MINI REVIEW

Accurate differentiation of *Escherichia coli* and *Shigella* serogroups: challenges and strategies

N. K. Devanga Ragupathi, D. P. Muthuirulandi Sethuvel, F. Y. Inbanathan and B. Veeraraghavan

Department of Clinical Microbiology, Christian Medical College, Vellore, India

Abstract

Shigella spp. and *Escherichia coli* are closely related; both belong to the family *Enterobacteriaceae*. Phenotypically, *Shigella* spp. and *E. coli* share many common characteristics, yet they have separate entities in epidemiology and clinical disease, which poses a diagnostic challenge. We collated information for the best possible approach to differentiate clinically relevant *E. coli* from *Shigella* spp. We found that a molecular approach is required for confirmation. High discriminatory potential is seen with whole genome sequencing analysed for k-mers and single nucleotide polymorphism. Among these, identification using single nucleotide polymorphism is easy to perform and analyse, and it thus appears more promising. Among the nonmolecular methods, matrix-assisted desorption ionization–time of flight mass spectrometry may be applicable when data analysis is assisted with advanced analytic tools.

Keywords: 16S rRNA, k-mer, MALDI-TOF MS, single nucleotide polymorphism, whole genome sequencing
Original Submission: 28 June 2017; **Revised Submission:** 7 September 2017; **Accepted:** 19 September 2017
Article published online: 23 September 2017

Corresponding author. B. Veeraraghavan, Department of Clinical Microbiology, Christian Medical College, Vellore 4, India
E-mail: vbalaji@cmcvellore.ac.in

Introduction

Diarrhoeal disease is not uncommon in both developing and developed countries. *Shigella* spp. are among the most important enteric pathogens causing bacillary dysentery worldwide, mainly in humans. Differentiation of *Shigella* spp. from *Escherichia coli* is challenging because of their close genetic relatedness. Brenner *et al.* [1] determined that the nucleotide similarity between *Shigella* and *E. coli* was 80% to 90%, whereas other *Escherichia* species are genetically distant [2]. *Shigellae* are phylogenetically *E. coli* that were later classified as separate species on the bases of biochemical characteristics and clinical relevance [3,4].

Biochemical characteristics and serotyping are usually used to identify the species. However, many isolates cannot be distinguished as either *E. coli* or *Shigella* spp. Molecular methods such as 16S rRNA gene sequencing and protein signature–based matrix-assisted laser desorption/ionization–time of flight mass spectrometry (MALDI-TOF MS) are unable to differentiate *Shigella* spp. from *E. coli* [4]. Further, *Shigella*-like strains of *E. coli* (enteroinvasive *E. coli*, EIEC) causing invasive dysenteric diarrhoeal illness make clinical and laboratory diagnoses difficult. In addition, the change in antimicrobial resistance patterns with the change in the serogroup/serotype further highlights the need for accurate identification of *Shigella* spp. so that appropriate antimicrobial therapy may be administered [5].

We attempted to accurately identify *E. coli* and *Shigella* spp., and trace the evolution of facts contributing to the masking of discrimination between *E. coli* and *Shigella* spp. We discuss the challenges and the possible methods to differentiate *E. coli* and *Shigella* spp. using protein signature and molecular tools.

Evolution of *Shigella* Species

At present, *Shigella* and *Escherichia* genera are considered to be unique genomospecies. Unlike *E. coli, Shigella* strains are nonmotile as a result of deletion in the *fliF* operon (flagellar

https://doi.org/10.1016/j.nmni.2017.09.003

coding region) or an IS*1* insertion mutation in the *flhD* operon. Also, *Shigella* does not ferment lactose, as *S. flexneri* [1,3] and *S. bodyii* [2,4] do not contain any of the *lac* genes (*lac*Y, *lac*A and *lac*Z) required for fermentation. *S. dysenteriae* 1 was known to have only *lac*Y and *lac*A. *S. sonnei* has all three genes but is unable to ferment as a result of lack of permease activity. These observations are one such example for the multiple origins of the *Shigella* phenotype by convergent evolution [6].

Earlier reports suggested that the arrival of a virulence plasmid into an *E. coli* strain gave rise to a monophyletic group from which all *Shigella* and *E. coli* groups descended. This led to the occurrence of highly diversified and pathogenic virotypes, which includes EIEC, Shiga toxin–producing *E. coli* (STEC; includes enterohemorrhagic *E. coli*, EHEC), enteropathogenic *E. coli* (EPEC), enteroaggregative *E. coli* (EAEC) and enterotoxigenic *E. coli* (ETEC) [7]. Interestingly, commensal *E. coli* strains may not become pathogenic *Shigella* on acquiring a virulence plasmid, as it does not seem to transmit horizontally among *E. coli* and *Shigella* strains [7].

STEC that is able to cause haemorrhagic colitis and haemolytic uremic syndrome is referred to as EHEC. This causes pancolitis due to toxigenic noninvasive (EHEC) infection, whereas EIEC causes proctocolitis via a nontoxigenic invasive mechanism similar to *Shigella* [8]. EIEC serotypes have been suggested as being ancestral to the different *Shigella* serogroups contributing to these differences [9]. However, supporting evidence for evolution of STEC is not clear. Similarly, limited information is available on the origins of other virotypes of *E. coli*.

In the midst of changing evolution, there is a need for accurate identification of *E. coli* and *Shigella* spp. for appropriate clinical management and accurate epidemiologic data. The accuracy of identification using molecular methods (duplex real-time PCR, 16S rRNA, multilocus sequence typing (MLST) and whole genome sequencing (WGS)) and nonmolecular methods (matrix-assisted desorption ionization–time of flight mass spectrometry, MALDI-TOF MS) will be discussed.

Currently Used Molecular Methods for Differentiation of *E. coli* and *Shigella* spp.

Duplex real-time PCR

A duplex real-time PCR for differentiation of EIEC and *Shigella* spp. was reported by Pavlovic *et al.* [10]; this PCR amplified the genes encoding β-glucuronidase (*uid*A) and lactose permease (*lac*Y). The gene *uid*A is common for *E. coli* and *Shigella,* while the latter (*lac*Y) is present only in *E. coli.* Ninety-six isolates including 11 EIEC isolates of different serotypes and at least three representatives of each *Shigella* species were identified correctly. Likewise, Lobersli *et al.* [11] established a duplex real-time PCR (*ipa*H and *lac*Y) to differentiate EIEC and *Shigella* spp., where *lac*Y is specific to *E. coli.* This PCR target differentiated *Shigella* spp. and EIEC O121 and O124 groups, but not EIEC O164 group.

16S rRNA gene sequencing to differentiate *E. coli* from *Shigella* spp.

Molecular identification using 16S rRNA sequencing could not distinguish atypical *E. coli* and *Shigella* spp. [12,13]. The 16S rRNA sequence similarities between various pathogenic strains of *E. coli,* EPEC (KR476716), EHEC (CP018252), STEC (CP015229), EIEC (AB604198), *E. coli* ATCC 25922 (KC429776), *S. boydii* (JQ073777), *S. sonnei* (HQ591457), *S. flexneri* (NR026331), *S. flexneri* 2a (CP012137), *S. flexneri* 5a (NZCM001474) and *S. dysenteriae* (NR026332) were calculated using the available reference 16S rRNA sequences from the National Center for Biotechnology Information (NCBI) database (Table 1).

The differentiation of *E. coli* and *Shigella* spp. could not be achieved using 16S rRNA gene sequences as a result of the narrow (<1%) divergence between EHEC, EIEC and *Shigella* spp. Jenkins *et al.* [14] concur with this finding; their 16S rRNA gene comparison could not distinguish between *E. coli* and *Shigella* spp. as a result of >99% sequence identity. We

TABLE 1. 16S rRNA sequence similarity between closely related *Shigella* serogroups, serotypes and virotypes of *Escherichia coli*

	E. coli ATCC 25922	EPEC	EHEC	STEC	EIEC	*S. dysenteriae*	*S. flexneri* 2a	*S. flexneri* 5a	*S. flexneri*	*S. boydii*	*S. sonnei*
E. coli ATCC 25922	100										
EPEC	98.89	100									
EHEC	99.04	98.89	100								
STEC	98.97	98.55	99.42	100							
EIEC	99.63	98	98.41	98.47	100						
S. dysenteriae	98.97	98.2	98.92	98.99	98.72	100					
S. flexneri 2a	99.63	98.06	98.91	98.97	99.53	98.86	100				
S. flexneri 5a	99.63	98	98.84	99.03	99.07	98.92	99.55	100			
S. flexneri	99.78	98.2	98.99	99.13	99.6	99.13	99.73	99.8	100		
S. boydii	99.56	98	98.8	98.87	99.66	98.79	99.93	99.47	99.66	100	
S. sonnei	99.56	97.93	98.78	98.97	99	98.86	99.49	99.68	99.73	99.4	100

EHEC, enterohaemorrhagic *E. coli*; EIEC, enteroinvasive *E. coli*; EPEC, enteropathogenic *E. coli*; STEC, Shiga toxin–producing *E. coli*.

therefore deem this approach to be unacceptable to differentiate certain inter- and intraspecies identity.

Exploration of MLST for differentiation of *E. coli* and *Shigella* spp.

The Pasteur and Warwick MLST databases use highly conserved housekeeping genes that are the same for both *E. coli* and *Shigella* spp. Hence, sequence types are assigned irrespective of *E. coli* and *Shigella* spp. A study by Li *et al.* [15] involving MLST for clinical *S. flexneri* isolates found that different serotypes (1–5, X and Y) were clustered together in a group, while a single serotype formed a distinct group. Li *et al.* reported the inability of MLST method to differentiate the evolutionary relationship between virotypes of *E. coli* and *Shigella* spp. However, there have been reports focusing directly on sequence data from the housekeeping genes rather than the allelic profile for clonal diversification. The discrimination based on difference in one MLST housekeeping gene sequence from the founder genotype is termed single-locus variants, and diversification of two housekeeping genes is defined as double-locus variants (DLVs) [16–19]. Until now, these variants were used to categorize clonal complexes to relate the phylogeny. Taking a cue from this knowledge, we made an attempt to use the direct sequence data of housekeeping genes to differentiate *E. coli* from *Shigella* spp.

Interestingly, we could identify the variations among *Shigella* spp. and *E. coli* virotypes beyond their sequence types utilizing the DLV approach (Fig. 1). Accurate identification was achieved using *rpo*B and *mdh* genes. *rpo*B, a protein-encoding housekeeping gene, has several potential advantages over other molecular methods. The *rpo*B gene occurs as a single copy in all prokaryotes, it functions as a housekeeping gene, it is less susceptible to some lateral gene transfer and its genetic divergence provides enhanced resolution for species identification. 16S rRNA gene copy number, however, varies among species and shows heterogeneity among intragenomic gene copies. *rpo*B is therefore the better marker to distinguish interspecies relationships between and within *E. coli* and *Shigella* spp. than 16S rRNA sequences [20]. Similarly, housekeeping gene malate–lactate dehydrogenase (*mdh*) was reported to provide good subtype discrimination between various subspecies [21], which reveals the evolutionary histories of *Salmonella* and *E. coli* chromosomes.

WGS for differentiation of *E. coli* and *Shigella* spp.

Differentiation of species based on WGS can be attained by two methods, k-mers and whole genome single nucleotide polymorphism (SNP). Chattaway *et al.* utilized k-mers (substrings of *k* nucleotides in DNA sequence data) to predict the species based on the number of co-occurring k-mers in two bacterial genomes as a measure of evolutionary relatedness. This accurately identified the strains to the species level [22–24]. Among 1297 isolates, 18 were misidentified by conventional biochemicals and serotyping. Of these, 15 were intragenomic misidentifications and three were intergenomic misidentifications. These 18 isolates were then correctly identified by the k-mer approach. The phylogenetic relation of the clonal complexes derived from MLST and a minimum spanning tree confirmed that the k-mer method was accurate in discriminating *Shigella* spp. from *E. coli*.

Recently the use of whole genome SNPs for drawing phylogenetic relationships has been gaining attention. Pettengill *et al.* [25] reported the ability of SNPs to accurately identify EIEC and *Shigella* spp. from WGS data. This method used 404

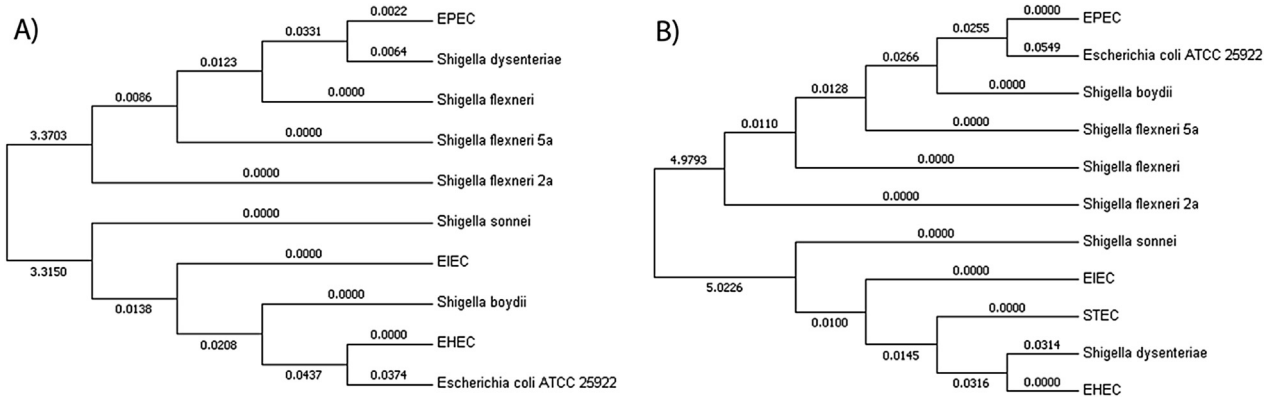

FIG. 1. Genotypic diversification of various *Escherichia coli* and *Shigella* spp. based on highly conserved housekeeping genes *mdh* (A) and *rpo*B (B). EHEC, EIEC, EPEC, STEC and ATCC 25922 *E. coli* form *E. coli* group; *S. dysenteriae, S. flexneri* 2a, *S. flexneri* 5a, *S. flexneri, S. boydii* and *S. sonnei* from *Shigella* group were used to construct double-locus variant–based phylogeny. EHEC, enterohaemorrhagic *E. coli*; EIEC, enteroinvasive *E. coli*; EPEC, enteropathogenic *E. coli*; STEC, Shiga toxin–producing *E. coli*.

TABLE 2. List of nonmolecular and molecular methods for accurate differentiation of *Escherichia coli* and *Shigella* spp

Method for differentiation	Target	Level of differentiation between *E. coli* and *Shigella* spp.	References
MALDI-TOF MS	Biomarker-based classifiers using their protein signature	Conventional MALDI-TOF MS fails to distinguish *E. coli* from *Shigella* spp. However, advanced software analytic tools like ClinPro could distinguish inactive and other non-lactose-fermenting *E. coli* from *Shigella* spp.	Francisco *et al.* [16]; Khot and Fischer [4]
Duplex real-time PCR	*uid*A, *lac*Y	Method is based on target-specific real-time PCR. EIEC and *Shigella* spp. can be differentiated because *lac*Y is specific for *E. coli*	Pavlovic *et al.* [10]
16S rRNA sequencing	16S rRNA	Unacceptable for discrimination of *E. coli* and *Shigella* spp. because sequence similarities were >99% for EIEC, EHEC and *Shigella* spp.	Edwards *et al.* [12]; Chen *et al.* [13]
MLST (conventional)	Housekeeping genes (*adk*, *fum*C, *gyr*B, *icd*, *mdh*, *pur*A, rec*A*)	Allele-based sequence type identification within *E. coli* and *Shigella* spp. without differentiating between them	Li *et al.* [15]
Specific locus variants	Housekeeping genes (*adk*, *fum*C, *gyr*B, *icd*, *mdh*, *pur*A, rec*A*)	Uses sequence data of housekeeping genes rather than MLST allelic profiles. Can differentiate within sequence types using single-locus variant and double-locus variant approach	Gibreel *et al.* [17]; Otero *et al.* [18]; Shahsavan *et al.* [19]
k-mer	k-mer regions	Serotype-level identification and differentiation of *E. coli* and *Shigella* spp. is performed using co-occurring k-mers	Hasman *et al.* [23]; Larsen *et al.* [24]; Chattaway *et al.* [22]
SNP	SNP markers	Specific SNP markers were used for classification using SNPs with their evolutionary phylogenetic relationships	Ashton *et al.* [26]; Pettengill *et al.* [25]

EHEC, enterohaemorrhagic *Escherichia coli*; EIEC, enteroinvasive *Escherichia coli*; MALDI-TOF MS, matrix-assisted desorption ionization–time of flight mass spectrometry; MLST, multilocus sequence typing; SNP, single nucleotide polymorphism.

SNP markers for differentiating *Shigella* and EIEC lineages. Further, Ashton *et al.* [26] proved classification of *Shigella* serotypes using SNPs with their evolutionary phylogenetic relationships. This seems to be an easier and more promising approach.

Identification based on ribosomal protein signature

MALDI-TOF MS is used for early species-level identification. However, the power of discrimination is still considered to be low for *Shigella* spp. [27]. In 2013, Khot and Fisher [4] reported that conventional MALDI-TOF MS failed to distinguish *Shigella* spp. from *E. coli.* However, they reported that MALDI-TOF MS with an automated data analysis approach could distinguish inactive and other non-lactose-fermenting *E. coli* from *Shigella* species [4]. This special approach included the use of ClinPro software's database and analysis tool functions like data preparation, model generation and spectra classification. Classification of unknown spectra for identification was achieved by using the 'Classify' function in ClinProTools, in which, if two or more of three spectra per isolate were assigned to the same class, the identification was accepted [16].

Table 2 compares the ability of each molecular method to differentiate *E. coli* and *Shigella* serogroups.

Conclusion

Among the molecular methods, we deem 16S rRNA to be unacceptable, while duplex real-time PCR and DLV using sequence data of the conserved housekeeping genes *rpo*B and *mdh* may be used. A high discriminatory potential is evident with WGS that analyses k-mers and SNPs. Among these two WGS modalities, identification using SNPs is easy to perform and analyse, and we think it is more promising. Among the nonmolecular methods, MALDI-TOF MS may be applicable when data analysis is assisted with advanced analytic tools.

Acknowledgement

Supported in part by the Indian Council of Medical Research, New Delhi, India (AMR/TF/55/13ECDII).

Conflict of Interest

None declared.

References

[1] Brenner DJ, Fanning GR, Steigerwalt AG, Orskov I, Orskov F. Polynucleotide sequence relatedness among three groups of pathogenic *Escherichia coli* strains. Infect Immun 1972;6:308–15.

[2] Van den Beld MJ, Reubsaet FA. Differentiation between *Shigella*, enteroinvasive *Escherichia coli* (EIEC) and noninvasive *Escherichia coli*. Eur J Clin Microbiol Infect Dis 2012;31:899–904.

[3] Connor TR, Barker CR, Baker KS, Weill FX, Talukder KA, Smith AM, et al. Species-wide whole genome sequencing reveals historical global spread and recent local persistence in *Shigella flexneri*. Elife 2015;4:e07335.

[4] Khot PD, Fisher MA. Novel approach for differentiating *Shigella* species and *Escherichia coli* by matrix-assisted laser desorption ionization–time of flight mass spectrometry. J Clin Microbiol 2013;51:3711–6.

[5] Shakya G, Acharya J, Adhikari S, Rijal N. Shigellosis in Nepal: 13 years review of nationwide surveillance. J Health Popul Nutr 2016;35:36.

[6] Pupo GM, Lan R, Reeves PR. Multiple independent origins of *Shigella* clones of *Escherichia coli* and convergent evolution of many of their characteristics. Proc Natl Acad Sci U S A 2000;197:10567–72.

[7] Escobar-Paramo P, Giudicelli C, Parsot C, Denamur E. The evolutionary history of *Shigella* and enteroinvasive *Escherichia coli* revised. J Mol Evol 2003;57:140–8.
[8] Johnson JR. *Shigella* and *E. coli* at the crossroads: Machiavellian masqueraders or taxonomic treachery? J Med Microbiol 2000;49:583–5.
[9] Lukjancenko O, Wassenaar TM, Ussery DW. Comparison of 61 sequenced *Escherichia coli* genomes. Microb Ecol 2010;60:708–20.
[10] Pavlovic M, Luze A, Konrad R, Berger A, Sing A, Busch U. Development of a duplex real-time PCR for differentiation between *E. coli* and *Shigella* spp. J Appl Microbiol 2011;110:1245–51.
[11] Lobersli I, Wester AL, Kristiansen A, Brandal LT. Molecular differentiation of *Shigella* spp. from enteroinvasive *E. coli*. Eur J Microbiol Immunol 2016;6:197–205.
[12] Edwards KJ, Logan JMJ, Langham S, Swift C, Gharbia SE. Utility of real-time amplification of selected 16S rRNA gene sequences as a tool for detection and identification of microbial signatures directly from clinical samples. J Med Microbiol 2012;61:645–52.
[13] Chen L, Cai Y, Zhou G, Shi X, Su J, Chen G, et al. Rapid Sanger sequencing of the 16S rRNA gene for identification of some common pathogens. PLoS One 2014;9:e88886.
[14] Jenkins C, Ling CL, Ciesielczuk HL, Lockwood J, Hopkins S, McHugh TD, et al. Detection and identification of bacteria in clinical samples by 16S rRNA gene sequencing: comparison of two different approaches in clinical practice. J Med Microbiol 2012;61:483–8.
[15] Li S, Sun Q, Wei X, Klena JD, Wang J, Liu Y. Genetic characterization of *Shigella flexneri* isolates in Guizhou province, China. PLoS One 2015;10:e0116708.
[16] Francisco AP, Bugalho M, Ramirez M, Carriço JA. Global optimal eBURST analysis of multilocus typing data using a graphic matroid approach. BMC Bioinformatics 2009;10:152.
[17] Gibreel TM, Dodgson AR, Cheesbrough J, Fox AJ, Bolton FJ, Upton M. Population structure, virulence potential and antibiotic susceptibility of uropathogenic *Escherichia coli* from Northwest England. J Antimicrob Chemother 2012;67:346–56.
[18] Otero V, Rodriguez-Calleja JM, Otero A, Garcia-Lopez ML, Santos JA. Genetic characterization of atypical enteropathogenic *Escherichia coli* isolates from ewes' milk, sheep farm environments, and humans by multilocus sequence typing and pulsed-field gel electrophoresis. Appl Environ Microbiol 2013;79:5864–9.
[19] Shahsavan S, Nobakht M, Rastegar-Lari A, Owlia P, Bakhshi B. Multilocus sequence type analysis of *Shigella* spp. isolates from Tehran, Iran. Iranian J Microbiol 2016;8:298–306.
[20] Vos M, Quince C, Pijl AS, de Hollander M, Kowalchuk GA. A Comparison of *rpo*B and 16S rRNA as markers in pyrosequencing studies of bacterial diversity. PLoS One 2012;7:e30600.
[21] Brown EW, Mammel MK, LeClerc JE, Cebula TA. Limited boundaries for extensive horizontal gene transfer among *Salmonella* pathogens. Proc Natl Acad Sci U S A 2003;100:15676–81.
[22] Chattaway MA, Schaefer U, Tewolde R, Dallman TJ, Jenkins C. Identification of *Escherichia coli* and *Shigella* species from whole-genome sequences. J Clin Microbiol 2017;55:616–23.
[23] Hasman H, Saputra D, Sicheritz-Ponten T, Lund O, Svendsen CA, Frimodt-Moller N, et al. Rapid whole-genome sequencing for detection and characterization of microorganisms directly from clinical samples. J Clin Microbiol 2014;52:139–46.
[24] Larsen MV, Cosentino S, Lukjancenko O, Saputra D, Rasmussen S, Hasman H, et al. Benchmarking of methods for genomic taxonomy. J Clin Microbiol 2014;52:1529–39.
[25] Pettengill EA, Pettengill JB, Binet R. Phylogenetic analyses of *Shigella* and enteroinvasive *Escherichia coli* for the identification of molecular epidemiological markers: whole-genome comparative analysis does not support distinct genera designation. Front Microbiol 2015;6:1573.
[26] Ashton PM, Baker KS, Gentle A, Wooldridge DJ, Thomson NR, Dallman TJ, et al. Draft genome sequences of the type strains of *Shigella flexneri* held at Public Health England: comparison of classical phenotypic and novel molecular assays with whole genome sequence. Gut Pathog 2014;6:7.
[27] Dekker J, Frank K. Salmonella, Shigella, and *Yersinia*. Clin Lab Med 2015;35:225–46.