Cluster-specific gene marker enhance *Shigella* and Enteroinvasive *Escherichia coli in silico* serotyping

Xiaomei Zhang[1], Michael Payne[1], Thanh Nguyen[1], Sandeep Kaur[1], Ruiting Lan[1]*

[1]School of Biotechnology and Biomolecular Sciences, University of New South Wales, Sydney, New South Wales, Australia

*Corresponding Author

Email: r.lan@unsw.edu.au

Phone: 61-2-9385 2095

Fax: 61-2-9385 1483

Keywords: Phylogenetic clusters, cluster-specific gene markers, *Shigella*/EIEC serotyping

Running title: *Shigella* EIEC Cluster Enhanced Serotyping

Repositories: Raw sequence data are available from NCBI under the BioProject number PRJNA692536.

Abstract

Shigella and enteroinvasive *Escherichia coli* (EIEC) cause human bacillary dysentery with similar invasion mechanisms and share similar physiological, biochemical and genetic characteristics. The ability to differentiate *Shigella* and EIEC from each other is important for clinical diagnostic and epidemiologic investigations. The existing genetic signatures may not discriminate between *Shigella* and EIEC. Phylogenetically, *Shigella* and EIEC strains composed of multiple clusters and are different forms of *E. coli*. In this study, we identified 10 *Shigella* clusters, 7 EIEC clusters and 53 sporadic types of EIEC by examining over 17,000 publicly available *Shigella*/EIEC genomes. We compared *Shigella* and EIEC accessory genomes to identify the cluster-specific gene markers or marker sets for the 17 clusters and 53 sporadic types. The gene markers showed 99.63% accuracy and more than 97.02% specificity. In addition, we developed a freely available *in silico* serotyping pipeline named *Shigella* EIEC Cluster Enhanced Serotype Finder (ShigEiFinder) by incorporating the cluster-specific gene markers and established *Shigella*/EIEC serotype specific O antigen genes and modification genes into typing. ShigEiFinder can process either paired end Illumina sequencing reads or assembled genomes. ShigEiFinder provided nearly perfect differentiation of *Shigella* from EIEC with 99.70% and 99.81% accuracy to assign isolates to the correct clusters for the assembled genomes and reads mapping respectively. ShigEiFinder was able to serotype over 59 *Shigella* serotypes and 22 EIEC serotypes and provided a high specificity with 99.40% for assembled genomes and 99.38% for reads mapping for serotyping. The cluster markers and our new serotyping tool, ShigEiFinder, will be useful for epidemiologic and diagnostic investigations.

Data summary

Sequencing data have been deposited at the National Center for Biotechnology Information under BioPreoject number PRJNA692536.

Introduction

Shigella is one of the most common etiologic agents of foodborne infections worldwide and can cause diarrhea with a very low infectious dose (1, 2). The infections can vary from mild diarrhea to severe bloody diarrhea referred to as bacillary dysentery. The estimated cases of *Shigella* infections are 190 million with at least 210,000 deaths annually, predominantly in children younger than 5 years old in developing countries (3-7). *Shigella* infections also have a significantly impact on public health in developed countries although most cases are travel-associated (8).

The *Shigella* genus consists of four species, *Shigella sonnei*, *Shigella flexneri*, *Shigella boydii* and *Shigella dysenteriae* (9). Serological testing further classifies *Shigella* species into more than 55 serotypes through the agglutination reaction of antisera to *Shigella* serotype specific O-antigens (10, 11). Up to 89.6% *Shigella* infections were caused by *S. flexneri* (65.9%) and *S. sonnei* (23.7%) globally (12, 13). The predominant serotype reported in *Shigella* infections has been *S. flexneri* serotype 2a while *S. dysenteriae* serotype 1 has caused the most severe disease (11, 14). Note that for brevity, in all references to *Shigella* serotypes below, *S. sonnei*, *S. flexneri*, *S. boydii* and *S. dysenteriae* are abbreviated as SS, SF, SB and SD respectively and a serotype is designated with abbreviated “species” name plus the serotype number e.g. *S. dysenteriae* serotype 1 is abbreviated as SD1.

Enteroinvasive *Escherichia coli* (EIEC) is a pathovar of *E. coli* that causes diarrhoea with less severe symptoms to *Shigella* infections in humans worldwide, particularly in developing countries (8, 13, 15-18). EIEC infections in developed countries are mainly imported (19). EIEC has more than 18 specific E. coli O-serotypes (19, 20). Although the incidence of EIEC is low (17), EIEC serotypes have been associated with outbreaks and sporadic cases of infections (20-22). In contrast to *Shigella*, EIEC infections are not notifiable in many countries (23, 24).

Shigella and EIEC have always been considered very closely related and share several characteristics (25-28). *Shigella* and EIEC are both non-motile and lack the ability of ferment lactose (24). Some of EIEC O antigens are identical or similar to *Shigella* O antigens (O112ac, O124, O136, O143, O152 and O164) (26, 29-31). Furthermore, *Shigella* and EIEC both carry the virulence plasmid pINV, which encodes virulence genes required for invasion (32, 33) and contains *ipaH* (invasion plasmid antigen H) genes with the exception of some SB13 isolates

(10, 23, 24, 34, 35). *Shigella* and EIEC have arisen from *E. coli* in multiple independent events and should be regarded as a single pathovar of E. coli (25, 26, 28, 36-38). Previous phylogenetic studies suggested that *Shigella* isolates were divided into 3 clusters (C1, C2 and C3) with 5 outliers (SS, SB13, SD1, SD8 and SD10) (25, 38) whereas EIEC isolates were grouped into four clusters (C4, C5, C6 and C7) (26). The seven *Shigella*/EIEC clusters and 5 outliers of *Shigella* are within the broader non-enteroinvasive *E. coli* species except for SB13 which is closer to *Escherichia albertii* (39, 40). Alternative WGS-based phylogenomic studies have also defined multiple clusters of *Shigella* and EIEC (23, 28, 41).

The traditional biochemical test for motility and lysine decarboxylase (LDC) activity (42) and molecular test for the presence of *ipaH* gene have been used to differentiate *Shigella* and EIEC from non-enteroinvasive E. coli (24, 43-45). Agglutination with *Shigella*/EIEC associated antiserum further classify *Shigella* or EIEC to serotype level. However, cross-reactivity, strains not producing O antigens, and newly emerged *Shigella* serotypes may all prevent accurate serotyping (10, 46). Serotyping by antigenic agglutination is being replaced by molecular serotyping (47, 48), which can be achieved through examination of the sequences of O antigen biosynthesis and modification genes (8, 24, 49-52).

Recently, PCR-based molecular detection methods targeting the gene *lacY* were developed to distinguish *Shigella* from EIEC (53, 54). However, the ability of the primers described in these methods to accurately differentiate between *Shigella* and EIEC was later questioned (23, 28). With the uptake of whole-genome sequencing technology, several studies have identified phylogenetic clade specific markers, species specific markers and EIEC lineage-specific genes for discrimination between *Shigella* and EIEC and between *Shigella* species (23, 27, 28, 41, 55, 56). More recently, genetic markers *lacY*, *cadA*, *Ss_methylase* were used for identification of *Shigella* and EIEC (10). However, these markers failed to discriminate between *Shigella* and EIEC when a larger genetic diversity is considered (23, 28, 55). A Kmer-based approach can identify *Shigella* isolates to the species level but misidentification was also observed (56).

In this study, we aimed to i), identify phylogenetical clusters of *Shigella* and EIEC through large scale examination of publicly available genomes; ii), identify cluster-specific gene markers using comparative genomic analysis of *Shigella* and EIEC accessory genomes for differentiation of *Shigella* and EIEC; iii), develop a pipeline for *Shigella* and EIEC *in silico* serotyping based on the cluster-specific gene markers combined with *Shigella* and EIEC serotype-specific O antigen and H antigen genes. We demonstrate that these cluster-specific gene markers enhance *in silico* serotyping using genomic data. We also developed an automated pipeline for cluster typing and serotyping of *Shigella*/EIEC from WGS data.

Materials and Methods

Identification of *Shigella*/EIEC isolates from NCBI database

E. coli/Shigella isolates from the NCBI SRA (National Center for Biotechnology Information Sequence Read Archive) as May of 2019 were queried. Raw reads were retrieved from ENA (European Nucleotide Archive). The *ipaH* gene (GenBank accession number M32063.1) was used to screen E. coli/*Shigella* reads using Salmon v0.13.0 (57). Taxonomic classification for *E. coli/Shigella* was confirmed by Kraken v1.1.1 (58). Molecular serotype prediction of *ipaH* negative *Shigella* isolates was performed by ShigaTyper v1.0.6 (10). Isolates that were *ipaH* positive and isolates with designation of SB13 by ShigaTyper were selected as *Shigella*/EIEC database.

The sequence types (STs) and ribosomal STs (rSTs) of *ipaH* gene negative E. coli (non-enteroinvasive E. coli) isolates were examined. STs and rSTs for these isolates were obtained from the E. coli/*Shigella* database in the Enterobase (59) as of May 2019. For STs and rSTs with only one isolate, the isolates were selected. For STs and rSTs with more than one isolates, one representative isolate for each ST and rST were randomly selected. In total, 12,743 ipaH negative E. coli isolates representing 3,800 STs and 11,463 rSTs were selected as non-enteroinvasive E. coli control database.

Genome sequencing

Whole-genome sequencing (WGS) of 31 EIEC strains used in a previous study (26) was performed by Illumina NextSeq (Illumina, Scoresby, VIC, Australia). DNA libraries were constructed using Nextera XT Sample preparation kit (Illumina Inc., San Diego, CA, USA) and sequenced using the NextSeq sequencer (Illumina Inc.). FASTQ sequences of the strains sequenced in this study were deposited in the NCBI under the BioProject (PRJNA692536).

Genome assembly and data processing

Raw reads were *de novo* assembled using SPADES v3.14.0 assembler with default settings [http://bioinf.spbau.ru/spades] (60). The metrics of assembled genomes were obtained with

QUAST v5.0.0 (61). Three standard deviations (SD) from the mean for contig number, largest contig, total length, GC, N50 and genes were used as quality filter for assembled genomes.

The STs for isolates in *Shigella*/EIEC database was checked by using mlst (https://github.com/tseemann/mlst) with the *E. coli* scheme from PubMLST (62). rSTs were extracted from the E. coli/*Shigella* rMLST database in Enterobase (59) as of May 2019. Serotype prediction for isolates in *Shigella*/EIEC was performed by ShigaTyper v1.0.6 (10). Serotyping of E. coli O and H antigens were predicted by using SerotypeFinder v2.0.1 (63).

Selection of isolates for *Shigella*/EIEC identification dataset

The selection of isolates for the identification dataset was based on the representative isolates for each ST, rST and serotype of *Shigella* and EIEC in the *Shigella*/EIEC database. For STs and rSTs with only one isolate, the isolate was selected. For STs and rSTs with more than one isolates, one representative isolate for each ST, rST was randomly selected. A representative experimentally confirmed isolate of each serotype of *Shigella* and EIEC was also randomly selected. 72 ECOR strains downloaded from Enterobase (59) and 18 *E. albertii* strains were used as controls for the identification dataset. The details of the identification dataset are listed in Table S1. The remaining isolates in *Shigella*/EIEC database were referred as validation dataset (Table S2).

The identification dataset was used for identification of phylogenetic relationships of *Shigella* and EIEC. The identification dataset was also used for identification of cluster-specific genes. The validation dataset was used to evaluate the performance of cluster-specific gene markers using the *in-silico* serotyping pipeline.

Phylogeny of *Shigella* and EIEC based on WGS

Three phylogenetic trees including identification tree, confirmation tree and validation tree were constructed by Quicktree v1.3 (64) with default parameters to identify and confirm the phylogenetic clustering of *Shigella* and EIEC isolates. The phylogenetic trees were visualised by Grapetree's interactive mode and ITOL v5 (65, 66).

The identification phylogenetic tree was generated based on isolates in the identification dataset for identification of clusters of *Shigella* and EIEC isolates (Fig. 1). A subset of 485 isolates known to represent each identified cluster from the identification dataset were then selected. The subset of 485 isolates from the identification dataset and 1,872 non-enteroinvasive E. coli isolates from non-enteroinvasive E. coli control dataset (2,357 isolates total) were used to construct a confirmation tree. This tree was used for confirmation of the phylogenetic relationships between identified *Shigella*/EIEC clusters in the identification dataset and non-enteroinvasive E. coli isolates. The validation tree was generated based on 1,159 representative isolates from the validation dataset that were selected in the same way as the identification dataset and a subset of 485 isolates from the identification dataset to assign validation dataset isolates to clusters.

Investigation of *Shigella* virulence plasmid pINV

The presence of *Shigella* virulence plasmid pINV in isolates were investigated by using BWA-MEM v0.7.17 (Burrows-Wheeler Aligner) (67) to align isolate raw reads onto the reference sequence of pINV (68) (NC_024996.1). Mapped reads were sorted and indexed using Samtools v1.9 (69). The individual gene coverage from mapping was obtained using Bedtools coverage v2.27.1 (70).

Identification of the cluster-specific gene markers

Cluster-specific gene markers were identified from *Shigella/*EIEC accessory genomes. The genomes from the identification dataset were annotated using PROKKA v1.13.3 (71). Pan- and core-genomes were analysed by roary v3.12.0 (72) using an 80% sequence identity threshold. The genes specific to each cluster were identified from the accessory genes with an in-house python script. In this study, the number of genomes from a given cluster containing all specific genes for that cluster was termed true positives (TP), the number of genomes from the same cluster lacking any of those same genes was termed false negatives (FN). The number of genomes from other clusters containing all of those same genes was termed false positives (FP).

The sensitivity (True positive rate, TPR) of each cluster-specific gene marker was defined as TP/(TP+FN). The specificity (True negative rate, TNR) was defined as TN/(TN+FP).

Validation of the cluster-specific gene markers

The ability of cluster-specific gene markers to assign *Shigella*/EIEC isolates was examined by using BLASTN to search against the validation dataset (Table S2) and non-enteroinvasive E. coli control database for the presence of any of the cluster-specific gene marker or a set of cluster-specific gene markers. The BLASTN thresholds were defined as 80% sequence identity and 50% gene length coverage.

Development an automated pipeline for molecular serotyping of *Shigella*/EIEC

The pipeline was developed using paired end illumina genome sequencing reads or assembled genomes identify cluster-specific gene markers combined with *Shigella*/EIEC serotype specific O antigen genes (wzx and wzy) and modification genes (Fig. 2, Data S1). We used the same signature O and H sequences from ShigaTyper and SerotypeFinder (Data S2) (10, 63). These includes *Shigella* serotype-specifc wzx/wzy genes and modification genes from ShigaTyper and E. coli O antigen and *fliC* (H antigen) genes from SerotypeFinder. *ipaH* gene and 38 virulence genes used in analysis of virulence of 59 sporadic EIEC isolates were also included in the typing reference sequences database. Seven House Keeping (HK) genes -*recA*, *purA*, *mdh*, *icd*, *gyrB*, *fumC* and *adk* downloaded from NCBI were used for contamination checking.

Raw reads were aligned to the typing reference sequences by using BWA-MEM v0.7.17 (67). The mapping length percentage and the mean mapping depth for all genes were calculated using Samtools coverage v1.10 (69). To determine whether the genes present or absent, 50% of mapping length for all cluster-specific genes, virulence genes and O antigen genes and 10% for *ipaH* gene were used as cutoff value. The ratio of mean mapping depth to the mean mapping depth of the 7 HK genes was used to determine a contamination threshold with ratios less than 1% for *ipaH* gene and less than 10% for other genes assigned as contamination. Reads coverage mapped to particular regions of genes were checked by using samtools mpileup v1.10.

Assembled genomes were BLASTN v2.9.0 (73) searched against the typing reference sequences with 80% sequence identity and 50% gene length coverage for all genes with exception of *ipaH* gene which was defined as 10% gene length coverage.

The pipeline was tested with the identification dataset and validated with the *Shigella*/EIEC validation dataset and non-enteroinvasive E. coli control database. The specificity defined as (1 - the number of non-enteroinvasive E. coli isolates being detected / the total number of non-enteroinvasive E. coli isolates) * 100.

Results

Screening sequenced genomes for *Shigella*/EIEC isolates

We first screened available *E. coli* and *Shigella* genomes based on the presence of *ipaH* gene. We examined 122,361 isolates with the species annotation of E. coli (104,256) or *Shigella* (18,105) with paired end illumina sequencing reads available in NCBI SRA database. Of 122,361 isolates, 17,989 isolates were positive to the *ipaH* gene including 455 out of 104,256 E. coli isolates and 17,434 out of 18,105 *Shigella* isolates. The 17,989 *ipaH* positive *E. coli* and *Shigella* genomes and 571 *ipaH* negative "*Shigella"* genomes were checked for taxonomic classification and genome assembly quality. 17,320 *ipaH* positive *E. coli* and *Shigella* genomes and 246 *ipaH* negative "*Shigella"* genomes passed quality filters. Among 246 *ipaH* negative "*Shigella"* genomes, 11 isolates belonged to SB13 by using ShigaTyper (10) while the remaining 235 isolates were classified with taxonomic identifier of E. coli by Kraken v1.1.1 (58) and were removed from analysis. A total of 17,331 genomes including 17,320 *ipaH* positives and 11 SB13 genomes were selected to form the *Shigella*/EIEC database, which contained 429 genomes with species identifier of *E. coli* and 16,902 genomes with species identifier of *Shigella.*

Isolates in *Shigella*/EIEC database were typed using MLST, ShigaTyper and serotypeFinder. MLST and rMLST divided the 17,331 *Shigella*/EIEC isolates into 252 STs (73 isolates untypeable by MLST) and 1,128 rSTs (3,513 isolates untypeable by rMLST). Of 16,902 genomes with species identifier of *Shigella*, 8,313 isolates and 8,189 isolates were typed as *Shigella* and EIEC respectively by ShigaTyper while 400 isolates were untypeable. ShigaTyper typed the majority of the 8,313 isolates as SF (66.82%) including 25.43% SF2a isolates, followed by SS (19.69%), SB (7.22%) and SD (6.27%).

SerotypeFinder typed 293 of the 429 *E. coli* genomes into 71 *E. coli* O/H antigen types. Among these 293 isolates with typable O/H antigen types, 190 isolates belonged to 22 known EIEC serotypes (O28ac:H-, O28ac:H7, O29:H4, O112ac:H26, O121:H30, O124:H30, O124:H24, O124:H7, O132:H7, O132:H21, O135:H30, O136:H7, O143:H26, O144:H25, O152:H-, O152:H30, O164:H-, O164:H30, O167:H26, O173:H7 and 2 newly emerged EIEC serotypes O96:H19 and O8:H19) (20-22). The remaining 136 of 429 genomes were O antigen untypable and typed to 15 H antigen types only by SerotypeFinder, of which H16 was the predominant H antigen type.

Identification of *Shigella* and EIEC clusters

Shigella and EIEC are known to have been derived from *E. coli* independently. To identify previously defined clusters (25, 26) and any new clusters from the 17,331 *Shigella*/EIEC genomes, we selected representative genomes to perform phylogenetic analysis as it was impractical to construct a tree with all genomes. The selection was based on ST, rST and serotype of the 17,331 *Shigella*/EIEC genomes. One isolate was selected to represent each ST, rST and serotype for a total of 1,830 isolates. The selection included 252 STs, 1,128 rSTs, 59 *Shigella* serotypes (21 SB serotypes, 20 SF serotypes, 17 SD serotypes and SS), 22 EIEC known serotypes and 31 other or partial antigen types. A further 31 in-house sequenced EIEC isolates, 18 EIEC isolates used in a previous typing study (41), 72 ECOR strains and 18 *E. albertii* strains were also included to form the identification dataset of 1,969 isolates. Details are listed in Table S1. A phylogenetic tree was constructed based on the identification dataset to identify the clusters (Fig. 1).

All known clusters were identified (Fig. 1) including 3 *Shigella* clusters (C1, C2, C3) and 5 outliers (SD1, SD8, SD10, SB13 and SS) as defined by Pupo et al (25) and 4 EIEC clusters (C4, C5, C6 and C7) defined by Lan et al. (26). Each of these clusters was supported by a bootstrap value of 80% or greater (Fig. S1). 1,789 isolates of the 1,879 *Shigella*/EIEC isolates (1,830 isolates from the *Shigella*/EIEC database, 31 in-house sequenced EIEC isolates and 18 EIEC isolates from Hazen *et al.*) fell within these clusters.

Of the remaining 90 *Shigella*/EIEC unclustered isolates, 31 belonged to 5 *Shigella*/EIEC serotypes including 5 SB13 isolates, 8 SB12 isolates, 2 EIEC O135:H30 isolates, 12 EIEC serotype O96:H19 isolates and 4 EIEC O8:H19 isolates, while 59 isolates were sporadic EIEC isolates which are described in detail in the separate section below. The 5 SB13 isolates were grouped into one lineage within *E. coli* and close to known *Shigella*/EIEC clusters rather than the established SB13 cluster outside *E. coli* which was within the *E. albertii* lineage. The former was previously named as atypical SB13 while the latter was previously named as typical SB13 (39). The 8 SB12 isolates formed one single cluster close to SD1 and atypical SB13 clusters. Two EIEC O135:H30 isolates were grouped as a separate cluster close to C5. Twelve isolates belonging to EIEC serotype O96:H19 and 4 isolates typed as O8:H19 were clustered into two separate clusters, both of which were more closely related to SD8 than other *Shigella*/EIEC clusters. Therefore, atypical SB13 and SB12 were defined as new clusters of

Shigella while EIEC O96:H19, EIEC O8:H19 and EIEC O135:H30 were defined as C8, C9 and C10 respectively. In total there were 10 *Shigella* clusters and 7 EIEC clusters (Table 1).

Analysis of the 59 sporadic EIEC isolates

To determine the phylogenetic relationships of the above defined clusters and the remaining 59 sporadic EIEC isolates within the larger non-enteroinvasive *E. coli* population a confirmation tree was generated using 485 isolates representing the known clusters and 1,872 representative non-*Shigella*/EIEC isolates (Fig. S2). The 59 sporadic EIEC isolates including 2 EIEC isolates M2330 (O152:H51) and M2339 (O124:H7) sequenced in this study and 57 isolates were interspersed among non-*Shigella*/EIEC isolates and did not form large clusters. Groups of these isolates that were not previously identified were named as sporadic EIEC lineage followed by their serotype. For example, M2339 (O124:H7) grouped together with one other EIEC isolate with the same O and H antigens O124:H7 and were named ‘sporadic EIEC lineage O124:H7’. There were 53 sporadic EIEC lineages including 5 lineages with 2 or more isolates and 48 lineages with only one isolate. The STs, rSTs and antigen types of these 59 isolates were listed in the Table S1.

Some of the sporadic EIEC isolates fell into STs containing *ipaH* negative isolates. We therefore examined the presence of the pINV virulence plasmid in the sporadic EIEC isolates. We selected 38 genes that are essential for virulence including 35 genes (12 *mxi* genes, 9 *spa* genes, 5 *ipaA-J* genes, 6 *ipgA-F* genes as well as *acp*, *virB*, *icsB*) in the conserved entry region encoding the Mxi-Spa-Ipa type III secretion system and its effectors and 3 regulator genes (*virF*, *virA* and *icsA*/*virG*) (24, 33, 68) and determined the presence of pINV in the 59 sporadic EIEC isolates by mapping the sequence reads onto a pINV reference sequence (68). Reads from 18 non-*Shigella/EIEC* isolates that shared the same ST as one of 58 sporadic isolates were positive for these genes.

The number of essential virulence genes with mapped reads in the 59 sporadic EIEC isolates were analysed (Fig. S3). Those isolates containing more than 25 of the 38 essential virulence genes were defined as virulence plasmid positive. While isolates containing between 13 and 25 were defined as intermediate and less than 13 were defined as virulence plasmid negative.

The 2 newly sequenced sporadic EIEC isolates (M2330 and M2339) were positive for the virulence plasmid and of the other 57 sporadic EIEC isolates, 39 isolates were positive, 9 isolates were negative and 9 isolates were intermediate (Table S1). The results were compared with those non-*Shigella/*EIEC isolates belonging to the same ST. The virulence plasmid was absent in all non-*Shigella/*EIEC isolates while all sporadic EIEC isolates in these STs were either positive or intermediate. Therefore, this analysis confirmed the sporadic isolates belonged to EIEC and the STs contained both EIEC and non- EIEC isolates.

Identification of cluster-specific gene markers

In this study, cluster-specific gene markers were either a single gene present in all isolates of a cluster and absent in all other isolates or a set of genes (two or more) that as a combination were only found in one cluster. For the marker sets, a subset of cluster-specific gene markers for a given cluster could be found in other clusters but the entire set was only found in the target cluster.

Comparative genomic analysis on 1,969 accessory genomes from the identification dataset was used to identify cluster-specific gene markers or marker sets. Multiple candidate cluster-specific gene markers or marker sets of markers for each of 17 *Shigella*/EIEC clusters and 53 sporadic EIEC lineages were identified through screening the accessory genes from 1,969 genomes . These gene markers or marker sets were 100% sensitive to clusters but with varying specificity. The cluster-specific gene markers or marker sets of markers with the lowest FP rates were then selected from candidate cluster-specific gene markers by BLASTN searches against genomes in the identification dataset using 80% sequence identity and 50% gene length coverage threshold.

Five single cluster-specific gene markers (C7, C10, SB12, SB13 and atypical SB13) and 12 sets of cluster-specific gene markers (C1, C2, C3, C4, C5, C6, C8, C9, SS, SD1, SD8 and SD10) were selected for *Shigella*/EIEC cluster typing. The sensitivity and specificity for each cluster-specific gene marker or a set of cluster-specific gene markers for the identification dataset were listed in Table 2. The cluster-specific gene markers or marker sets of markers were all 100% sensitive and 100% specific with exception of C1 (99.94%), C3 (99.91% specificity) and SS (99.8% specificity). A single specific gene for each of 53 sporadic EIEC lineages were also selected with the exception of one lineage which has a set of 2 genes. These genes were all 100% sensitive and specific for a given sporadic EIEC lineage.

All cluster-specific gene markers, 37 in total (5 single, 32 genes in 12 sets) and 54 sporadic EIEC lineages specific gene markers were located on chromosome but one of C4 gene markers and 5 sporadic EIEC lineages specific genes were located on plasmid. None of the cluster-specific gene markers were contiguous in the genomes. The location of these cluster-specific gene markers was determined by BLASTN against representative complete genomes of *Shigella*/EIEC containing gene features downloaded from NCBI GenBank. In those cluster or sporadic lineages with no representative complete genome specific gene markers were named using their cluster or sporadic EIEC lineage followed by the cluster or lineage number. For example, C7 specific gene marker was named “C7 specific gene”.

The functional characterization of these specific gene markers were identified from RAST annotation (74). For 37 cluster-specific gene markers, 22 had known functions and 15 encoded hypothetical proteins with unknown functions, while 11 sporadic EIEC lineages specific gene markers were identified with known functions and 43 were hypothetical proteins with unknown functions. The location and functions of specific gene markers are listed in Table S3.

Validation of cluster-specific gene markers

The ability of cluster-specific gene markers to correctly assign *Shigella*/EIEC isolates was evaluate with 15,501 *Shigella*/EIEC isolates in the validation dataset, 12,743 isolates from non-enteroinvasive *E. coli* control database.

Using cluster-specific gene markers, 15,443 of the 15,501 (99.63%) *Shigella*/EIEC isolates were correctly assigned to clusters which included 15,337 *Shigella* isolates, 102 EIEC isolates, 4 sporadic EIEC isolates, and 38 (0.24%) isolates with more than one clusters. Twenty of the 15,501 (0.13%) *Shigella*/EIEC isolates were not assigned to any of identified clusters.

To confirm the assignment of cluster-specific gene markers, we constructed a “validation” phylogenetic tree (Fig. S4) using 1,159 representative isolates from the validation dataset and a subset of 485 isolates from each cluster from the identification dataset. Isolates that grouped with known cluster isolates (from identification dataset) with strong bootstrap support were assigned to that cluster. All 1,159 isolates were grouped into known clusters on the validation phylogenetic tree. The cluster-specific gene markers assignments were entirely consistent with cluster assignments by phylogenetic tree.

We tested cluster-specific gene markers with the 12,743 non-enteroinvasive *E. coli* isolates. The *Shigella*/EIEC cluster-specific gene markers were highly specific with specificity varying from 98.8% to 100% for cluster-specific genes and 97.02% to 100% for sporadic EIEC specific genes. Details are listed in Table S4.

Development an automated pipeline for molecular serotyping of *Shigella*/EIEC

Above results showed that cluster-specific gene markers were sensitive and specific and can distinguish *Shigella* and EIEC isolates. We therefore used these genes combined with established *Shigella*/EIEC serotype specific O antigen and H antigen genes to develop an automated pipeline for *in silico* serotyping of *Shigella*/EIEC (Fig. 2).

The pipeline is named *Shigella* EIEC Cluster Enhanced Serotype Finder (ShigEiFinder). ShigEiFinder can process either paired end Illumina sequencing reads or assembled genomes (https://github.com/LanLab/ShigEiFinder). ShigEiFinder classifies isolates into Non-*Shigella*/EIEC*,* *Shigella* or EIEC clusters based on the presence of *ipaH* gene, number of virulence genes, cluster specific genes. The "Not *Shigella*/EIEC" assignment was determined by the absence of *ipaH* gene, virulence genes (>25/38) and cluster-specific gene markers. The "*Shigella* or EIEC clusters" assignments were made based on the presence of *ipaH* gene, and/or more than 25 virulence genes together with the presence of any of cluster-specific gene markers or marker set, whereas the presence of *ipaH* gene and/or more than 25 virulence genes with absence of any of cluster-specific gene markers were assigned as "*Shigella*/EIEC unclustered".

Shigella and EIEC isolates were differentiated and serotypes were assigned after cluster assignment. ShigEiFinder predicts a serotype through examining the presence of any of established *Shigella* serotype specific O antigen and modification genes and E. coli O and H antigen genes that differentiate the serotypes as ShigaTyper and SerotypeFinder (10, 63). A "novel serotype" is assigned if no match to known serotypes.

Two pairs of *Shigella* serotypes, SB1/SB20 and SB6/SB10, are known to be difficult to differentiate as they share identical O antigen genes (10, 46, 75). ShigaTyper used a heparinase gene for the differentiation of SB20 from SB1 and *wbaM* gene for the separation of SB6 from SB10. We found that fragments of the heparinase and *wbaM* genes may be present in other serotypes and cannot accurately differentiate SB1/SB20 and SB6/SB10. We found a SB20 specific gene which encoded hypothetical proteins with unknown functions and located on a plasmid by comparative genomic analysis of all isolates in C1 accessory genome. The SB20 specific gene can reliably differentiate SB20 from SB1and also one SNP each in *wzx* and *wzy* genes that can differentiate SB6 from SB10. We used these differences (Data S1) in ShigEiFinder for the prediction of these serotypes.

The accuracy and specificity of ShigEiFinder in cluster typing

The accuracy of ShigEiFinder was tested with 1,969 isolates (1,969 assembled genomes and 1,951 Illumina reads [note no reads available for 18 EIEC isolates from NCBI) from the identification dataset and 15,501 isolates from the validation dataset. The results are listed in Table 3.

ShigEiFinder was able to assign 99.54% and 99.28% of the isolates in the identification dataset to clusters for assembled genomes and reads mapping respectively. The accuracy was 99.70% and 99.81% for assembled genomes and reads mapping respectively when applied to the validation dataset. Discrepancies were observed between assembled genomes and reads mapping (Table 3). There were more isolates assigned to "*Shigella*/EIEC unclustered" in reads mapping, in contrast there were more isolates assigned to multiple clusters in genome assemblies. The specificity of ShigEiFinder was 99.40% for assembled genomes and 99.38% for reads mapping when evaluated with 12,743 non-*Shigella*/EIEC *E. coli* isolates. An additional 2 isolates were detected as sporadic EIEC lineages by reads mapping.

Comparison of ShigEiFinder and ShigaTyper

To demonstrate ShigEiFinder for differentiation of *Shigella* from EIEC and enhancement of cluster based serotyping, the comparison of reads mapping results between ShigEiFinder and the existing *in silico Shigella* identification pipeline ShigaTyper (10) was performed with 488 isolates used in ShigaTyper and 15,501 isolates from *Shigella*/EIEC validation dataset used in the present study.

The 488 isolates used in ShigaTyper consisted of 23 other species, 45 *E. coli* isolates and 420 *Shigella* isolates. ShigEiFinder identified 23 other species isolates and 453 out of 465 *E. coli* and *Shigella* isolates correctly, in agreement with ShigaTyper assignment. ShigEiFinder also assigned the remaining 12 *E. coli* and *Shigella* isolates including 3 EIEC isolates and 9 untypable (either multiple *wzx* or no *wzx* genes found) isolates by ShigaTyper to *Shigella*/EIEC clusters.

ShigEiFinder assigned 15,471 of 15,501 *Shigella*/EIEC isolates to *Shigella* or EIEC clusters and then to a serotype. The accuracy of ShigEiFinder to correctly assign isolates to *Shigella* or EIEC clusters was 99.81% (15,471/15,501). By contrast, ShigaTyper assigned 7,277 isolates (46.95%) to *Shigella*, 7.976 isolates (51.45%) to EIEC, 177 (1.14%) isolates to multiple *wzx* genes and failed to type 71 (0.46%) isolates.

The predicted serotype of 7,277 (46.96%) *Shigella* isolates by ShigaTyper agreed with the results of ShigEiFinder. For 8,224 isolates typed as EIEC or untypable by ShigaTyper, 99.73% (8,202/8,224) of the isolates were assigned to *Shigella* or EIEC clusters by ShigEiFinder (Table 4). Of these isolates, the majority belonged to SS, SD1 and SF which were erroneously predicted as EIEC by ShigaTyper.

Discussion

Shigella and EIEC cause human bacillary dysentery with similar invasion mechanisms, however the pathogenicity of these 2 groups varies (8, 43). The prevalence of each of the four *Shigella* "species" also varies (11-13). Differentiation of *Shigella* and EIEC from each other is important for epidemiologic and diagnostic investigations. However, their similar physiological, biochemical and genetic characteristics make this differentiation difficult.

Determining phylogenetic clusters for better separation *Shigella* isolates from EIEC

From phylogenetic perspective, *Shigella* and EIEC strains consisted of multiple phylogenetic lineages derived from commensal *E. coli*, which do not reflect the nomenclature of *Shigella* and EIEC (23, 25, 26, 28, 38, 41). In the present study, we identified all phylogenetic clusters of *Shigella* and EIEC through large scale examination of publicly available genomes. Phylogenetic results demonstrated that *Shigella* isolates had at least 10 clusters while EIEC isolates had at least 7 clusters. The 10 *Shigella* clusters included the 7 previously defined lineages including 3 major clusters (C1, C2 and C3) and 5 outliers (SD1, SD8, SD10, SB13 and SS) (25) and 2 newly identified clusters (SB12 and SB13-atypical). The 7 EIEC clusters consisted of 4 previously defined EIEC clusters (C4, C5, C6 and C7) (26) and 3 newly identified EIEC clusters (C8 EIEC O96:H19, C9 EIEC O8:H19 and C10 EIEC O135:H30).

Our WGS-based phylogeny provided high resolution for assigning *Shigella* and EIEC isolates to clusters. Several serotypes that are currently increasing in frequency (SB19, SB20, SD14, SD15, SD provisional serotype 96-626) (76-79) were assigned to clusters and five new clusters/outliers were identified. SB13 isolates in this study formed two known lineages. One lineage was located outside of *Shigella*/EIEC clusters and represented the outlier SB13 which is in fact belonging to the newly defined species *E. albertii* (25, 26, 38, 39). The second lineage was with *E. coli*, and was defined as atypical SB13 previously (39). The newly identified *Shigella* outlier SB12 was previously grouped into C3 based on housekeeping gene trees (25, 38) but was seen as outliers in two other studies (28, 56).

Newly identified clusters C8 (EIEC O96:H19) and C9 (EIEC O8:H19) represented the emergence of novel EIEC serotypes. A recent study revealed that EIEC serotype O96:H19 (C8) could be the result of a recent acquisition of the invasion plasmid by commensal E. coli (80). The EIEC serotype O8:H19 (C9) had not been reported previously.

Apart from the 17 major and outlier clusters of *Shigella* and EIEC, the presence of 53 sporadic EIEC lineages indicated greater genetic diversity than has been observed previously. Isolates belonging to these sporadic EIEC groups were more closely related to non-enteroinvasive E. coli isolates than to major *Shigella*/EIEC lineages. However, 41 of the isolates representing 38 sporadic EIEC lineages that carried pINV. *Shigella* and EIEC both carry the *Shigella* virulence plasmid pINV which is vital for virulence and distinguishes *Shigella*/EIEC from other *E. coli* (24, 33, 68). Therefore, these isolates may represent recently formed EIEC lineages through acquisition of the pINV. The remaining 18 isolates contained the *ipaH* gene but may or may not carry pINV. It is possible that these strains carried very low copy number of pINV or the pINV plasmid was lost during culture.

Highly sensitive and specific cluster-specific gene markers for differentiation of *Shigella* and EIEC isolates

Several studies have identified phylogenetic related genomic markers for discrimination of *Shigella* and EIEC and between *Shigella* species (23, 27, 28, 41, 55, 56). However, these phylogenetic analyses were performed only with a small number of genomes (23, 28, 55). In addition, non-invasive E. coli isolates were included in some of the phylogenetic clusters identified (28) which led to non-invasive E. coli isolates being identified by the markers.

We identified cluster-specific gene markers for each respectively clusters which were only composed of *Shigella* or EIEC isolates. Sets of cluster-specific gene markers were identified for those clusters where no single suitable marker is present. The combination of genes enhances the specificity of cluster-specific gene markers as demonstrated by the 100% sensitivity and very high specificity in this analysis (Table 2). Genes specific to each of the 53 sporadic EIEC lineages were also identified and they were sensitive and specific, although it should be noted that these values are based on very small sample sizes.

The cluster-specific gene markers or marker sets can be used to differentiate *Shigella*/EIEC from non-enteroinvasive *E. coli* independent of *ipaH* gene. The *ipaH* gene as a molecular target has been used to differentiate *Shigella* and EIEC from non-enteroinvasive E. coli (24, 43-45). In our study, the cluster-specific gene markers were specific to *Shigella*/EIEC with 98.8% to 100% specificity when evaluated on non-enteroinvasive *E. coli* control database, giving us the confidence that the cluster-specific genes or sets are robust markers to identify *Shigella*/EIEC. 53 sporadic EIEC lineage specific gene markers also have very high specificity (97.02% to 100%) against non-enteroinvasive *E. coli* control database.

The cluster-specific gene markers or marker sets are able to assign *Shigella*/EIEC isolates correctly in 99.63% of cases and can clearly distinguish *Shigella* isolates from EIEC when applied to the validation dataset. While ShigaTyper assigned 46.95% isolates to *Shigella* and 51.45% isolates to EIEC in the same dataset we tested, leading to a large proportion of isolates incorrectly assigned. The majority of the isolates predicted as EIEC by ShigaTyper were SS or SD1 as they belonged to SS and SD1 specific STs and were positive to a set of SS or SD1 specific gene markers and grouped into SS or SD1 cluster on our phylogenetic tree. The genes used in ShigaTyper were SS specific marker Ss_methylase gene (81, 82) together with SS O antigen wzx gene. However, SS specific marker Ss_methylase gene was found in other *Shigella* serotypes and EIEC (10) and SS O antigen wzx gene were located on a plasmid which is frequently lost (83). Similarly, the SD1 O antigen genes used in ShigaTyper were plasmid-borne which may also lead to inconsistent detection (84, 85). A previous study identified 6 loci to distinguish EIEC from *Shigella* (23). We searched the 6 loci against our *Shigella*/EIEC database and found that some *Shigella* isolates were misidentified as EIEC isolates, such as SD8 isolates incorrectly identified as EIEC subtype 13. Our cluster-specific genes can differentiate SD8 isolates from EIEC with 100% accuracy. Therefore, the cluster-specific gene markers marker sets provided nearly perfect differentiation of *Shigella* from EIEC.

The cluster-specific gene markers or marker sets are able to differentiate SS and SF (with exception of SF6) from SB and SD. SF and SS are the major cause of *Shigella* infections, accounting for up to 89.6% annual cases (11-13). Differentiation of SS and SF isolates from SB and SD is also beneficial for diagnosis and surveillance. A recent study identified "species" specific markers for the detection of each of the four *Shigella* "species" and validated with only one isolate per species (55). A molecular algorithm based on *Shigella* O antigen genes can detect 85% of SF isolates (52). In contrast, a set of SF specific genes in our study can correctly identify SF isolates with 99.62% accuracy.

The cluster-specific gene markers or marker sets can also assign *Shigella*/EIEC isolates to serotype level if the cluster has single serotype such as SD1, SD8, SD10, SB13, SB12, EIEC O144:H25 (C7), EIEC O96:H19 (C8), EIEC O8:H19 (C9) and EIEC O135:H30 (C10). The remaining EIEC, SF, SB and SD serotypes were distributed over the major clusters C4-6, C3, C1 and C2 respectively. Cluster-specific gene markers combined with serotype associated O antigen and modification genes can further identify these isolates to serotype level.

Cluster-specific gene marker based ShigEiFinder can accurately type *Shigella* and EIEC

To facilitate the use of cluster-specific gene markers or marker sets for typing, we developed an automated pipeline, ShigEiFinder, for *in silico* molecular serotyping of *Shigella*/EIEC. ShigEiFinder provided *Shigella*/ EIEC differentiation as well as serotype prediction by yielding "presence or absence" of cluster-specific gene markers or marker sets combined with *Shigella*/EIEC O antigen genes and modification genes in a query isolate (either reads or assembled genomes). We showed 99.70% and 99.81% accuracy to assign isolates to the correct clusters from 15,501 *Shigella*/EIEC isolates in validation dataset for the assembled genomes and reads mapping respectively. In contrast, the existing *in silico Shigella* serotyping pipeline ShigaTyper had 46.95% accuracy for reads mapping when tested with the same validation dataset, with 51.45% of isolates in validation dataset being predicted as EIEC by ShigaTyper.

The genetic determinants used in ShigaTyper for differentiation of *Shigella* from EIEC and identification of SS were *lacY*, *cadA*, *Ss_methylase,* SS and SD1 O antigen *wzx* genes (10). As discussed above some of these genes were found to be non-specific in this study. Compared with ShigaTyper, the cluster-specific gene markers used in ShigEiFinder for identification of *Shigella* and EIEC provided higher discriminatory power than ShigaTyper. ShigEiFinder also provided a high specificity with 99.40% for assembled genomes and 99.38% for reads mapping.

ShigEiFinder can differentiate *Shigella* isolates from EIEC and distinguish SS and SF (with exception of SF6) isolates from SB and SD accurately. It also can identify SD1 isolates directly. ShigEiFinder was able to serotype over 59 *Shigella* serotypes and 22 EIEC serotypes. Therefore, ShigEiFinder will be useful for clinical, epidemiological and diagnostic investigations and the cluster-specific gene markers identified could be adapted for metagenomics or culture independent typing.

Conclusion

This study analysed over 17,000 publicly available *Shigella*/EIEC genomes and identified 10 clusters of *Shigella*, 7 clusters of EIEC and 53 sporadic types of EIEC. Cluster-specific gene markers or marker sets for the 17 major clusters and 53 sporadic types were identified and found to be valuable for *in silico* typing. We additionally developed a freely available *in silico* serotyping pipeline incorporating the cluster-specific gene markers to facilitate serotyping of *Shigella*/EIEC isolates using genome sequences with very high specificity and sensitivity.

Funding information

This work was funded in part by a National Health and Medical Research Council project grant (grant number 1129713) and an Australian Research Council Discovery Grant (DP170101917).

Acknowledgements

The authors thank Duncan Smith and Robin Heron from UNSW Research Technology Services for computing assistance.

Author contributions

Conceptualization: R.L, M.P.; Investigation: X.Z., M.P., T.N., S.K.; Methodology: M.P., R.L. Writing – original draft: X.Z.; Writing – review and editing: M.P., R.L.

Conflicts of interest

The authors declare that there are no conflicts of interest.

Data bibliography

Zhang X, Payne M, Nguyen T, Kaur S, Lan R. All the sequencing data generated within this study, NCBI BioProject number (PRJNA692536).

Abbreviations

SS, *Shigella sonnei*; SF, *Shigella flexneri*; SB, *Shigella boydii*; SD, *Shigella dysenteriae*; EIEC, Enteroinvasive *Escherichia coli*; NCBI SRA, National Center for Biotechnology Information Sequence Read Archive; ST, sequence type; rST, ribosomal ST; MLST, Multilocus sequence typing; rMLST, Ribosomal MLST; ECOR, *Escherichia coli* reference collection; WGS, whole-genome sequencing; TP, true positive; FN, false negative; FP, false positive; HK, House Keeping.

References

1. DuPont HL, Levine MM, Hornick RB, Formal SBJTJoid. Inoculum size in shigellosis and implications for expected mode of transmission. *The Journal of infectious diseases*. 1989;159(6):1126-8.

2. Troeger C, Forouzanfar M, Rao PC, Khalil I, Brown A, Reiner Jr RC, et al. Estimates of global, regional, and national morbidity, mortality, and aetiologies of diarrhoeal diseases: a systematic analysis for the Global Burden of Disease Study 2015. *The Lancet Infectious Diseases*. 2017;17(9):909-48.

3. World Health Organization. Guidelines for the control of shigellosis, including epidemics due to *Shigella dysenteriae* type 1. 2005.

4. Kirk MD, Pires SM, Black RE, Caipo M, Crump JA, Devleesschauwer B, et al. World Health Organization estimates of the global and regional disease burden of 22 foodborne bacterial, protozoal, and viral diseases, 2010: a data synthesis. *PLoS medicine*. 2015;12(12):e1001921.

5. Brengi SP, Sun Q, Bolaños H, Duarte F, Jenkins C, Pichel M, et al. PCR-based method for *Shigella flexneri* serotyping: international multicenter validation. *Journal of clinical microbiology*. 2019;57(4):e01592-18.

6. Kotloff KL, Nataro JP, Blackwelder WC, Nasrin D, Farag TH, Panchalingam S, et al. Burden and aetiology of diarrhoeal disease in infants and young children in developing countries (the Global Enteric Multicenter Study, GEMS): a prospective, case-control study. *The Lancet* 2013;382(9888):209-22.

7. Khalil IA, Troeger C, Blacker BF, Rao PC, Brown A, Atherly DE, et al. Morbidity and mortality due to *shigella* and enterotoxigenic *Escherichia coli* diarrhoea: the Global Burden of Disease Study 1990–2016. *The Lancet Infectious Diseases*. 2018;18(11):1229-40.

8. van den Beld MJ, Warmelink E, Friedrich AW, Reubsaet FA, Schipper M, de Boer RF, et al. Incidence, clinical implications and impact on public health of infections with *Shigella spp*. and entero-invasive *Escherichia coli* (EIEC): results of a multicenter cross-sectional study in the Netherlands during 2016–2017. *BMC Infectious Diseases*. 2019;19(1):1037.

9. Edwards PR, Ewing WH. Identification of enterobacteriaceae. Identification of Enterobacteriaceae.. 1972(Third edition).

10. Wu Y, Lau HK, Lee T, Lau DK, Payne J. *In silico* serotyping based on Whole-Genome sequencing improves the accuracy of *shigella* identification. *Appl Environ Microbiol*. 2019;85(7):e00165-19.

11. The HC, Thanh DP, Holt KE, Thomson NR, Baker SJNRM. The genomic signatures of *Shigella* evolution, adaptation and geographical spread. *Nature Reviews Microbiology*. 2016;14(4):235.

12. Livio S, Strockbine NA, Panchalingam S, Tennant SM, Barry EM, Marohn ME, et al. *Shigella* isolates from the global enteric multicenter study inform vaccine development. *Clinical Infectious Diseases*. 2014;59(7):933-41.

13. Group OW. Monitoring the incidence and causes of diseases potentially transmitted by food in Australia: Annual report of the OzFoodNet network, 2011. *Communicable diseases intelligence quarterly report*. 2015;39(2):E236.

14. Connor TR, Barker CR, Baker KS, Weill F-X, 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.

15. Taylor D, Echeverria P, Sethabutr O, Pitarangsi C, Leksomboon U, Blacklow N, et al. Clinical and microbiologic features of *Shigella* and enteroinvasive *Escherichia coli* infections detected by DNA hybridization. *Journal of clinical microbiology*. 1988;26(7):1362-6.

16. Levine MM. *Escherichia coli* that cause diarrhea: enterotoxigenic, enteropathogenic, enteroinvasive, enterohemorrhagic, and enteroadherent. *Journal of infectious Diseases*. 1987 Mar 1;155(3):377-89.

17. Tai AY, Easton M, Encena J, Rotty J, Valcanis M, Howden BP, et al. A review of the public health management of shigellosis in Australia in the era of culture-independent diagnostic testing. *Australian New Zealand journal of public health*. 2016;40(6):588-91.

18. Gomes TA, Elias WP, Scaletsky IC, Guth BE, Rodrigues JF, Piazza RM, et al. Diarrheagenic *Escherichia coli*. *brazilian journal of microbiology* 2016;47:3-30.

19. Pasqua M, Michelacci V, Di Martino ML, Tozzoli R, Grossi M, Colonna B, et al. The Intriguing Evolutionary Journey of Enteroinvasive *E. coli* (EIEC) toward Pathogenicity. *Frontiers in microbiology*. 2017;8:2390.

20. Herzig CT, Fleischauer AT, Lackey B, Lee N, Lawson T, Moore ZS, et al. Notes from the Field: Enteroinvasive *Escherichia coli* Outbreak Associated with a Potluck Party—North Carolina, June–July 2018. *Morbidity and Mortality Weekly Report*. 2019;68(7):183.

21. Pettengill EA, Hoffmann M, Binet R, Roberts RJ, Payne J, Allard M, et al. Complete genome sequence of enteroinvasive *Escherichia coli* O96: H19 associated with a severe foodborne outbreak. *Genome Announcements*. 2015;3(4):e00883-15.

22. Escher M, Scavia G, Morabito S, Tozzoli R, Maugliani A, Cantoni S, et al. A severe foodborne outbreak of diarrhoea linked to a canteen in Italy caused by enteroinvasive *Escherichia coli*, an uncommon agent. *Epidemiology and infecttion*. 2014;142(12):2559-66.

23. Dhakal R, Wang Q, Lan R, Howard P, Sintchenko VJJomm. Novel multiplex PCR assay for identification and subtyping of enteroinvasive *Escherichia coli* and differentiation from *Shigella* based on target genes selected by comparative genomics. *Journal of medical microbiology*. 2018;67(9):1257-64.

24. Van den Beld M, Reubsaet FJEjocm, diseases i. Differentiation between *Shigella*, enteroinvasive *Escherichia coli* (EIEC) and noninvasive *Escherichia coli*. *European Journal of Clinical Microbiology & Infectious Diseases*. 2012;31(6):899-904.

25. Pupo GM, Lan R, Reeves PRJPotNAoS. Multiple independent origins of *Shigella* clones of *Escherichia coli* and convergent evolution of many of their characteristics. *Proceedings of the National Academy of Sciences* 2000;97(19):10567-72.

26. Lan R, Alles MC, Donohoe K, Martinez MB, Reeves PRJI, immunity. Molecular evolutionary relationships of enteroinvasive *Escherichia coli* and *Shigella spp*. *Infection and immunity*. 2004;72(9):5080-8.

27. Sahl JW, Morris CR, Emberger J, Fraser CM, Ochieng JB, Juma J, et al. Defining the phylogenomics of *Shigella* species: a pathway to diagnostics. *Journal of clinical microbiology*. 2015;53(3):951-60.

28. Pettengill EA, Pettengill JB, Binet RJFim. 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. *Frontiers in microbiology*. 2016;6:1573.

29. Cheasty T, Rowe BJJocm. Antigenic relationships between the enteroinvasive *Escherichia coli* O antigens O28ac, O112ac, O124, O136, O143, O144, O152, and O164 and *Shigella* O antigens. *Journal of clinical microbiology*. 1983;17(4):681-4.

30. Landersjö C, Weintraub A, Widmalm GJCr. Structure determination of the O-antigen polysaccharide from the enteroinvasive *Escherichia coli* (EIEC) O143 by component analysis and NMR spectroscopy. *Carbohydrate research*. 1996;291:209-16.

31. Linnerborg M, Weintraub A, Widmalm GJEjob. Structural studies of the O-antigen polysaccharide from the enteroinvasive *Escherichia coli* O164 cross-reacting with *Shigella* dysenteriae type 3. *European journal of biochemistry*. 1999;266(2):460-6.

32. Sansonetti P, d'Hauteville H, Écobichon Ct, Pourcel C, editors. Molecular comparison of virulence plasmids in *Shigella* and enteroinvasive *Escherichia coli*. Annales de l'Institut Pasteur/Microbiologie; 1983: Elsevier.

33. Lan R, Lumb B, Ryan D, Reeves PRJI, immunity. Molecular Evolution of Large Virulence Plasmid in *Shigella* Clones and Enteroinvasive *Escherichia coli*. *Infection and immunity*. 2001;69(10):6303-9.

34. Venkatesan MM, Buysse JM, Kopecko DJJJoCM. Use of *Shigella flexneri* ipaC and ipaH gene sequences for the general identification of *Shigella spp*. and enteroinvasive *Escherichia coli*. *Journal of Clinical Microbiology*. 1989;27(12):2687-91.

35. Hale TLJM, Reviews MB. Genetic basis of virulence in *Shigella* species. Microbiology *Microbiology and Molecular Biology Reviews*. 1991;55(2):206-24.

36. Jin Q, Yuan Z, Xu J, Wang Y, Shen Y, Lu W, et al. Genome sequence of *Shigella flexneri* 2a: insights into pathogenicity through comparison with genomes of *Escherichia coli* K12 and O157. *Nucleic acids research*. 2002;30(20):4432-41.

37. Yang F, Yang J, Zhang X, Chen L, Jiang Y, Yan Y, et al. Genome dynamics and diversity of *Shigella* species, the etiologic agents of bacillary dysentery. *Nucleic acids research*. 2005;33(19):6445-58.

38. Yang J, Nie H, Chen L, Zhang X, Yang F, Xu X, et al. Revisiting the molecular evolutionary history of *Shigella spp*. *Journal of molecular evolution*. 2007;64(1):71-9.

39. Hyma KE, Lacher DW, Nelson AM, Bumbaugh AC, Janda JM, Strockbine NA, et al. Evolutionary genetics of a new pathogenic *Escherichia* species: *Escherichia albertii* and related *Shigella boydii* strains. *Journal of bacteriology*. 2005;187(2):619-28.

40. Walters LL, Raterman EL, Grys TE, Welch RA. Atypical *Shigella boydii* 13 encodes virulence factors seen in attaching and effacing *Escherichia coli*. *FEMS Microbiol Lett*. 2012;328(1):20-5.

41. Hazen TH, Leonard SR, Lampel KA, Lacher DW, Maurelli AT, Rasko DAJI, et al. Investigating the relatedness of enteroinvasive *Escherichia coli* to other *E. coli* and *Shigella* isolates by using comparative genomics. *Infection and immunity*. 2016;84(8):2362-71.

42. Silva RM, Toledo M, Trabulsi LRJJoCM. Biochemical and cultural characteristics of invasive *Escherichia coli*. *Journal of Clinical Microbiology*. 1980;11(5):441-4.

43. van den Beld MJ, Friedrich AW, van Zanten E, Reubsaet FA, Kooistra-Smid MA, Rossen JWJJomm. Multicenter evaluation of molecular and culture-dependent diagnostics for *Shigella* species and Entero-invasive *Escherichia coli* in the Netherlands. *Journal of microbiological methods*. 2016;131:10-5.

44. Van Lint P, De Witte E, Ursi J, Van Herendael B, Van Schaeren JJDm, disease i. A screening algorithm for diagnosing bacterial gastroenteritis by real-time PCR in combination with guided culture. *Diagnostic microbiology*. 2016;85(2):255-9.

45. De Boer RF, Ott A, Kesztyüs B, Kooistra-Smid AMJJocm. Improved detection of five major gastrointestinal pathogens by use of a molecular screening approach. *Journal of clinical microbiology*. 2010;48(11):4140-6.

46. Liu B, Knirel YA, Feng L, Perepelov AV, Senchenkova SyN, Wang Q, et al. Structure and genetics of *Shigella* O antigens. *FEMS microbiology reviews*. 2008;32(4):627-53.

47. Wattiau P, Boland C, Bertrand S. Methodologies for *Salmonella enterica ssp enterica* subtyping: gold standards and alternatives. *Applied and environmental microbiology*. 2011:AEM. 05527-11.

48. Cai H, Lu L, Muckle C, Prescott J, Chen S. Development of a novel protein microarray method for serotyping *Salmonella enterica* strains. *Journal of clinical microbiology*. 2005;43(7):3427-30.

49. van der Ploeg CA, Rogé AD, Bordagorría XL, de Urquiza MT, Castillo ABC, Bruno SB. Design of Two Multiplex PCR Assays for Serotyping *Shigella flexneri*. *Foodborne pathogens and disease*. 2018;15(1):33-8.

50. Sun Q, Lan R, Wang Y, Zhao A, Zhang S, Wang J, et al. Development of a multiplex PCR assay targeting O-antigen modification genes for molecular serotyping of *Shigella flexneri*. *Journal of clinical microbiology*. 2011;49(11):3766-70.

51. Li Y, Cao B, Liu B, Liu D, Gao Q, Peng X, et al. Molecular detection of all 34 distinct O-antigen forms of *Shigella*. J Med Microbiol. 2009;58(Pt 1):69-81.

52. van den Beld MJ, de Boer RF, Reubsaet FA, Rossen JW, Zhou K, Kuiling S, et al. Evaluation of a Culture-Dependent Algorithm and a Molecular Algorithm for Identification of

Shigella spp., *Escherichia coli*, and Enteroinvasive *E. coli*. *Journal of clinical microbiology*. 2018;56(10):e00510-18.

53. Pavlovic M, Luze A, Konrad R, Berger A, Sing A, Busch U, et al. Development of a duplex real-time PCR for differentiation between *E. coli* and *Shigella spp*. *Journal of applied microbiology*. 2011;110(5):1245-51.

54. Løbersli I, Wester AL, Kristiansen Å, Brandal LTJEJoM, Immunology. Molecular differentiation of *Shigella spp*. from enteroinvasive *E. coli*. *European Journal of Microbiology*. 2016;6(3):197-205.

55. Kim H-J, Ryu J-O, Song J-Y, Kim H-YJFp, disease. Multiplex polymerase chain reaction for identification of shigellae and four *Shigella* species using novel genetic markers screened by comparative genomics. *Foodborne pathogens*. 2017;14(7):400-6.

56. Chattaway MA, Schaefer U, Tewolde R, Dallman TJ, Jenkins CJJocm. Identification of *Escherichia coli* and *Shigella* species from whole-genome sequences. *Journal of clinical microbiology*. 2017;55(2):616-23.

57. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods*. 2017;14(4):417-9.

58. Wood DE, Salzberg SLJGb. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome biology*. 2014;15(3):R46.

59. Alikhan N-F, Zhou Z, Sergeant MJ, Achtman M. A genomic overview of the population structure of *Salmonella*. *PLoS genetics*. 2018;14(4):e1007261.

60. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of computational biology*. 2012;19(5):455-77.

61. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics*. 2013;29(8):1072-5.

62. Jolley KA, Maiden MCJ. BIGSdb: Scalable analysis of bacterial genome variation at the population level. *BMC Bioinformatics*. 2010;11(1):595.

63. Joensen KG, Tetzschner AM, Iguchi A, Aarestrup FM, Scheutz FJJocm. Rapid and easy *in silico* serotyping of *Escherichia coli* isolates by use of whole-genome sequencing data. *Journal of clinical microbiology*. 2015;53(8):2410-26.

64. Hu D, Liu B, Wang L, Reeves PR. Living Trees: High-Quality Reproducible and Reusable Construction of Bacterial Phylogenetic Trees. *Molecular Biology and Evolution*. 2019;37(2):563-75.

65. Letunic I, Bork PJNar. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic acids research*. 2019;47(W1):W256-W9.

66. Zhou Z, Alikhan N-F, Sergeant MJ, Luhmann N, Vaz C, Francisco AP, et al. GrapeTree: visualization of core genomic relationships among 100,000 bacterial pathogens. *Genome research*. 2018;28(9):1395-404.

67. Li HJapa. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv*. 2013.

68. Buchrieser C, Glaser P, Rusniok C, Nedjari H, d'Hauteville H, Kunst F, et al. The virulence plasmid pWR100 and the repertoire of proteins secreted by the type III secretion apparatus of *Shigella flexneri*. *Molecular microbiology*. 2000;38(4):760-71.

69. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 2009;25(16):2078-9.

70. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*. 2010;26(6):841-2.

71. Seemann T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics*. 2014;30(14):2068-9.

72. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MT, et al. Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics*. 2015;31(22):3691-3.

73. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. *BMC bioinformatics*. 2009;10(1):421.

74. Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, et al. The RAST Server: rapid annotations using subsystems technology. *BMC genomics*. 2008;9(1):75.

75. Sof'ya NS, Feng L, Yang J, Shashkov AS, Cheng J, Liu D, et al. Structural and genetic characterization of the *Shigella boydi*i type 10 and type 6 O antigens. *Journal of bacteriology*. 2005;187(7):2551-4.

76. Ansaruzzaman M, Kibriya A, Rahman A, Neogi P, Faruque A, Rowe B, et al. Detection of provisional serovars of *Shigella* dysenteriae and designation as *S*. dysenteriae serotypes 14 and 15. *Journal of clinical microbiology*. 1995;33(5):1423-5.

77. Balows AJDM, Disease I. Manual of clinical microbiology 8th edition: PR Murray, EJ Baron, JH Jorgenson, MA Pfaller, and RH Yolken, eds., ASM Press, 2003, 2113 pages, 2 vol, 2003+ subject & author indices, ISBN: 1-555810255-4, US $189.95. *Diagnostic Microbiology*. 2003;47(4):625.

78. Woodward DL, Clark CG, Caldeira RA, Ahmed R, Soule G, Bryden L, et al. Identification and characterization of *Shigella boydii* 20 serovar nov., a new and emerging *Shigella* serotype. *J Med Microbiol*. 2005;54(8):741-8.

79. Kim J, Lindsey RL, Garcia-Toledo L, Loparev VN, Rowe LA, Batra D, et al. High-quality whole-genome sequences for 59 historical *Shigella* strains generated with PacBio sequencing. *Genome Announcements*. 2018;6(15).

80. Michelacci V, Prosseda G, Maugliani A, Tozzoli R, Sanchez S, Herrera-León S, et al. Characterization of an emergent clone of enteroinvasive *Escherichia coli* circulating in Europe. *Clinical Microbiology*. 2016;22(3):287. e11-. e19.

81. Liu J, Platts-Mills JA, Juma J, Kabir F, Nkeze J, Okoi C, et al. Use of quantitative molecular diagnostic methods to identify causes of diarrhoea in children: a reanalysis of the GEMS case-control study. *The Lancet*. 2016;388(10051):1291-301.

82. Cho MS, Ahn T-Y, Joh K, Kwon O-S, Jheong W-H, Park DSJJMB. A novel marker for the species-specific detection and quantitation of *Shigella sonnei* by targeting a methylase gene. *J Microbiol Biotechnol*. 2012;22(8):1113-7.

83. Sansonetti PJ, Kopecko DJ, Formal SBJI, immunity. *Shigella sonnei* plasmids: evidence that a large plasmid is necessary for virulence. *Infection and immunity*. 1981;34(1):75-83.

84. Feng L, Perepelov AV, Zhao G, Shevelev SD, Wang Q, Sof'ya NS, et al. Structural and genetic evidence that the *Escherichia coli* O148 O antigen is the precursor of the *Shigella dysenteriae* type 1 O antigen and identification of a glucosyltransferase gene. *Microbiology*. 2007;153(1):139-47.

85. Göhmann S, Manning P, Alpert C-A, Walker M, Timmis KJMp. Lipopolysaccharide O-antigen biosynthesis in *Shigella* dysenteriae serotype 1: analysis of the plasmid-carried rfp determinant. *Microbial pathogenesis*. 1994;16(1):53-64.

Table 1: The summary of identified *Shigella*/EIEC clusters and outliers in identification dataset

Clusters (no of serotypes)[#]	No of isolates	No. STs	No. rSTs	Serotypes
C1 (25)	288	36	166	SB1-4, SB6, SB8, SB10, SB14, SB18, SB11[b], SB19-20[b]; SD3-7, SD9, SD11-13, SD14-15[*], SD-96-265[*]; SF6
C2 (9)	101	19	56	SB5, SB7, SB9, SB11, SB15, SB16, SB17; SD2, SD-E670-74[b]; SD2
C3 (20)	744	81	437	SF1a, SF1b, SF1c (7a), SF2a, SF2b, SF3a, SF3b, SF4a, SF4av, SF4b, SF4bv, SF5a, SF5b, SF7b, SFX, SFXv (4c), SFY, SFYv, SF novel serotype; SB-E1621-54*
C4 (9)	51	6	21	O28ac:H-, O28ac:H7, O136:H7, O164:H-, O164:H7, O29:H4, O173:H7, O124:H7, O132:H7[*]
C5 (6)	62	4	15	O121:H30, O124:H30, O164:H30, O132:H21, O152:H30, O152:H-
C6 (3)	20	2	6	O143:H26, O167:H26, O112ac:H26[b]
C7	10	1	3	O144:H25
C8[a]	12	2	1	O96:H19
C9[a]	4	1	2	O8:H19
C10 [#]	2	1	1	O135:H30
CSS	427	39	294	
CSD1	70	8	56	SD1
CSD8	7	3	3	SD8
CSD10	2	2	1	SD10
CSB12[a]	8	2	6	SB12
CSB13	7	3	3	SB13

Clusters (no of serotypes)[#]	No of isolates	No. STs	No. rSTs	Serotypes
CSB13-atypical[a]	5	3	3	SB13
Sporadic EIEC lineages[a] (53)	59	49	53	53 antigen types

[#]Numbers in parentheses are the number of serotypes within that cluster.

[a]: Clusters identified as new clusters in this study.

[b]: Serotypes were inconsistent with previous analyses.

Table 2: The sensitivity and specificity of cluster-specific genes

Clusters	**Cluster-specific genes (Single/sets)**[b]	**Identification dataset (1969 isolates)**		
		No of isolates	**Sensitivity**	**Specificity**
C1	Set of 4 genes	288	100	99.94[a]
C2	Set of 3 genes	101	100	100
C3	Set of 3 genes	744	100	99.59[a]
C4	Set of 2 genes	51	100	100
C5	Set of 3 genes	62	100	100
C6	Set of 2 genes	20	100	100
C7	Single gene	10	100	100
C8	Set of 2 genes	12	100	100
C9	Set of 2 genes	4	100	100
C10	Single gene	2	100	100
CSS	Set of 5 genes	427	100	99.87[a]
CSD1	Set of 2 genes	70	100	100
CSD8	Single gene	7	100	100
CSD10	Single gene	2	100	100
CSB12	Single gene	8	100	100
CSB13	Single gene	7	100	100
CSB13-atypical	Single gene	5	100	100
53 Sporadic EIEC lineages	Single gene / lineage	59	100	100

[a]:The specificity of cluster-specific gene set less than 100% was due to at least one FP found in that set.

[b]: The sequences of these genes were listed in Data S1.

Table 3: The accuracy of ShigEiFinder with identification dataset and validation dataset

ShigEiFinder assignments	**Identification Dataset (n=1,969)[a]**		**Validation dataset (n=15,501)**	
	Genomes	**Reads mapping**	**Genomes**	**Reads mapping**
Shigella/EIEC clusters	1871	1848	15,455	15,471
Multiple *Shigella*/EIEC clusters	9	6	33	7
Shigella/EIEC unclustered	0	8	13	23
Not *Shigella*/EIEC	89	89	0	0
Accuracy[b]	99.54%	99.28%	99.70%	99.81%

[a]: Identification dataset has 90 non-*Shigella*/EIEC strains including 72 ECOR strains and 18 *E.albertii* strains. 1,969 assembled genomes and 1,951 reads (reads not available for 18 EIEC isolates downloaded from NCBI) in identification dataset. One of *E.albertii* strain was assigned as SB13 which was grouped into SB13 cluster on the phylogenetic tree.

[b]: The accuracy was defined as the number of *Shigella*/EIEC isolates being correctly assigned to cluster over the total number of tested.

Table 4: Discrepant assignment of 8,224 isolates by ShigEiFinder and Shigatyper

ShigEiFinder Assignment	ShigaTyper assignment			Total
	EIEC	Multiple wzx	Non-prediction	
SS	7,465	12	7	7,484
SF	117	61	10	188
C1 and C2 (SB/SD)	17	99	51	167
SB12	0	2	0	2
SD1	244	1	1	246
SD8	1	0	0	1
SD10	0	0	2	2
EIEC	97	0	0	97
Sporadic EIEC lineages	15	0	0	15
Multiple clusters	5	2	0	7
Shigella/EIEC unclustered	15	0	0	15
Total	7,976	177	71	8,224

Figure legends:

Figure 1: *Shigella*/EIEC cluster **Identification phylogenetic tree**

Representative isolates from the identification dataset were used to construct the phylogenetic tree by Quicktree v1.3 (64) to identify *Shigella* and EIEC clusters and visualised by Grapetree's interactive mode. The dendrogram tree shows the phylogenetic relationships of 1879 *Shigella* and EIEC isolates represented in the identification dataset. Branch lengths are log scale for clarity. The tree scales indicated the 0.2 substitutions per locus. *Shigella* and EIEC clusters are coloured. Numbers in square brackets indicate the number of isolates of each identified cluster. CSP is sporadic EIEC lineages.

Figure 2: *in silico* serotyping pipeline workflow

Schematic of *in silico* serotyping *Shigella* and EIEC by cluster-specific genes combined with the *ipaH* gene and O antigen and modification genes and H antigen genes, implemented in ShigEiFinder. Both assembled genomes and raw reads are accepted as data input.

Supplementary Material

Figure S1: **Identification phylogenetic tree**

An identification phylogenetic tree constructed by Quicktree v1.3 (64) and visualised by ITOL v5 shows the phylogenetic relationships of 1879 *Shigella* and EIEC isolates in identification dataset. The tree scales indicated the 0.01 substitutions per locus. *Shigella* and EIEC clusters are colored. The internal branches are colored to represent the bootstrap values. Green color indicates the maximum bootstrap value (1). The red color shows the minimum bootstrap value (0). Each of cluster is well supported by bootstrap value. CSP is sporadic EIEC lineages.

Figure S2-A: **Confirmation phylogenetic tree**

A confirmation phylogenetic tree was constructed by Quicktree v1.3 (64) based on 2375 isolates and visualised by Grapetree's interactive mode. The tree shows the phylogenetic relationships between identified *Shigella*/EIEC clusters in identification dataset and non-enteroinvasive E. coli isolates. Branch lengths are log scale for clarity. The tree scales indicated the 0.1 substitutions per locus. Known *Shigella* and EIEC clusters from identification dataset are colored. Numbers in square brackets indicate the number of isolates of each identified cluster. CSP is sporadic EIEC lineages.

Figure S2-B: **Confirmation phylogenetic tree**

A confirmation phylogenetic tree constructed by Quicktree v1.3 (64) and visualised by ITOL v5 shows the phylogenetic relationships between identified *Shigella*/EIEC clusters in identification dataset and non-enteroinvasive E. coli isolates. The tree scales indicated the 0.01 substitutions per locus. *Shigella* and EIEC clusters are colored. The internal branches are colored to represent the bootstrap values. Green color indicates the maximum bootstrap value (1). The red color shows the minimum bootstrap value (0). Each of cluster is well supported by bootstrap value. CSP is sporadic EIEC lineages.

Figure S3: **Distribution of mapped 38 virulence genes in 58 sporadic isolates**

The presence of *Shigella* virulence plasmid pINV in 58 sporadic isolates in identification dataset was determined by the mapped 38 virulence genes. Detailed genes were described in Results "**Investigation of *Shigella* virulence plasmid pINV in 58 sporadic isolates**". Three categories were defined based on the number of virulence genes mapped to isolate. Virulence plasmid positive: > 25 genes mapped to isolate; Intermediate: 13 to 25 genes mapped to isolate; Virulence plasmid negative: less than 13 genes mapped to isolate.

Figure S4 (A): **Validation phylogenetic tree**

A validation tree was generated by Quicktree v1.3 (64) and visualised by Grapetree's interactive mode to assign representative isolates in validation dataset to clusters. Branch lengths are log scale for clarity. The tree scales indicated the 0.2 substitutions per locus. Known *Shigella* and EIEC clusters from identification dataset are colored. Numbers in square brackets indicate the number of isolates of each identified cluster. Isolates in validation dataset are colored white. The isolates are assigned to clusters if they grouped into known cluster isolates. CSP is sporadic EIEC lineages.

Figure S4 (B): **Validation phylogenetic tree**

A validation phylogenetic tree was constructed by Quicktree v1.3 (64) and visualised by ITOL v5 to assign representative isolates in validation dataset to clusters. The tree scales indicated the 0.01 substitutions per locus. *Shigella* and EIEC clusters are colored. The internal branches are colored to represent the bootstrap values. Green color indicates the maximum bootstrap value (1). The red color shows the minimum bootstrap value (0). Each of cluster is well supported by bootstrap value. Isolates that grouped with known cluster isolates (from identification dataset) with strong bootstrap support are categorised into that cluster. CSP is sporadic EIEC lineages.

Table S1: 1,969 isolates used in identification dataset

Table S2: 15,501 isolates used in validation dataset

Table S3: The location and function of cluster-specific genes

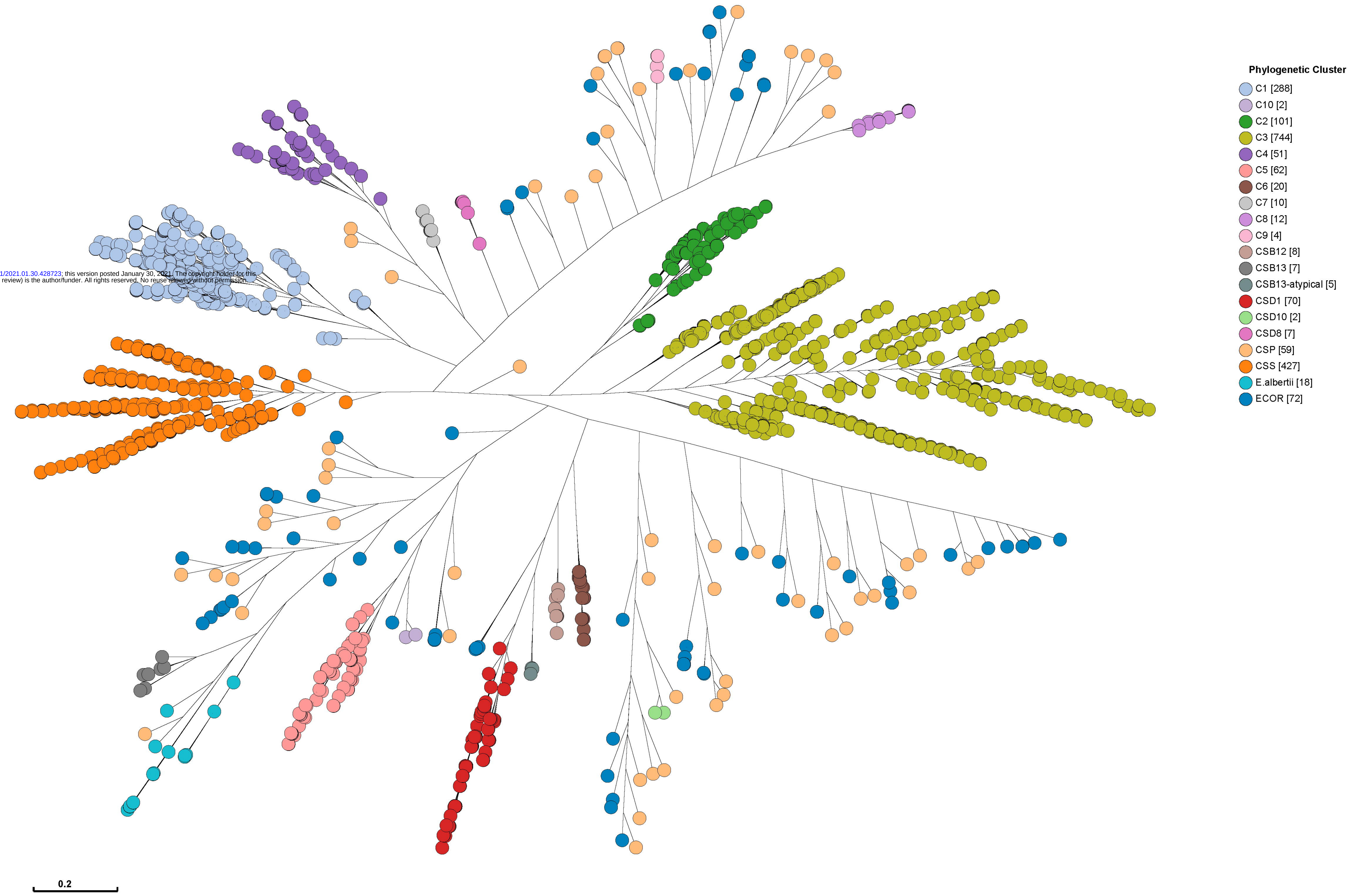
Table S4: The results of cluster-specific gene markers tested with 12,743 non-enteroinvasive E. coli isolates

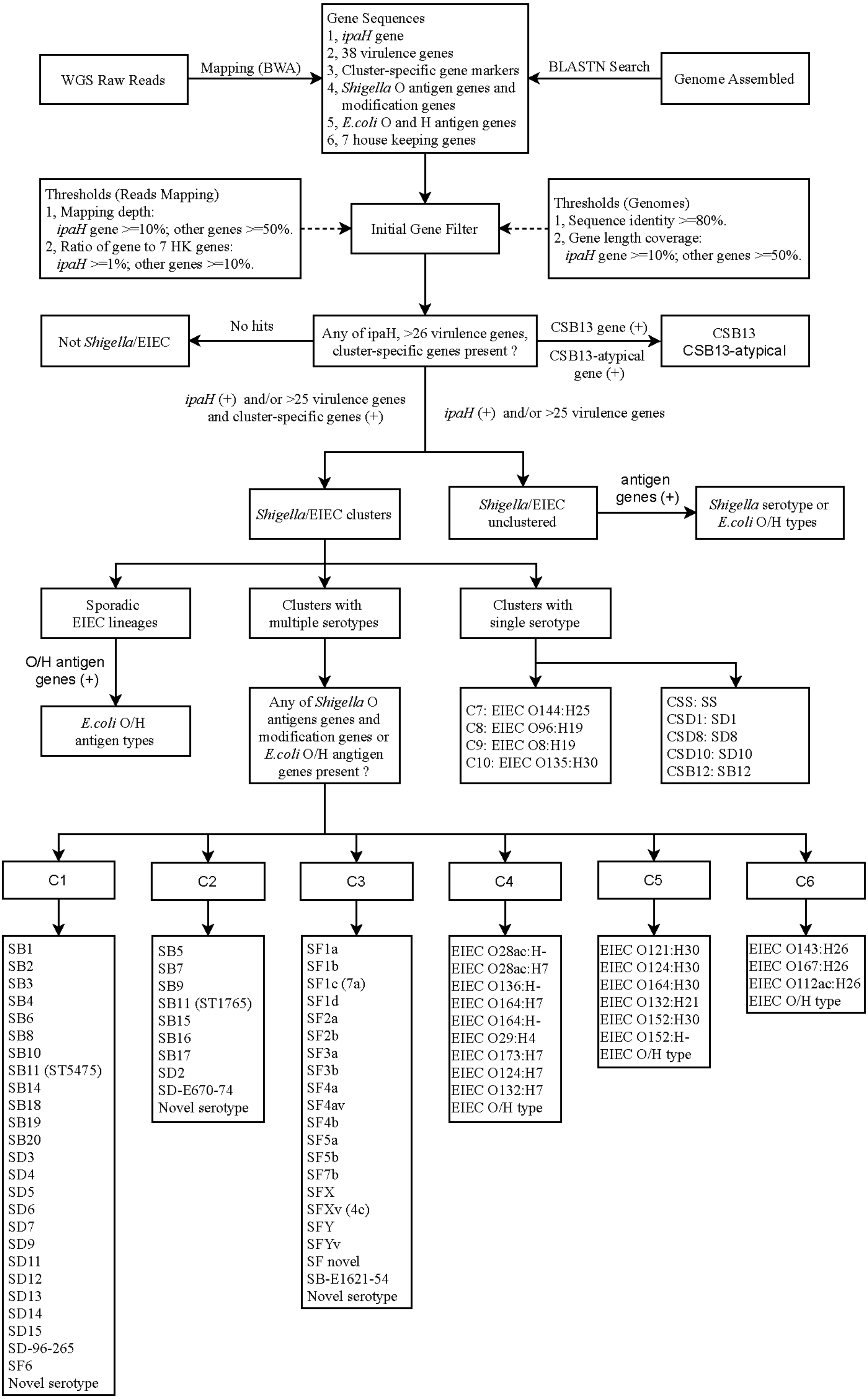
Data S1: Algorithms incorporated into the ShigEiFinder

Data S2: Genetic signature O and H genes from ShigaTyper and SerotypeFinder

Data Availability Statement

Custom python scripts used in this study are available from the authors on request.


Phylogenetic Cluster
C1 [288]
C10 [2]
C2 [101]
C3 [744]
C4 [51]
C5 [62]
C6 [20]
C7 [10]
C8 [12]
C9 [4]
CSB12 [8]
CSB13 [7]
CSB13-atypical [5]
CSD1 [70]
CSD10 [2]
CSD8 [7]
CSP [59]
CSS [427]
E.albertii [18]
ECOR [72]
0.2


WGS Raw Reads
Mapping (BWA)
Gene Sequences
1, *ipaH* gene
2, 38 virulence genes
3, Cluster-specific gene markers
4, *Shigella* O antigen genes and modification genes
5, *E.coli* O and H antigen genes
6, 7 house keeping genes
BLASTN Search
Genome Assembled
Thresholds (Reads Mapping)
1, Mapping depth:
ipaH gene >=10%; other genes >=50%.
2, Ratio of gene to 7 HK genes:
ipaH >=1%; other genes >=10%.
Initial Gene Filter
Thresholds (Genomes)
1, Sequence identity >=80%.
2, Gene length coverage:
ipaH gene >=10%; other genes >=50%.
Not *Shigella*/EIEC
No hits
Any of ipaH, >26 virulence genes, cluster-specific genes present ?
CSB13 gene (+)
CSB13-atypical gene (+)
CSB13
CSB13-atypical
ipaH (+) and/or >25 virulence genes and cluster-specific genes (+)
ipaH (+) and/or >25 virulence genes
Shigella/EIEC clusters
Shigella/EIEC unclustered
antigen genes (+)
Shigella serotype or *E.coli* O/H types
Sporadic EIEC lineages
Clusters with multiple serotypes
Clusters with single serotype
O/H antigen genes (+)
E.coli O/H antigen types
Any of *Shigella* O antigens genes and modification genes or *E.coli* O/H angtigen genes present ?
C7: EIEC O144:H25
C8: EIEC O96:H19
C9: EIEC O8:H19
C10: EIEC O135:H30
CSS: SS
CSD1: SD1
CSD8: SD8
CSD10: SD10
CSB12: SB12
C1
C2
C3
C4
C5
C6
SB1
SB2
SB3
SB4
SB6
SB8
SB10
SB11 (ST5475)
SB14
SB18
SB19
SB20
SD3
SD4
SD5
SD6
SD7
SD9
SD11
SD12
SD13
SD14
SD15
SD-96-265
SF6
Novel serotype
SB5
SB7
SB9
SB11 (ST1765)
SB15
SB16
SB17
SD2
SD-E670-74
Novel serotype
SF1a
SF1b
SF1c (7a)
SF1d
SF2a
SF2b
SF3a
SF3b
SF4a
SF4av
SF4b
SF5a
SF5b
SF7b
SFX
SFXv (4c)
SFY
SFYv
SF novel
SB-E1621-54
Novel serotype
EIEC O28ac:H-
EIEC O28ac:H7
EIEC O136:H-
EIEC O164:H7
EIEC O164:H-
EIEC O29:H4
EIEC O173:H7
EIEC O124:H7
EIEC O132:H7
EIEC O/H type
EIEC O121:H30
EIEC O124:H30
EIEC O164:H30
EIEC O132:H21
EIEC O152:H30
EIEC O152:H-
EIEC O/H type
EIEC O143:H26
EIEC O167:H26
EIEC O112ac:H26
EIEC O/H type