Evaluation of genetic diversity among aquatic and fecal isolates of *Escherichia coli* using multilocus variable number of tandem repeat analysis

Neha Khare¹ · Megha Kaushik¹ · Sanjay Kumar¹ · Pooja Gulati¹

Received: 8 August 2019 / Accepted: 10 January 2020 / Published online: 23 January 2020
© King Abdulaziz City for Science and Technology 2020

Abstract

In developing countries like India, fecal pollution of surface waters is a major threat to public and environmental health. The aim of the study was to assess serological, phylogenetic and molecular diversity among aquatic *Escherichia coli* isolates from Yamuna river and their comparison with the animal fecal isolates. A total of 97 *E. coli* isolates from Yamuna river and domesticated animals were characterized by multilocus variable number tandem repeat analysis (MLVA) using four VNTR loci. The pathogenicity of these strains by serological and phylogenetic analysis was also determined. *E. coli* strains were differentiated into 53 distinct MLVA types with high discriminatory power, Simpson's index of 0.95 (95% CI 0.923–0.978). Cluster analysis and population modeling using minimum spanning tree suggested a possible epidemiological linkage among aquatic and fecal isolates. The study also reported the presence of highly diverse and pathogenic serotypes belonging to STEC and EPEC strains, particularly O157 and high prevalence of pathogenic phylogroups (phylogroup, B2 and D). The presence of such a high molecular heterogeneity among aquatic and fecal *E. coli* isolates emphasizes upon the need to develop proper fecal pollution abatement strategies for Indian natural bodies.

Keywords *E. coli* · Serotype · Phylogroup · VNTR · MLVA · Clonal relationship

Introduction

Surface water ecosystem like rivers, lakes, and ponds are important natural resources exploited for household, irrigation and recreational purposes. Contamination of these surface waters with municipal sewage waste, effluents from the wastewater treatment plant, storm sewer drainage, domestic animals, agricultural or urban effluents has extensively increased the microbial load of these aquatic environments (Sasakova et al. 2018). Fecal coliforms like *E. coli* isolated from the natural water bodies have immense genetic diversity allowing for the better species adaptability and enhanced resistance to environmental changes (Berthe et al. 2013). Routine microbiological monitoring fails to provide information on the heterogeneity of fecal *E. coli*.

Several typing methods have been employed for effective differentiation of related bacterial species in epidemiological surveillance. Conventional typing methods such as serotyping, biochemical typing and antibiogram-based typing of bacterial species have been used extensively (van Belkum 2007). Traditionally, *E. coli* are serotyped on the basis of their O (lipopolysaccharide), H (flagellar) and K (capsular) surface antigen profiles. With more than 180 different O-serotypes, many of these serotypes are associated with disease causing *E. coli* pathotypes, including Shiga toxin-producing *Escherichia coli* (STEC) [eg. Enterohemorrhagic *E. coli* (EHEC)], Enteropathogenic *E. coli* (EPEC) and Uropathogenic *E. coli* (UPEC) (Croxen et al. 2013). In addition, phylogenetic analysis of *E. coli* is determined using the combination of three virulence gene markers *chuA*, *yjaA*, and *TspE4.C2*. The presence or absence of these genes differentiate the isolates into four main phylogroups: A [*chuA* (–), *yjaA* (–/+), *TSPE4.C2* (–)]; B1 [*chuA* (–), *yjaA* (–), *TSPE4.C2* (+)]; B2 [*chuA* (+), *yjaA* (+), *TSPE4.C2* (–/+)] and D [*chuA* (+), *yjaA* (–), *TSPE4.C2* (+)] (Clermont et al. 2000). However, these phenotypic methods should be validated with genotypic methods for efficient bacterial typing.

✉ Pooja Gulati
gulatipooja1@gmail.com; pooja.micro@mdurohtak.ac.in

¹ Medical Microbiology and Bioprocess Technology Laboratory, Department of Microbiology, Maharshi Dayanand University, Rohtak, Haryana, India

Molecular methods like pulse field gel electrophoresis (PFGE), ribotyping, repetitive element-based PCR analysis (rep-PCR), PCR-RFLP, MLVA and newer techniques like matrix-assisted laser desorption ionization time-of-flight (MALDI-TOF) mass spectrometry (MS) have been used successfully for typing of various bacterial species (Shangkuan et al. 2000; Mohapatra et al. 2007; Murray 2010). PFGE has been considered as a 'gold standard' for bacterial subtyping, however, the technique is laborious and unable to process a large sampling size (Kawamori et al. 2008). Recently, multilocus variable-number tandem repeat analysis (MLVA) has also emerged as an efficient epidemiological typing tool and considered as a potential next-generation typing system by CDC (Cooley et al. 2007). MLVA is a PCR-based method which analyses polymorphism at several variable numbers of tandem repeats (VNTR) containing loci (van Belkum et al. 1998). This method is fast, inexpensive and does not require any specialized instrument. Moreover the results obtained are easy to interpret and allows unambiguous inter-laboratory comparison (Dahyot et al. 2018). However, the need to design new primers and optimization of protocols for every species is its major disadvantage (<https://www.cdc.gov/pulsenet/pathogens/mlva.html>). MLVA typing has been used extensively in differentiating bacterial strains of *Yersinia*, *Salmonella*, *Mycobacterium* and *E. coli* (Cowan et al. 2002; Lindstedt et al. 2003; Noller et al. 2003; Gulati et al. 2009). However, there is a paucity of data on the use of such technique in developing countries like India especially in bacterial isolates from surface water (Hagedorn et al. 2011).

The River Yamuna serves as a principal source of water to the metropolitan city of National Capital Territory of India. Heavy loads of fecal contamination have been routinely detected in it (Jamwal et al. 2011). Contamination through various point and non-point sources, particularly partially treated and untreated wastewater discharge, increases the microbial load of the river tremendously, thereby representing a reservoir of diverse fecal coliforms like *E. coli* (Finley et al. 2013). It is therefore imperative to identify the predominant serotypes and phylogroups among *E. coli* strains of New Delhi. Unraveling the heterogeneity among the strains of *E. coli* will help to understand their evolution and behavior in natural water bodies. Moreover, analyzing the determinants of genetic diversity can also provide insights into the relationship between aquatic coliforms and its originating sources and thereby helping in the development of effective fecal pollution abatement strategies.

Thus, the present study was undertaken to analyze the diversity of several strains of *E. coli* isolated from Yamuna river water and animal origin from New Delhi using multilocus VNTR analysis (MLVA). The distribution of different serotypes and phylogroups among various animal and water-borne *E. coli* isolates was also studied. The efficacy of MLVA as an important molecular tool for epidemiological

investigations of *E. coli* in natural environments was also discerned.

Material and methods

E. coli isolates used in the study

A total of 97 *E. coli* isolates of water and animal origin were used. The study included 81 *E. coli* strains isolated earlier (October 2012 to June 2013) from Yamuna River, New Delhi from eight different sampling sites (Palla, Wazirabad, ISBT, ITO, Nizamuddin Bridge, DND Flyway, Okhla Barrage, Najafgarh Drain) (Kaushik et al. 2019). The fecal samples of domestic animals were collected (October 2013 to December 2013) from the rural area of New Delhi around the Yamuna river. Briefly, 0.5 g of each animal feces was directly inoculated in 50 mL of tryptone phosphate broth (TPB) and MacConkey broth with an overnight incubation at 37 °C under shaking conditions. The overnight grown broth culture was further streaked on selective and differential media viz; eosin methylene blue (EMB) agar and MacConkey agar. Initially, 27 presumptive *E. coli* were identified based on the appearance of metallic green sheen colonies on EMB agar and/or dark pink colonies on MacConkey agar. The cultures were confirmed by standard biochemical IMViC test series and later based on the carbon utilization capability of *E. coli* isolates using KB010 Hi *E. coli*TM Identification Kit from HiMedia. Sixteen fecal isolates were confirmed as *E. coli* and preserved on nutrient agar slants at 4 °C. The pure cultures and the glycerol stocks (50% v/v) were maintained and preserved at −20 °C. The purity of preserved isolates was also checked periodically on EMB agar.

Serotyping of *E. coli* isolates

The *E. coli* isolates were confirmed and serotyped by slide agglutination tests with antiserum following conventional method. The isolates were serotyped on the basis of heat-stable O surface antigen by the National *Salmonella* and *Escherichia* Centre at the Central Research Institute at Kasauli, Himachal Pradesh, India.

Genomic DNA preparation

The genomic DNA was extracted from the overnight grown broth culture for each isolate using Wizard[®] Genomic DNA Purification Kit from Promega according to the manufacturer's instructions. Briefly, each isolate was grown at 37 °C in nutrient broth. One millilitre of the overnight grown culture was harvested at 8000 rpm for 10 min and the whole genomic DNA was isolated as per the manufacturer's guidelines. The purity of the DNA was checked

spectrophotometrically. The extracted genomic DNA was also quantitated spectrophotometrically and stored at -4°C .

Phylogrouping of isolates

Phylogrouping analysis of *E. coli* isolates was performed by a triplex PCR targeting virulence genes *chuA*, *yjaA*, and *TspE4C2*. The presence or absence of these genes classified the isolates as extraintestinal pathogenic (belonging to B2 and D phylogroup) and commensal strains (belonging to phylogroup A and B1) (Clermont et al. 2000). The primers and PCR conditions used were same as described previously (Kaushik et al. 2019).

VNTR loci and PCR conditions

Four VNTR loci namely CVN001, CVN002, CVN003 and CVN015 described previously by Lindstedt et al. (2007) were selected. The VNTR loci with a unit length ≥ 8 , tandem repeats of ≥ 4 and exhibiting $> 85\%$ homology were selected for MLVA analysis. The primer pairs used and detailed information about each VNTR locus is presented in Table 1. The VNTR loci were amplified each in uniplex PCR. 25 μL reaction mixture consisted of bacterial DNA as a template, 100 ng; each primer (25 pmol) (Sigma Aldrich, India) (Table 1), 1 μL ; 10 $\times$ PCR buffer, 2.5 μL ; 10 mM each dNTPs mix, 0.5 μL and 2 U *Taq* DNA Polymerase (ThermoFisher, US). The PCR amplification was performed in 48 well MJ-MiniTM Thermal Cycler (Bio-Rad). The PCR conditions for all the VNTR loci (CVN001, CVN002, CVN003, and CVN015) were as follows: pre-denaturation at 95°C for 15 min, followed by 30 cycles of denaturation at 94°C for 30 s, annealing (refer to Table 1 for annealing temperature) for 90 s and an extension at 72°C for 90 s and a final extension at 72°C for 10 min. The amplified PCR products were resolved on 3% agarose gel in 1X Tris-acetate EDTA buffer at 80 V. 100 bp DNA Ladder (HiMedia, Mumbai) was used as a size marker to normalize the agarose gel. The gels were visualized under gel documentation system and images were stored in TIFF format.

Sequencing of the VNTR amplicons

DNA sequencing of PCR amplicons from representative isolates (one for each VNTR loci) was performed. The amplified PCR products were purified using Thermo Scientific GeneJet Gel Extraction Kit. The purified PCR products were sequenced using Genetic Analyzer ABI 3730DA by the dye termination method. The sequences thus obtained were aligned and the number of repeats in each isolate was determined. The sequences were submitted to GenBank under the accession numbers MN895235 (CVN001 locus),

Table 1 Characteristics of VNTR containing loci, amplification conditions and the number of alleles obtained among *E. coli* strains

S. no.	Locus	Position in <i>E. coli</i> K-12	Array length ^a in <i>E. coli</i> K-12	Primer	Sequence ^b (5'–3')	Annealing temp. ($^{\circ}\text{C}$)	Expected amplicon size ^c (bp)	No. of alleles	No. of repeats		Nei's diversity index (D_N)
									Min	Max	
1	CVN001	934,551–934,976	234	VN01-F VN01-R	AACCGGCTGGGGCGAATCC GGCGCGGTGTCAGCAATC	63.6	445	9	4	24	0.61
2	CVN002	4,473,452–4,473,919	216	VN02-F VN02-R	AACCGTTATGAAGTCT TCGCCAGTAAGTATGAAATC	56	488	4	6	22	0.59
3	CVN003	58,780–59,235	90	VN03-F VN03-R	AAAAATCCGGATGAGWTGGTC TTGCGTTGTCAGTAATTTGTTTCAG	65	479	6	3	17	0.49
4	CVN015	143–341	36	VN15-F VN15-R	TAGGCATAGCGCACAGACAGATAA GTACCGCCGAACTTCAACACTC	63.6	220	5	2	6	0.55

^aRepeat length * copy number

^bThe letter in bold and italics denotes the degenerate base with IUPAC nomenclature

^cPCR amplicon based on *E. coli* K-12

MN901986 (CVN002 locus), MN895234 (CVN003 locus) and MN890135 (CVN015 locus).

Phylogenetic and statistical analyses

The gels were analyzed using Quantity One software (Bio-Rad) and the exact band position of the amplicons was determined. The numbers of repeats were also calculated using tandem repeat calculator of this program. For each locus, different numbers of tandem repeats were assigned a separate allele number (Table 2). When no amplicon was observed, the allele number was assigned as '0'. Thus, each *E. coli* isolate was identified by a string of four allele numbers in the order CVN001–CVN002–CVN003–CVN015. These allele number strings were imported into Bionumerics software version 7.6 (Applied Maths, Belgium) as character type data and a dendrogram was generated using the categorical coefficient of similarity and the Ward algorithm.

Phylogenic modeling of the strains was illustrated by the construction of a minimum spanning tree (MST) based on categorical coefficient and priority rules set at first maximum number of single-locus variants and then a maximal number of single- and double-locus variants. Two different MSTs were created showing the association of different MLVA types with serotypes (Fig. 2a) and phylogenetic groups (Fig. 2b).

Diversity index

The diversity index based on Nei's diversity (D_N) was calculated for each locus using the formula $D_N = 1 - \sum(f)^2$, where f is the allele frequency (Weir 1990). The discriminatory power of the MLVA typing method was calculated by Simpson's index of diversity (DI) (Hunter and Gaston 1988) using the formula $DI = 1 - [\sum n(n-1)/N(N-1)]$ where n is the number of isolates belonging to a single type and N is the total number of isolates. 95% confidence interval (CI)

was also calculated as described previously (Grundmann et al. 2001).

Results

E. coli isolates used in the study

The isolates used in the present study comprised of 81 *E. coli* isolated earlier from Yamuna river, New Delhi (Kaushik et al. 2019). In addition, 16 fecal *E. coli* strains from domesticated animals viz. goat (six isolates), cow (five isolates), sheep (two isolates) and one each from buffalo, horse, and pig were also isolated and included in the study.

Characterization of *E. coli* isolates

Both the aquatic and fecal *E. coli* isolates showed a remarkable diversity of serotypes. Approximately 60% of the 97 isolates were assigned an O type while 25 isolates remained untypable and 8 were auto-agglutinable. 23 O types were identified, with O126 and O88 serotypes being the most common (eight each; 8%), followed by O9 (5%), O7 and O22 (four each; 4%) (Fig. 1). Serotypes O157, O113 and O91, regarded as STEC, were also identified in aquatic *E. coli* and considered as potentially pathogenic strains.

The aquatic and animal *E. coli* isolates were differentiated into four phylogroups using triplex PCR. The frequency of different phylogroups in the present study was the following: B2 (38%, 37/97) followed by A (33%, 30/97), B1 (6%, 6/97) and D (2%, 2/97) (Fig. 1). About 21% of the isolates remained unassigned. Although, a comparable proportion of virulent (40%) and non-virulent (37%) phylogroups was observed in the present study, however, the study reports the predominance of virulent phylogroups (B2 and D) among these isolates. The incidence of virulent phylogroup B2 (38%) was observed only among water-borne *E. coli* isolates while phylogroup D (2%) was found only in *E. coli* isolates of animal origin.

Comparison of VNTR alleles among strains of *E. coli*

All the four VNTR loci studied were found polymorphic and the variations are listed in Table 1. The analysis of CVN001 locus in 97 *E. coli* strains revealed nine different alleles with 4–24 repeats. The allele with a copy number of 8 was the most predominant. In some of the isolates, multiple bands differing in tandem repeats were obtained for this locus. The multiple amplicons which were highly reproducible and distinct were considered for the study. Four isolates carried two amplicons differing in size by single repeat size ($n=6$ and $n=7$) while one isolate carried fragments with 7 and 8

Table 2 Conversion of copy number of tandem repeats into allele number

Allele number	Number of tandem repeats			
	CVN001	CVN002	CVN003	CVN015
1	4	6	3	2
2	7	8	6	3
3	8	22	7	4
4	9	6, 8	6, 7	5
5	6, 7		3, 6, 7	6
6	7, 8		3, 6, 7, 17	
7	4, 5, 6			
8	4, 5, 8, 17			
9	6, 7, 8, 17, 18, 24

Fig. 1 Dendrogram based on MLVA profiles of *E. coli* strains. The clustering between MLVA genotypes was performed by using the categorical coefficient and Ward's method. The allele string represents the allele numbers obtained at four loci in the order CVN001–CVN002–CVN003–CVN0015. ND not determined, AA auto agglu-

tinable, UT untypeable. ^aRefer to Kaushik et al. (2019) where * represents the presence of variable region of that particular integron only and ** depicts the presence of both integron along with its variable region

Fig. 2 Population modeling using the minimum spanning tree (MST) for multilocus variable number tandem repeat analysis of *E. coli* strains using Bionumerics v7.6. The MST was constructed based on different MLVA allele strings using the single locus variants as the priority rule where each circle represents different MLVA type with the size proportional to the number of isolates within that type. Thick dark lines connecting two circles (MLVA type) indicates difference in a single VNTR locus (denoted by 1 on each line) while thin dotted line indicates difference by two VNTR loci (denoted by 2 on each line). *E. coli* isolated from different fecal and water sam-

ples are depicted by different colors. Green, red, navy blue, mustard, turquoise, teal and brown indicating *E. coli* isolated from water, goat, cow, sheep, pig, buffalo and horse, respectively. **a** Association between the MLVA types and different serotypes is depicted. Letter in the circle denotes serotypes. Pink zone surrounds *E. coli* strains of fecal origin. **b** Association between the MLVA types and different phylogroups is depicted. Pink zone surrounds *E. coli* strains belonging to pathogenic phylogroups. Letter in the circle denotes phylogenetic groups

repeats. Three isolates were reported to carry more than two fragments (varying in their repeat numbers).

The locus CVN002 produced only four different alleles (Table 2). Majority of the isolates (57/97; 59%) carried the allele with eight repeats. A single isolate generated two amplicons varying in size by two repeats ($n=6$, $n=8$). However, 20 isolates failed to produce an amplicon for this locus. All the isolates possessing allele with 22 repeats belonged to the same sampling region. Both the *E. coli* strains belonging to O157 serotype were untypeable using this locus.

PCR amplification targeting the locus CVN003 resulted in typing of only 30% (29 of 97) of the isolates while the rest were untypeable. This locus was not amplified in any of the isolates from two sampling sites (Palla region and Najafgarh drain of Yamuna river) and was thus designated as 'null allele'. For CVN003 locus, a total of six alleles were produced with repeats ranging from 3 to 17. Most of the isolates produced single amplicon with 3 or 6 or 7 repeats. Some of the isolates were found to carry more than one CVN003 loci, with different combinations of alleles with 3, 6 and 7 repeats. One isolate showing four bands with an additional locus with 17 repeats was also reported.

VNTR locus CVN015 showed the presence of five alleles with 63% of the isolates possessing an allele with 3 repeats. 11% of isolates carried allele with 2 and 3 repeats each while 7% of the isolates carried 5 and 6 repeats. None of the isolates produced multiple bands for this locus.

Nei's diversity (D_N) calculated for each locus was determined as the potential marker to discriminate the diversity among *E. coli* strains. The value for D_N ranged between 0.49 and 0.61. CVN001 locus was identified as the most polymorphic locus ($D_N=0.61$) while locus CVN003 had the lowest diversity index ($D_N=0.49$) (Table 1).

Reproducibility of MLVA profile

To access reproducibility and stability of MLVA genotypes, five *E. coli* isolates were subcultured in trypticase soy broth and VNTR PCR amplification was performed (result not shown). Indistinguishable MLVA genotypes were generated for CVN001, CVN002, CVN003 and CVN015 locus up to five subcultures, suggesting stability of VNTR typing when performed with *E. coli* subjected to repeated subculturing.

Sequences submitted at GenBank

On sequencing, the obtained sequences were compared with the available genome sequences of reference strain *E. coli* K12 (U00096) to verify the number of repeats. The sequence for CVN001 locus was found to carry 7 repeats and has been submitted under the accession number MN895235. The sequence for CVN002 locus was carrying 8 repeats and has been submitted under the accession number MN901986. Accession numbers for sequences of CVN003 and CVN015 locus each carrying 3 repeats are MN895234 and MN890135, respectively.

MLVA-based cluster analysis

Each isolate has been characterized by a distinct MLVA allele string where each new multiple of repeat was assigned a distinct allele number. The resulting allele numbers were imported into BioNumerics v7.6 as character values for analysis and dendrogram was constructed. The allele string containing four numbers define a bacterial isolate.

A total of 53 MLVA genotypes were observed among 97 *E. coli* isolates where 37 of them were unique to a single isolate. MLVA clustered *E. coli* isolates into two major groups—A and B (Fig. 1). Although both clusters A and B carried almost the same number of genotypes, Group B carried majority (14/16) of the *E. coli* isolates of animal origin while two animal isolates are clustered in group A. No co-relation in context to serotyping could be discerned as similar serotypes were distributed in both the major clusters. However, it was noticed that isolates with Shiga toxin-producing *E. coli* (STEC) serotype O157 were clustered in the same sub-cluster A1 (IT 16 and NJ 7) while the other non-O157 STEC were distributed in both the major clusters. The isolates exhibiting 100% similarity in MLVA patterns were predominantly observed in group B where some of the *E. coli* isolates of animal origin generated the same MLVA patterns as the water isolates (Fig. 1). The MLVA pattern 3-2-0-2 was the most predominant type which comprised of 18 *E. coli* isolates including 3 animal isolates viz. cow, buffalo, and goat. The second most common MLVA pattern (2-2-0-2), differs from predominant MLVA type in copy number of tandem repeats at a single locus, CVN001. This MLVA pattern included 11 *E. coli* isolates where 2 animal isolates sheep and cow shared the same pattern with other aquatic *E. coli* isolates. Simpson's index of the MLVA typing method at four VNTR loci for *E. coli* isolates was 0.95 (95% confidence interval 0.923–0.978).

Population modelling using minimum spanning tree (MLVA and serotypes) depicted that MLVA genotypes were shared between fecal and aquatic *E. coli* belonging to diverse serotypes (Fig. 2a). Serotype O157 created a separate branch from other serotypes. MST between MLVA and phylogroups

depicted that the central branch had a core of virulent phylogroups (B2 and D), and the proportion of other commensal phylogenetic groups (A and B1) increases in an outward direction from them (Fig. 2b). It was also observed that the central core of MST was formed by B2 phylogroup while phylogroup D was present only at the periphery.

Discussion

The present study investigated the prevalence of *E. coli* in urban water of a major river, the Yamuna crossing the national capital, New Delhi and compared them with animal fecal isolates. The isolates were characterized by serotyping and phylogrouping to assess variation in their antigenic determinants and virulence factors. Multilocus variable number tandem repeat analysis using four different VNTR loci was used to characterize these *E. coli* isolates.

Serotyping of *E. coli* isolates depicted the presence of diverse serotypes in both aquatic and animal *E. coli*. In the current study, O serotypes could not be determined for 25 *E. coli* isolates, which may represent emergence of novel serotypes (Bai et al. 2016). However, maximum diversity of serotypes was found in aquatic *E. coli* isolated from industrial regions of ISBT, ITO and Okhla Barrage. Several isolates belonging to pathogenic serotypes like O157 and O91 (STEC), O22 (UPEC), O113 (EHEC), O88, O126, and O9 were identified. Serotypes O91 identified in the present study have been recognized elsewhere as non-O157 STEC linked to severe human diseases (Bettelheim 2007). Studies from other aquatic ecosystem like River Ganga, India have also reported high incidence of *E. coli* exhibiting a diverse range of pathogenic strains viz. Shiga-toxin producing *E. coli* (STEC) and Enterotoxigenic *E. coli* (ETEC) (Hamner et al. 2007; Ram et al. 2007). Another Indian waterbody, Cochin estuary, has documented the prevalence of *E. coli* belonging to STEC, UPEC, EHEC and suggested a plausible release of these pathogenic strains from hospital wastes (Chandran et al. 2008). Thus the natural river systems of developing countries act as a reservoir as well as route of transmission of fecal coliforms, notably *E. coli*.

The study also assessed phylogroup distribution among *E. coli* isolates which showed that almost 40% of all the *E. coli* isolates belonged to the virulent phylogroup B2, which is generally associated with extraintestinal infections (Micenková et al. 2016). Our results were in accordance with a study from surface waters of Detroit River (bordering Canada and United States) which also reported high prevalence of phylogroup B2 (Hamelin et al. 2007). However predominance of commensal phylogroups (A and B1) is generally reported from aquatic environments across the globe (Walk et al. 2007; Luna et al. 2010; Tymensen et al. 2015; Bajaj et al. 2015). The high frequency of virulent phylogroup B2

in the present study could be anticipated due to the discharge of municipal sewage water into the river. Majority of the animal fecal *E. coli* isolates belonged to commensal phylogroups A and B1. This result is inconsistent with previous studies on animal isolates in other parts of the world (Higgins et al. 2007; Carlos et al. 2010).

MLVA is a simple and rapid method, commonly used for monitoring outbreaks and clonal spread of bacterial isolates. Due to its cost-effectiveness and good reproducibility, it has been successfully employed to discern the clonal relationship amongst *E. coli* isolates (McLellan 2004; Gorgé et al. 2008). These studies have highlighted the fact that a high amount of diversity exists between environmental *E. coli* isolates, thus, accurate strain characterization is important in bacteria isolated from surface waters to elucidate fecal pollution. In the current study, VNTR profiling of individual locus was performed by PCR amplification, which resulted in an array of MLVA genotype characteristic of each isolate. Among the four loci studied, CVN001 was the most variable locus while CVN003 was least diverse. In the present study, only 30% of the isolates yielded allele corresponding to CVN003 locus, whereas the majority of isolates (70%) showed 'null allele'. Earlier studies have also reported a high incidence of 'null allele' for CVN003 in *E. coli* isolated from diverse origin (Løbersli et al. 2012; Krüger et al. 2015). These results suggest that either the corresponding locus may not be present or sequence divergence resulted in mispriming (Onteniente et al. 2003; Camélina et al. 2019). Few isolates were characterized with more than one band for a specific VNTR locus. Double or multiple fragments produced may be due to an error during PCR amplification or due to the presence of multiple numbers of the allele of that particular sequence (Gulati et al. 2009; Lazzarini et al. 2012).

MLVA analysis of the isolates yielded 53 distinct genotypes based on the allele string generated by four VNTR loci. The most common MLVA profile, 3-2-0-2, was shared by 14 aquatic *E. coli* and 4 animal isolates. The second frequent profile (2-2-0-2) was found in nine aquatic *E. coli* and two animal isolates. The presence of animal isolates with 100% identical MLVA pattern to *E. coli* isolates of Yamuna river indicates clonal relatedness. Among the animal isolates, *E. coli* isolates from cow, goat, pig, and sheep showed similar MLVA patterns as aquatic *E. coli* suggesting that these domesticated animals might act as a major source of fecal pollution to surface waters.

Escherichia coli isolates used in the present study were previously assessed for the presence of different classes of integrons (Kaushik et al. 2019). Integrons are mobile genetic elements responsible for the dissemination of antibiotic-resistant determinants via discrete gene cassettes and are differentiated into four classes namely class 1, class 2, class 3 and class 4 based on variations in integrase gene

(Stalder et al. 2012). The MLVA pattern observed in the present study was also compared to the presence of integrons. MLVA-based cluster analysis differentiated class 1 and class 2 integron-carrying antibiotic resistant genes (ARGs) in separate clusters. Although cluster A exclusively comprised of class 1 integron, however, few class 1 integron positive isolates were also grouped along with class 2 integron in cluster B. Reports from previous studies have also described that integrons were not restricted but dispersed over a variety of MLVA genotypes in different bacterial species (Saffari et al. 2017; Taghadosi et al. 2019).

Population modeling of these *E. coli* isolates by minimum spanning tree explained the divergence of different serotypes and probable route of its transmission amongst different animal and water-borne *E. coli*. The current study depicted a separate branch for the STEC O157 isolates. Similar results were reported from Norway where serotypes like O157, O55 and O103 represented as separate branches from other *E. coli* serotypes (Løbersli et al. 2012). Thus, the present study depicted co-carriage of virulent phylogroups (B2 and D) and serotypes of potentially pathogenic strains (STEC, EPEC and UPEC) by various animal and water-borne *E. coli*.

Conclusion

The present study characterized 97 *E. coli* isolates by serotyping, phylogrouping and multilocus VNTR analysis. A wide variety of pathogenic serotypes including O157 and non-O157 STEC serotypes were reported. The phylogenetic analysis resulted in the predominance of virulent phylogroup B2 in the Yamuna river. Such a high incidence of pathogenic strains of *E. coli* in surface waters of a major river of North India could act as a great threat to public health. Four VNTR loci were used for typing of fecal and aquatic *E. coli* isolates. Some of the animal and water-borne *E. coli* isolates shared identical MLVA pattern showing clonal relatedness among them. MLVA typing resulted in 53 distinct genotypes and was found to be an efficient discriminatory tool (DI=0.95) which can be exploited for the surveillance of natural water bodies in developing countries like India.

Acknowledgements The present study was carried out with the financial support provided by UGC-RGNF [F1-17.1/2016-17/RGNF-2015-17-SC-HAR-18413/(SA-III/Website)] and UGC major research project (Grant No. UGC-41-1172/2012). NK and MK thank Maharshi Dayanand University for providing University Research Scholarship. MK expresses sincere gratitude to CSIR for Senior Research Fellowship (No. 09/382(0212)/19-EMR-I). The authors also wish to acknowledge FIST Grant funded by Department of Science and Technology (No. 1196 SR/FST/LS-I/2017/4) for improvement of infrastructural facilities. The authors are also grateful to National Salmonella and Escherichia Centre, Central Research Institute, Kasauli, Himachal Pradesh, India for serotyping *E. coli* isolates.

Author contribution P.G. conceived and planned the experiments. N.K. carried out the experiments. P.G., N.K. and M.K. contributed to the interpretation of the results. All the authors provided critical feedback and helped shape the research, analysis and manuscript.

Compliance with ethical standards

Conflict of interest The authors declare that there is no conflict of interest.

References

- Bai X, Hu B, Xu Y et al (2016) Molecular and phylogenetic characterization of non-O157 Shiga toxin-producing *Escherichia coli* strains in China. *Front Cell Infect Microbiol* 6:143
- Bajaj P, Singh NS, Kanaujia PK et al (2015) Distribution and molecular characterization of genes encoding CTX-M and AmpC β -lactamases in *Escherichia coli* isolated from an Indian urban aquatic environment. *Sci Total Environ* 505:350–356
- Berthe T, Ratajczak M, Clermont O et al (2013) Evidence for coexistence of distinct *Escherichia coli* populations in various aquatic environments and their survival in estuary water. *Appl Environ Microbiol* 79:4684–4693
- Bettelheim KA (2007) The non-O157 Shiga-toxigenic (verocytotoxinogenic) *Escherichia coli*; under-rated pathogens. *Crit Rev Microbiol* 33:67–87
- Camlena F, Birgy A, Smail Y et al (2019) Rapid and simple universal *Escherichia coli* genotyping method based on multiple-locus variable-number tandem-repeat analysis using single-tube multiplex PCR and standard gel electrophoresis. *Appl Environ Microbiol* 85:e02812–2818
- Carlos C, Pires MM, Stoppe NC et al (2010) *Escherichia coli* phylogenetic group determination and its application in the identification of the major animal source of fecal contamination. *BMC Microbiol* 10:161
- Chandran A, Hatha AA, Varghese S et al (2008) Prevalence of multiple drug resistant *Escherichia coli* serotypes in a Tropical Estuary, India. *Microbes Environ* 23:153–158
- Clermont O, Bonacorsi P, Bingen E (2000) Rapid and simple determination of the *Escherichia coli* phylogenetic group. *Appl Environ Microbiol* 66:4555–4558
- Cooley M, Carychao D, Miksa L et al (2007) Incidence and tracking of *Escherichia coli* O157:H7 in a major produce production region in California. *PLoS ONE* 2:e1159
- Cowan LS, Mosher L, Diem L et al (2002) Variable-number tandem repeat typing of *Mycobacterium tuberculosis* isolates with low copy numbers of IS6110 by using mycobacterial interspersed repetitive units. *J Clin Microbiol* 40:1592–1602
- Croxen MA, Law RJ, Scholz R et al (2013) Recent advances in understanding enteric pathogenic *Escherichia coli*. *Clin Microbiol Rev* 26:822–880
- Dahyot S, Lebeurre J, Argemi X et al (2018) Multiple-locus variable number tandem repeat analysis (MLVA) and tandem repeat sequence typing (TRST), helpful tools for subtyping *Staphylococcus lugdunensis*. *Sci Rep* 8:11669
- Finley RL, Collignon P, Larsson DG et al (2013) The scourge of antibiotic resistance: the important role of the environment. *Clin Infect Dis* 57:704–710
- Gorg O, Lopez S, Hilaire V et al (2008) Selection and validation of a multilocus variable-number tandem-repeat analysis panel for typing *Shigella* spp. *J Clin Microbiol* 46:1026–1036
- Grundmann H, Hori S, Tanner G (2001) Determining confidence intervals when measuring genetic diversity and the discriminatory abilities of typing methods for microorganisms. *J Clin Microbiol* 39:4190–4192
- Gulati P, Varshney RK, Virdi JS (2009) Multilocus variable number tandem repeat analysis as a tool to discern genetic relationships among strains of *Yersinia enterocolitica* biovar 1A. *J Appl Microbiol* 107:875–884
- Hagedorn C, Lepo JE, Hellein KN et al (2011) Microbial source tracking in China and developing nations. In: Hagedorn C, Blanch AR, Harwood VJ (eds) *Microbial source tracking: methods, applications and case studies*. Springer-Science, New York, pp 515–544
- Hamelin K, Bruant G, Shaarawi A et al (2007) Occurrence of virulence and antimicrobial resistance genes in *Escherichia coli* isolates from different aquatic ecosystems within the St. Clair River and Detroit River areas. *Appl Environ Microb* 73:477–484
- Hamner S, Broadaway SC, Mishra VB et al (2007) Isolation of potentially pathogenic *Escherichia coli* O157:H7 from the Ganges River. *Appl Environ Microbiol* 73:2369–2372
- Higgins J, Hohn C, Hornor S et al (2007) Genotyping of *Escherichia coli* from environmental and animal samples. *J Microbiol Methods* 70(2):227–235
- Hunter PR, Gaston MA (1988) Numerical index of the discriminatory ability of typing systems: an application of Simpson's index of diversity. *J Clin Microbiol* 26:2465–2466
- Jamwal P, Mittal AK, Mouchel JM (2011) Point and non-point microbial source pollution: a case study of Delhi. *Phys Chem Earth* 36:490–499
- Kaushik M, Khare N, Kumar S et al (2019) High prevalence of antibiotic resistance and integrons in *Escherichia coli* isolated from urban river water, India. *Microb Drug Resist* 25:359–370
- Kawamori F, Hiroi M, Harada T et al (2008) Molecular typing of Japanese *Escherichia coli* O157:H7 isolates from clinical specimens by multilocus variable-number tandem repeat analysis and PFGE. *J Med Microbiol* 57:58–63
- Krger A, Lucchesi PM, Sanso AM et al (2015) Genetic characterization of Shiga toxin-producing *Escherichia coli* O26:H11 strains isolated from animal, food, and clinical samples. *Front Cell Infect Microbiol* 5:74
- Lazzarini LC, Rosenfeld J, Huard RC et al (2012) *Mycobacterium tuberculosis* spoligotypes that may derive from mixed strain infections are revealed by a novel computational approach. *Infect Genet Evol* 12:798–806
- Lindstedt BA, Heir E, Gjernes E et al (2003) DNA fingerprinting of *Salmonella enterica* subsp. *enterica* serovar *Typhimurium* with emphasis on phage type DT104 based on variable number of tandem repeat loci. *J Clin Microbiol* 41:1469–1479
- Lindstedt BA, Brandal LT, Aas L et al (2007) Study of polymorphic variable-number of tandem repeats loci in the ECOR collection and in a set of pathogenic *Escherichia coli* and *Shigella* isolates for use in a genotyping assay. *J Microbiol Methods* 69:197–205
- Lbersli I, Haugum K, Lindstedt BA (2012) Rapid and high resolution genotyping of all *Escherichia coli* serotypes using 10 genomic repeat-containing loci. *J Microbiol Methods* 88:134–139
- Luna GM, Vignaroli C, Rinaldi C et al (2010) Extraintestinal *Escherichia coli* carrying virulence genes in coastal marine sediments. *Appl Environ Microbiol* 76:5659–5668
- McLellan SL (2004) Genetic diversity of *Escherichia coli* isolated from urban rivers and beach water. *Appl Environ Microbiol* 70:4658–4665
- Micnkov L, Bosk J, Vrba M et al (2016) Human extraintestinal pathogenic *Escherichia coli* strains differ in prevalence of virulence factors, phylogroups, and bacteriocin determinants. *BMC Microbiol* 16:218–225
- Mohapatra BR, Broersma K, Mazumder A (2007) Comparison of five rep-PCR genomic fingerprinting methods for differentiation of fecal *Escherichia coli* from humans, poultry and wild birds. *FEMS Microbiol Lett* 277:98–106

- Murray PR (2010) Matrix-assisted laser desorption ionization time-of-flight mass spectrometry: usefulness for taxonomy and epidemiology. *Clin Microbiol Infect* 16:1626–1630
- Noller AC, McEllistrem MC, Pacheco AGF et al (2003) Multilocus variable-number tandem repeat analysis distinguishes outbreak and sporadic *Escherichia coli* O157:H7 isolates. *J Clin Microbiol* 41:5389–5397
- Onteniente L, Brisse S, Tassios PT et al (2003) Evaluation of the polymorphisms associated with tandem repeats for *Pseudomonas aeruginosa* strain typing. *J Clin Microbiol* 41:4991–4997
- Ram S, Vajpayee P, Shanker R (2007) Prevalence of multiantimicrobial-agent resistant shiga toxin and enterotoxin producing *Escherichia coli* in surface waters of river Ganga. *Environ Sci Technol* 41:7383–7388
- Saffari F, Monsen T, Karmostaji A et al (2017) Significant spread of extensively drug resistant *Acinetobacter baumannii* genotypes of clonal complex 92 among intensive care unit patients in a university hospital in Southern Iran. *J Med Microbiol* 66:1656–1662
- Sasakova N, Gregova G, Takacova D et al (2018) Pollution of surface and ground water by sources related to agricultural activities. *Front Sustain Food Syst* 2:42
- Shangkuan YH, Yang JF, Lin HC et al (2000) Comparison of PCR-RFLP, ribotyping and ERIC-PCR for typing *Bacillus anthracis* and *Bacillus cereus* strains. *J Appl Microbiol* 89:452–462
- Stalder T, Barraud O, Casellas M et al (2012) Integron involvement in environmental spread of antibiotic resistance. *Front Microbiol* 3:119
- Taghadosi R, Shakibaie MR, -Nave H (2019) Antibiotic resistance, ESBL genes, integrons, phylogenetic groups and MLVA profiles of *Escherichia coli* pathotypes isolated from patients with diarrhea and farm animals in south-east of Iran. *Comp Immunol Microbiol Infect Dis* 63:117–126
- Tymensen LD, Pyrdok F, Coles D et al (2015) Comparative accessory gene fingerprinting of surface water *Escherichia coli* reveals genetically diverse naturalized population. *J Appl Microbiol* 119:263–277
- van Belkum A (2007) Tracing isolates of bacterial species by multilocus variable number of tandem repeat analysis (MLVA). *FEMS Immunol Med Microbiol* 49:22–27
- van Belkum A, Scherer S, van Alphen L et al (1998) Short-sequence DNA repeats in prokaryotic genomes. *Microbiol Mol Biol Rev* 62:275–293
- Walk ST, Alm EW, Calhoun LM et al (2007) Genetic diversity and population structure of *Escherichia coli* isolated from freshwater beaches. *Environ Microbiol* 9:2274–2288
- Weir BS (1990) Genetic data analysis: methods for discrete population genetic data analysis. Sinauer Associates Inc Publishers, Sunderland