of fecal
616 indicator bacteria and potential pathogens in animal feces in subtropical catchments. Sci Total Environ
617 656, 1427-1435.
618 Ahmed, W., Payyappat, S., Cassidy, M. and Besley, C. 2019c. A duplex PCR assay for the simultaneous
619 quantification of Bacteroides HF183 and crAssphage CPQ_056 marker genes in untreated sewage and
620 stormwater. Environ Int 126, 252-259.
621 Beghain, J., Bridier-Nahmias, A., Le Nagard, H., Denamur, E. and Clermont, O. 2018. ClermonTyping: an
622 easy-to-use and accurate in silico method for *Escherichia* genus strain phylotyping. Microb Genom
623 4(7).
624 Boehm, A.B., Graham, K.E. and Jennings, W.C. 2018. Can We Swim Yet? Systematic Review, Meta-Analysis,
625 and Risk Assessment of Aging Sewage in Surface Waters. Environ Sci Technol 52(17), 9634-9645.
626 Brynildsrud, O., Bohlin, J., Scheffer, L. and Eldholm, V. 2016. Rapid scoring of genes in microbial pan-
627 genome-wide association studies with Scoary. Genome Biol 17(1), 238.
628 Buchan, A., Alber, M. and Hodson, R.E. 2001. Strain-specific differentiation of environmental *Escherichia*
629 *coli* isolates via denaturing gradient gel electrophoresis (DGGE) analysis of the 16S-23S intergenic
630 spacer region. FEMS Microbiol Ecol 35(3), 313-321.
631 Carson, C.A., Shear, B.L., Ellersieck, M.R. and Asfaw, A. 2001. Identification of fecal *Escherichia coli* from
632 humans and animals by ribotyping. Appl Environ Microbiol 67(4), 1503-1507.
633 Castro-Hermida, J.A., Garcia-Presedo, I., Almeida, A., Gonzalez-Warleta, M., Correia Da Costa, J.M. and Mezo,
634 M. 2008. Contribution of treated wastewater to the contamination of recreational river areas with
635 *Cryptosporidium* spp. and *Giardia duodenalis*. Water Res 42(13), 3528-3538.
636 Chen, S., Zhou, Y., Chen, Y. and Gu, J. 2018. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics
637 34(17), i884-i890.
638 Cock, P.J., Antao, T., Chang, J.T., Chapman, B.A., Cox, C.J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F.,
639 Wilczynski, B. and de Hoon, M.J. 2009. Biopython: freely available Python tools for computational
640 molecular biology and bioinformatics. Bioinformatics 25(11), 1422-1423.
641 Constantinides, B., Chau, K.K., Quan, T.P., Rodger, G., Andersson, M.I., Jeffery, K., Lipworth, S., Gweon, H.S.,
642 Peniket, A., Pike, G., Millo, J., Byukusenge, M., Holdaway, M., Gibbons, C., Mathers, A.J., Crook, D.W.,

643 Peto, T.E.A., Walker, A.S. and Stoesser, N. 2020. Genomic surveillance of *Escherichia coli* and
644 *Klebsiella* spp. in hospital sink drains and patients. Microb Genom 6(7).
645 Di Cesare, A., Eckert, E.M., D'Urso, S., Bertoni, R., Gillan, D.C., Wattiez, R. and Corno, G. 2016. Co-
646 occurrence of integrase 1, antibiotic and heavy metal resistance genes in municipal wastewater
647 treatment plants. Water Res 94, 208-214.
648 Dolejska, M., Frolkova, P., Florek, M., Jamborova, I., Purgertova, M., Kutilova, I., Cizek, A., Guenther, S. and
649 Literak, I. 2011. CTX-M-15-producing *Escherichia coli* clone B2-O25b-ST131 and *Klebsiella* spp.
650 isolates in municipal wastewater treatment plant effluents. J Antimicrob Chemother 66(12), 2784-
651 2790.
652 Dombek, P.E., Johnson, L.K., Zimmerley, S.T. and Sadowsky, M.J. 2000. Use of repetitive DNA sequences and
653 the PCR To differentiate *Escherichia coli* isolates from human and animal sources. Appl Environ
654 Microbiol 66(6), 2572-2577.
655 Field, K.G. and Samadpour, M. 2007. Fecal source tracking, the indicator paradigm, and managing water
656 quality. Water Res 41(16), 3517-3538.
657 Gomi, R., Matsuda, T., Matsui, Y. and Yoneda, M. 2014. Fecal source tracking in water by next-generation
658 sequencing technologies using host-specific *Escherichia coli* genetic markers. Environ Sci Technol
659 48(16), 9616-9623.
660 Gomi, R., Matsuda, T., Matsumura, Y., Yamamoto, M., Tanaka, M., Ichiyama, S. and Yoneda, M. 2017a.
661 Occurrence of Clinically Important Lineages, Including the Sequence Type 131 C1-M27 Subclone,
662 among Extended-Spectrum-beta-Lactamase-Producing *Escherichia coli* in Wastewater. Antimicrob
663 Agents Chemother 61(9), e00564-00517.
664 Gomi, R., Matsuda, T., Matsumura, Y., Yamamoto, M., Tanaka, M., Ichiyama, S. and Yoneda, M. 2017b.
665 Whole-Genome Analysis of Antimicrobial-Resistant and Extraintestinal Pathogenic *Escherichia coli* in
666 River Water. Appl Environ Microbiol 83(5), e02703-02716.
667 Gonzalez-Fernandez, A., Symonds, E.M., Gallard-Gongora, J.F., Mull, B., Lukasik, J.O., Rivera Navarro, P., Badilla
668 Aguilar, A., Peraud, J., Brown, M.L., Mora Alvarado, D., Breitbart, M., Cairns, M.R. and Harwood, V.J.
669 2021. Relationships among microbial indicators of fecal pollution, microbial source tracking markers,
670 and pathogens in Costa Rican coastal waters. Water Res 188, 116507.
671 Harwood, V.J., Staley, C., Badgley, B.D., Borges, K. and Korajkic, A. 2014. Microbial source tracking markers
672 for detection of fecal contamination in environmental waters: relationships between pathogens and
673 human health outcomes. FEMS Microbiol Rev 38(1), 1-40.
674 Ishii, S. and Sadowsky, M.J. 2008. *Escherichia coli* in the Environment: Implications for Water Quality and
675 Human Health. Microbes Environ 23(2), 101-108.
676 Johnson, J.R. and Clabots, C. 2006. Sharing of virulent *Escherichia coli* clones among household members of
677 a woman with acute cystitis. Clin Infect Dis 43(10), e101-108.
678 Karkman, A., Do, T.T., Walsh, F. and Virta, M.P.J. 2018. Antibiotic-Resistance Genes in Waste Water. Trends
679 Microbiol 26(3), 220-228.
680 Khatib, L.A., Tsai, Y.L. and Olson, B.H. 2002. A biomarker for the identification of cattle fecal pollution in

681 water using the LTIIa toxin gene from enterotoxigenic *Escherichia coli*. Appl Microbiol Biotechnol 59(1),
682 97-104.
683 Khatib, L.A., Tsai, Y.L. and Olson, B.H. 2003. A biomarker for the identification of swine fecal pollution in
684 water, using the STII toxin gene from enterotoxigenic *Escherichia coli*. Appl Microbiol Biotechnol 63(2),
685 231-238.
686 Li, A.D., Li, L.G. and Zhang, T. 2015. Exploring antibiotic resistance genes and metal resistance genes in
687 plasmid metagenomes from wastewater treatment plants. Front Microbiol 6, 1025.
688 Li, J., Bi, Z., Ma, S., Chen, B., Cai, C., He, J., Schwarz, S., Sun, C., Zhou, Y., Yin, J., Hulth, A., Wang, Y., Shen, Z.,
689 Wang, S., Wu, C., Nilsson, L.E., Walsh, T.R., Borjesson, S., Shen, J., Sun, Q. and Wang, Y. 2019. Inter-
690 host Transmission of Carbapenemase-Producing *Escherichia coli* among Humans and Backyard
691 Animals. Environ Health Perspect 127(10), 107009.
692 Lindsey, R.L., Garcia-Toledo, L., Fasulo, D., Gladney, L.M. and Strockbine, N. 2017. Multiplex polymerase
693 chain reaction for identification of *Escherichia coli*, *Escherichia albertii* and *Escherichia fergusonii*. J
694 Microbiol Methods 140, 1-4.
695 Lupolova, N., Dallman, T.J., Holden, N.J. and Gally, D.L. 2017. Patchy promiscuity: machine learning applied
696 to predict the host specificity of *Salmonella enterica* and *Escherichia coli*. Microb Genom 3(10),
697 e000135.
698 Ma, C.-Y., Ihara, M., Liu, S., Sugie, Y. and Tanaka, H. 2022. Tracking the source of antibiotic-resistant
699 *Escherichia coli* in the aquatic environment in Shiga, Japan, through whole-genome sequencing.
700 Environmental Advances 8, 100185.
701 Maheux, A.F., Picard, F.J., Boissinot, M., Bissonnette, L., Paradis, S. and Bergeron, M.G. 2009. Analytical
702 comparison of nine PCR primer sets designed to detect the presence of *Escherichia coli*/*Shigella* in
703 water samples. Water Res 43(12), 3019-3028.
704 Mathai, P.P., Staley, C. and Sadowsky, M.J. 2020. Sequence-enabled community-based microbial source
705 tracking in surface waters using machine learning classification: A review. J Microbiol Methods 177,
706 106050.
707 Naidoo, S. and Olaniran, A.O. 2013. Treated wastewater effluent as a source of microbial pollution of
708 surface water resources. Int J Environ Res Public Health 11(1), 249-270.
709 Nicolas-Chanoine, M.H., Bertrand, X. and Madec, J.Y. 2014. *Escherichia coli* ST131, an intriguing clonal
710 group. Clin Microbiol Rev 27(3), 543-574.
711 Nopprapun, P., Boontanon, S.K., Harada, H., Surinkul, N. and Fujii, S. 2020. Evaluation of a human-
712 associated genetic marker for *Escherichia coli* (H8) for fecal source tracking in Thailand. Water Sci
713 Technol 82(12), 2929-2936.
714 Page, A.J., Cummins, C.A., Hunt, M., Wong, V.K., Reuter, S., Holden, M.T., Fookes, M., Falush, D., Keane, J.A. and
715 Parkhill, J. 2015. Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics 31(22),
716 3691-3693.
717 Riley, L.W. 2014. Pandemic lineages of extraintestinal pathogenic *Escherichia coli*. Clin Microbiol Infect
718 20(5), 380-390.

719 Sabar, M.A., Honda, R. and Haramoto, E. 2022. CrAssphage as an indicator of human-fecal contamination in
720 water environment and virus reduction in wastewater treatment. Water Res 221, 118827.
721 Scott, T.M., Rose, J.B., Jenkins, T.M., Farrah, S.R. and Lukasik, J. 2002. Microbial source tracking: current
722 methodology and future directions. Appl Environ Microbiol 68(12), 5796-5803.
723 Seemann, T. 2014. Prokka: rapid prokaryotic genome annotation. Bioinformatics 30(14), 2068-2069.
724 Senkbeil, J.K., Ahmed, W., Conrad, J. and Harwood, V.J. 2019. Use of *Escherichia coli* genes associated with
725 human sewage to track fecal contamination source in subtropical waters. Sci Total Environ 686, 1069-
726 1075.
727 Sghaier, S., Abbassi, M.S., Pascual, A., Serrano, L., Diaz-De-Alba, P., Said, M.B., Hassen, B., Ibrahim, C., Hassen,
728 A. and Lopez-Cerero, L. 2019. Extended-spectrum beta-lactamase-producing Enterobacteriaceae
729 from animal origin and wastewater in Tunisia: first detection of O25b-B23-CTX-M-27-ST131
730 *Escherichia coli* and CTX-M-15/OXA-204-producing *Citrobacter freundii* from wastewater. J Glob
731 Antimicrob Resist 17, 189-194.
732 Steele, M. and Odumeru, J. 2004. Irrigation water as source of foodborne pathogens on fruit and
733 vegetables. J Food Prot 67(12), 2839-2849.
734 Stenske, K.A., Bemis, D.A., Gillespie, B.E., D'Souza, D.H., Oliver, S.P., Draughon, F.A., Matteson, K.J. and Bartges,
735 J.W. 2009. Comparison of clonal relatedness and antimicrobial susceptibility of fecal *Escherichia*
736 *coli* from healthy dogs and their owners. Am J Vet Res 70(9), 1108-1116.
737 Stoesser, N., Sheppard, A.E., Pankhurst, L., De Maio, N., Moore, C.E., Sebra, R., Turner, P., Anson, L.W., Kasarskis,
738 A., Batty, E.M., Kos, V., Wilson, D.J., Phetsouvanh, R., Wyllie, D., Sokurenko, E., Manges, A.R., Johnson,
739 T.J., Price, L.B., Peto, T.E., Johnson, J.R., Didelot, X., Walker, A.S., Crook, D.W. and Modernizing Medical
740 Microbiology Informatics, G. 2016. Evolutionary History of the Global Emergence of the
741 *Escherichia coli* Epidemic Clone ST131. mBio 7(2), e02162.
742 Teixeira, P., Dias, D., Costa, S., Brown, B., Silva, S. and Valerio, E. 2020. *Bacteroides* spp. and traditional fecal
743 indicator bacteria in water quality assessment - An integrated approach for hydric resources
744 management in urban centers. J Environ Manage 271, 110989.
745 Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B.C., Remm, M. and Rozen, S.G. 2012. Primer3-
746 -new capabilities and interfaces. Nucleic Acids Res 40(15), e115.
747 Warish, A., Triplett, C., Gomi, R., Gyawali, P., Hodgers, L. and Toze, S. 2015. Assessment of Genetic Markers
748 for Tracking the Sources of Human Wastewater Associated *Escherichia coli* in Environmental Waters.
749 Environ Sci Technol 49(15), 9341-9346.
750 Wick, R.R., Judd, L.M., Gorrie, C.L. and Holt, K.E. 2017. Unicycler: Resolving bacterial genome assemblies
751 from short and long sequencing reads. PLoS Comput Biol 13(6), e1005595.
752 Ye, J., Coulouris, G., Zaretskaya, I., Cutcutache, I., Rozen, S. and Madden, T.L. 2012. Primer-BLAST: a tool to
753 design target-specific primers for polymerase chain reaction. BMC Bioinformatics 13, 134.
754 Yu, D., Banting, G. and Neumann, N.F. 2021. A review of the taxonomy, genetics, and biology of the genus
755 *Escherichia* and the type species *Escherichia coli*. Can J Microbiol 67(8), 553-571.
756 Zhi, S., Banting, G., Stothard, P., Ashbolt, N.J., Checkley, S., Meyer, K., Otto, S. and Neumann, N.F. 2019.

757 Evidence for the evolution, clonal expansion and global dissemination of water treatment-resistant
758 naturalized strains of *Escherichia coli* in wastewater. Water Res 156, 208-222.
759 Zhi, S., Stothard, P., Banting, G., Scott, C., Huntley, K., Ryu, K., Otto, S., Ashbolt, N., Checkley, S., Dong, T.,
760 Ruecker, N.J. and Neumann, N.F. 2020. Characterization of water treatment-resistant and
761 multidrug-resistant urinary pathogenic *Escherichia coli* in treated wastewater. Water Res 182, 115827.
762 Zhou, Z., Alikhan, N.F., Mohamed, K., Fan, Y., Agama Study, G. and Achtman, M. 2020. The EnteroBase
763 user's guide, with case studies on *Salmonella* transmissions, *Yersinia pestis* phylogeny, and *Escherichia*
764 core genomic diversity. Genome Res 30(1), 138-152.
765
766
767