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

83 (10, 23, 24, 34, 35). *Shigella* and EIEC have arisen from *E. coli* in multiple independent
84 events and should be regarded as a single pathovar of E. coli (25, 26, 28, 36-38). Previous
85 phylogenetic studies suggested that *Shigella* isolates were divided into 3 clusters (C1, C2 and
86 C3) with 5 outliers (SS, SB13, SD1, SD8 and SD10) (25, 38) whereas EIEC isolates were
87 grouped into four clusters (C4, C5, C6 and C7) (26). The seven *Shigella*/EIEC clusters and 5
88 outliers of *Shigella* are within the broader non-enteroinvasive *E. coli* species except for SB13
89 which is closer to *Escherichia albertii* (39, 40). Alternative WGS-based phylogenomic studies
90 have also defined multiple clusters of *Shigella* and EIEC (23, 28, 41).
91
92 The traditional biochemical test for motility and lysine decarboxylase (LDC) activity (42) and
93 molecular test for the presence of *ipaH* gene have been used to differentiate *Shigella* and EIEC
94 from non-enteroinvasive E. coli (24, 43-45). Agglutination with *Shigella*/EIEC associated
95 antiserum further classify *Shigella* or EIEC to serotype level. However, cross-reactivity, strains
96 not producing O antigens, and newly emerged *Shigella* serotypes may all prevent accurate
97 serotyping (10, 46). Serotyping by antigenic agglutination is being replaced by molecular
98 serotyping (47, 48), which can be achieved through examination of the sequences of O antigen
99 biosynthesis and modification genes (8, 24, 49-52).
100
101 Recently, PCR-based molecular detection methods targeting the gene *lacY* were developed to
102 distinguish *Shigella* from EIEC (53, 54). However, the ability of the primers described in these
103 methods to accurately differentiate between *Shigella* and EIEC was later questioned (23, 28).
104 With the uptake of whole-genome sequencing technology, several studies have identified
105 phylogenetic clade specific markers, species specific markers and EIEC lineage-specific genes
106 for discrimination between *Shigella* and EIEC and between *Shigella* species (23, 27, 28, 41, 55,
107 56). More recently, genetic markers *lacY*, *cadA*, *Ss_methylase* were used for identification of
108 *Shigella* and EIEC (10). However, these markers failed to discriminate between *Shigella* and
109 EIEC when a larger genetic diversity is considered (23, 28, 55). A Kmer-based approach can
110 identify *Shigella* isolates to the species level but misidentification was also observed (56).
111
112 In this study, we aimed to i), identify phylogenetical clusters of *Shigella* and EIEC through
113 large scale examination of publicly available genomes; ii), identify cluster-specific gene
114 markers using comparative genomic analysis of *Shigella* and EIEC accessory genomes for
115 differentiation of *Shigella* and EIEC; iii), develop a pipeline for *Shigella* and EIEC *in silico*
116 serotyping based on the cluster-specific gene markers combined with *Shigella* and EIEC

117 serotype-specific O antigen and H antigen genes. We demonstrate that these cluster-specific
118 gene markers enhance *in silico* serotyping using genomic data. We also developed an
119 automated pipeline for cluster typing and serotyping of *Shigella*/EIEC from WGS data.
120

121 **Materials and Methods**

122 **Identification of *Shigella*/EIEC isolates from NCBI database**

123 *E. coli/Shigella* isolates from the NCBI SRA (National Center for Biotechnology Information
124 Sequence Read Archive) as May of 2019 were queried. Raw reads were retrieved from ENA
125 (European Nucleotide Archive). The *ipaH* gene (GenBank accession number M32063.1) was
126 used to screen E. coli*/Shigella* reads using Salmon v0.13.0 (57). Taxonomic classification for
127 *E. coli/Shigella* was confirmed by Kraken v1.1.1 (58). Molecular serotype prediction of *ipaH*
128 negative *Shigella* isolates was performed by ShigaTyper v1.0.6 (10). Isolates that were *ipaH*
129 positive and isolates with designation of SB13 by ShigaTyper were selected as *Shigella*/EIEC
130 database.
131

132 The sequence types (STs) and ribosomal STs (rSTs) of *ipaH* gene negative E. coli (non-
133 enteroinvasive E. coli) isolates were examined. STs and rSTs for these isolates were obtained
134 from the E. coli*/Shigella* database in the Enterobase (59) as of May 2019. For STs and rSTs
135 with only one isolate, the isolates were selected. For STs and rSTs with more than one isolates,
136 one representative isolate for each ST and rST were randomly selected. In total, 12,743 ipaH
137 negative E. coli isolates representing 3,800 STs and 11,463 rSTs were selected as non-
138 enteroinvasive E. coli control database.
139

140 **Genome sequencing**

141 Whole-genome sequencing (WGS) of 31 EIEC strains used in a previous study (26) was
142 performed by Illumina NextSeq (Illumina, Scoresby, VIC, Australia). DNA libraries were
143 constructed using Nextera XT Sample preparation kit (Illumina Inc., San Diego, CA, USA) and
144 sequenced using the NextSeq sequencer (Illumina Inc.). FASTQ sequences of the strains
145 sequenced in this study were deposited in the NCBI under the BioProject (PRJNA692536).
146
147

148 **Genome assembly and data processing**

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

151 QUAST v5.0.0 (61). Three standard deviations (SD) from the mean for contig number, largest
152 contig, total length, GC, N50 and genes were used as quality filter for assembled genomes.
153
154 The STs for isolates in *Shigella*/EIEC database was checked by using mlst
155 (https://github.com/tseemann/mlst) with the *E. coli* scheme from PubMLST (62). rSTs were
156 extracted from the E. coli/*Shigella* rMLST database in Enterobase (59) as of May 2019.
157 Serotype prediction for isolates in *Shigella*/EIEC was performed by ShigaTyper v1.0.6 (10).
158 Serotyping of E. coli O and H antigens were predicted by using SerotypeFinder v2.0.1 (63).
159
160 **Selection of isolates for *Shigella*/EIEC identification dataset**
161 The selection of isolates for the identification dataset was based on the representative isolates
162 for each ST, rST and serotype of *Shigella* and EIEC in the *Shigella*/EIEC database. For STs
163 and rSTs with only one isolate, the isolate was selected. For STs and rSTs with more than one
164 isolates, one representative isolate for each ST, rST was randomly selected. A representative
165 experimentally confirmed isolate of each serotype of *Shigella* and EIEC was also randomly
166 selected. 72 ECOR strains downloaded from Enterobase (59) and 18 *E. albertii* strains were
167 used as controls for the identification dataset. The details of the identification dataset are listed
168 in Table S1. The remaining isolates in *Shigella*/EIEC database were referred as validation
169 dataset (Table S2).
170
171 The identification dataset was used for identification of phylogenetic relationships of *Shigella*
172 and EIEC. The identification dataset was also used for identification of cluster-specific genes.
173 The validation dataset was used to evaluate the performance of cluster-specific gene markers
174 using the *in-silico* serotyping pipeline.
175
176 **Phylogeny of *Shigella* and EIEC based on WGS**
177 Three phylogenetic trees including identification tree, confirmation tree and validation tree
178 were constructed by Quicktree v1.3 (64) with default parameters to identify and confirm the
179 phylogenetic clustering of *Shigella* and EIEC isolates. The phylogenetic trees were visualised
180 by Grapetree's interactive mode and ITOL v5 (65, 66).
181
182 The identification phylogenetic tree was generated based on isolates in the identification
183 dataset for identification of clusters of *Shigella* and EIEC isolates (Fig. 1). A subset of 485
184 isolates known to represent each identified cluster from the identification dataset were then

185 selected. The subset of 485 isolates from the identification dataset and 1,872 non-
186 enteroinvasive E. coli isolates from non-enteroinvasive E. coli control dataset (2,357 isolates
187 total) were used to construct a confirmation tree. This tree was used for confirmation of the
188 phylogenetic relationships between identified *Shigella*/EIEC clusters in the identification
189 dataset and non-enteroinvasive E. coli isolates. The validation tree was generated based on
190 1,159 representative isolates from the validation dataset that were selected in the same way as
191 the identification dataset and a subset of 485 isolates from the identification dataset to assign
192 validation dataset isolates to clusters.
193
194 **Investigation of *Shigella* virulence plasmid pINV**
195 The presence of *Shigella* virulence plasmid pINV in isolates were investigated by using BWA-
196 MEM v0.7.17 (Burrows-Wheeler Aligner) (67) to align isolate raw reads onto the reference
197 sequence of pINV (68) (NC_024996.1). Mapped reads were sorted and indexed using
198 Samtools v1.9 (69). The individual gene coverage from mapping was obtained using Bedtools
199 coverage v2.27.1 (70).
200
201 **Identification of the cluster-specific gene markers**
202 Cluster-specific gene markers were identified from *Shigella*/EIEC accessory genomes. The
203 genomes from the identification dataset were annotated using PROKKA v1.13.3 (71). Pan- and
204 core-genomes were analysed by roary v3.12.0 (72) using an 80% sequence identity threshold.
205 The genes specific to each cluster were identified from the accessory genes with an in-house
206 python script. In this study, the number of genomes from a given cluster containing all specific
207 genes for that cluster was termed true positives (TP), the number of genomes from the same
208 cluster lacking any of those same genes was termed false negatives (FN). The number of
209 genomes from other clusters containing all of those same genes was termed false positives
210 (FP).
211
212 The sensitivity (True positive rate, TPR) of each cluster-specific gene marker was defined as
213 TP/(TP+FN). The specificity (True negative rate, TNR) was defined as TN/(TN+FP).
214
215 **Validation of the cluster-specific gene markers**
216 The ability of cluster-specific gene markers to assign *Shigella*/EIEC isolates was examined by
217 using BLASTN to search against the validation dataset (Table S2) and non-enteroinvasive E.
218 coli control database for the presence of any of the cluster-specific gene marker or a set of

219 cluster-specific gene markers. The BLASTN thresholds were defined as 80% sequence identity
220 and 50% gene length coverage.
221
222 **Development an automated pipeline for molecular serotyping of *Shigella*/EIEC**
223 The pipeline was developed using paired end illumina genome sequencing reads or assembled
224 genomes identify cluster-specific gene markers combined with *Shigella*/EIEC serotype specific
225 O antigen genes (wzx and wzy) and modification genes (Fig. 2, Data S1). We used the same
226 signature O and H sequences from ShigaTyper and SerotypeFinder (Data S2) (10, 63). These
227 includes *Shigella* serotype-specifc wzx/wzy genes and modification genes from ShigaTyper
228 and E. coli O antigen and *fliC* (H antigen) genes from SerotypeFinder. *ipaH* gene and 38
229 virulence genes used in analysis of virulence of 59 sporadic EIEC isolates were also included
230 in the typing reference sequences database. Seven House Keeping (HK) genes -*recA*, *purA*,
231 *mdh*, *icd*, *gyrB*, *fumC* and *adk* downloaded from NCBI were used for contamination checking.
232
233 Raw reads were aligned to the typing reference sequences by using BWA-MEM v0.7.17 (67).
234 The mapping length percentage and the mean mapping depth for all genes were calculated
235 using Samtools coverage v1.10 (69). To determine whether the genes present or absent, 50% of
236 mapping length for all cluster-specific genes, virulence genes and O antigen genes and 10% for
237 *ipaH* gene were used as cutoff value. The ratio of mean mapping depth to the mean mapping
238 depth of the 7 HK genes was used to determine a contamination threshold with ratios less than
239 1% for *ipaH* gene and less than 10% for other genes assigned as contamination. Reads
240 coverage mapped to particular regions of genes were checked by using samtools mpileup
241 v1.10.
242
243 Assembled genomes were BLASTN v2.9.0 (73) searched against the typing reference
244 sequences with 80% sequence identity and 50% gene length coverage for all genes with
245 exception of *ipaH* gene which was defined as 10% gene length coverage.
246
247 The pipeline was tested with the identification dataset and validated with the *Shigella*/EIEC
248 validation dataset and non-enteroinvasive E. coli control database. The specificity defined as (1
249 - the number of non-enteroinvasive E. coli isolates being detected / the total number of non-
250 enteroinvasive E. coli isolates) * 100.
251
252

253 **Results**

254 **Screening sequenced genomes for *Shigella*/EIEC isolates**

255 We first screened available *E. coli* and *Shigella* genomes based on the presence of *ipaH* gene.
256 We examined 122,361 isolates with the species annotation of E. coli (104,256) or *Shigella*
257 (18,105) with paired end illumina sequencing reads available in NCBI SRA database. Of
258 122,361 isolates, 17,989 isolates were positive to the *ipaH* gene including 455 out of 104,256
259 E. coli isolates and 17,434 out of 18,105 *Shigella* isolates. The 17,989 *ipaH* positive *E. coli* and
260 *Shigella* genomes and 571 *ipaH* negative "*Shigella"* genomes were checked for taxonomic
261 classification and genome assembly quality. 17,320 *ipaH* positive *E. coli* and *Shigella* genomes
262 and 246 *ipaH* negative "*Shigella"* genomes passed quality filters. Among 246 *ipaH* negative
263 "*Shigella"* genomes, 11 isolates belonged to SB13 by using ShigaTyper (10) while the
264 remaining 235 isolates were classified with taxonomic identifier of E. coli by Kraken v1.1.1
265 (58) and were removed from analysis. A total of 17,331 genomes including 17,320 *ipaH*
266 positives and 11 SB13 genomes were selected to form the *Shigella*/EIEC database, which
267 contained 429 genomes with species identifier of *E. coli* and 16,902 genomes with species
268 identifier of *Shigella.*

269

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

277

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

286

287 **Identification of *Shigella* and EIEC clusters**

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

300

301 All known clusters were identified (Fig. 1) including 3 *Shigella* clusters (C1, C2, C3) and 5
302 outliers (SD1, SD8, SD10, SB13 and SS) as defined by Pupo et al (25) and 4 EIEC clusters
303 (C4, C5, C6 and C7) defined by Lan et al. (26). Each of these clusters was supported by a
304 bootstrap value of 80% or greater (Fig. S1). 1,789 isolates of the 1,879 *Shigella*/EIEC isolates
305 (1,830 isolates from the *Shigella*/EIEC database, 31 in-house sequenced EIEC isolates and 18
306 EIEC isolates from Hazen *et al.*) fell within these clusters.

307

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

320 *Shigella* while EIEC O96:H19, EIEC O8:H19 and EIEC O135:H30 were defined as C8, C9
321 and C10 respectively. In total there were 10 *Shigella* clusters and 7 EIEC clusters (Table 1).
322
323 **Analysis of the 59 sporadic EIEC isolates**
324 To determine the phylogenetic relationships of the above defined clusters and the remaining 59
325 sporadic EIEC isolates within the larger non-enteroinvasive *E. coli* population a confirmation
326 tree was generated using 485 isolates representing the known clusters and 1,872 representative
327 non-*Shigella*/EIEC isolates (Fig. S2). The 59 sporadic EIEC isolates including 2 EIEC isolates
328 M2330 (O152:H51) and M2339 (O124:H7) sequenced in this study and 57 isolates were
329 interspersed among non-*Shigella*/EIEC isolates and did not form large clusters. Groups of these
330 isolates that were not previously identified were named as sporadic EIEC lineage followed by
331 their serotype. For example, M2339 (O124:H7) grouped together with one other EIEC isolate
332 with the same O and H antigens O124:H7 and were named 'sporadic EIEC lineage O124:H7'.
333 There were 53 sporadic EIEC lineages including 5 lineages with 2 or more isolates and 48
334 lineages with only one isolate. The STs, rSTs and antigen types of these 59 isolates were listed
335 in the Table S1.
336
337 Some of the sporadic EIEC isolates fell into STs containing *ipaH* negative isolates. We
338 therefore examined the presence of the pINV virulence plasmid in the sporadic EIEC isolates.
339 We selected 38 genes that are essential for virulence including 35 genes (12 *mxi* genes, 9 *spa*
340 genes, 5 *ipaA-J* genes, 6 *ipgA-F* genes as well as *acp*, *virB*, *icsB*) in the conserved entry region
341 encoding the Mxi-Spa-Ipa type III secretion system and its effectors and 3 regulator genes
342 (*virF*, *virA* and *icsA*/*virG*) (24, 33, 68) and determined the presence of pINV in the 59 sporadic
343 EIEC isolates by mapping the sequence reads onto a pINV reference sequence (68). Reads
344 from 18 non-*Shigella/EIEC* isolates that shared the same ST as one of 58 sporadic isolates
345 were positive for these genes.
346
347 The number of essential virulence genes with mapped reads in the 59 sporadic EIEC isolates
348 were analysed (Fig. S3). Those isolates containing more than 25 of the 38 essential virulence
349 genes were defined as virulence plasmid positive. While isolates containing between 13 and 25
350 were defined as intermediate and less than 13 were defined as virulence plasmid negative.
351
352 The 2 newly sequenced sporadic EIEC isolates (M2330 and M2339) were positive for the
353 virulence plasmid and of the other 57 sporadic EIEC isolates, 39 isolates were positive, 9

354 isolates were negative and 9 isolates were intermediate (Table S1). The results were compared
355 with those non-*Shigella/*EIEC isolates belonging to the same ST. The virulence plasmid was
356 absent in all non-*Shigella/*EIEC isolates while all sporadic EIEC isolates in these STs were
357 either positive or intermediate. Therefore, this analysis confirmed the sporadic isolates
358 belonged to EIEC and the STs contained both EIEC and non- EIEC isolates.
359
360 **Identification of cluster-specific gene markers**
361 In this study, cluster-specific gene markers were either a single gene present in all isolates of a
362 cluster and absent in all other isolates or a set of genes (two or more) that as a combination
363 were only found in one cluster. For the marker sets, a subset of cluster-specific gene markers
364 for a given cluster could be found in other clusters but the entire set was only found in the
365 target cluster.
366
367 Comparative genomic analysis on 1,969 accessory genomes from the identification dataset was
368 used to identify cluster-specific gene markers or marker sets. Multiple candidate cluster-
369 specific gene markers or marker sets of markers for each of 17 *Shigella*/EIEC clusters and 53
370 sporadic EIEC lineages were identified through screening the accessory genes from 1,969
371 genomes . These gene markers or marker sets were 100% sensitive to clusters but with varying
372 specificity. The cluster-specific gene markers or marker sets of markers with the lowest FP
373 rates were then selected from candidate cluster-specific gene markers by BLASTN searches
374 against genomes in the identification dataset using 80% sequence identity and 50% gene length
375 coverage threshold.
376
377 Five single cluster-specific gene markers (C7, C10, SB12, SB13 and atypical SB13) and 12
378 sets of cluster-specific gene markers (C1, C2, C3, C4, C5, C6, C8, C9, SS, SD1, SD8 and
379 SD10) were selected for *Shigella*/EIEC cluster typing. The sensitivity and specificity for each
380 cluster-specific gene marker or a set of cluster-specific gene markers for the identification
381 dataset were listed in Table 2. The cluster-specific gene markers or marker sets of markers
382 were all 100% sensitive and 100% specific with exception of C1 (99.94%), C3 (99.91%
383 specificity) and SS (99.8% specificity). A single specific gene for each of 53 sporadic EIEC
384 lineages were also selected with the exception of one lineage which has a set of 2 genes. These
385 genes were all 100% sensitive and specific for a given sporadic EIEC lineage.
386

387 All cluster-specific gene markers, 37 in total (5 single, 32 genes in 12 sets) and 54 sporadic
388 EIEC lineages specific gene markers were located on chromosome but one of C4 gene markers
389 and 5 sporadic EIEC lineages specific genes were located on plasmid. None of the cluster-
390 specific gene markers were contiguous in the genomes. The location of these cluster-specific
391 gene markers was determined by BLASTN against representative complete genomes of
392 *Shigella*/EIEC containing gene features downloaded from NCBI GenBank. In those cluster or
393 sporadic lineages with no representative complete genome specific gene markers were named
394 using their cluster or sporadic EIEC lineage followed by the cluster or lineage number. For
395 example, C7 specific gene marker was named "C7 specific gene".
396
397 The functional characterization of these specific gene markers were identified from RAST
398 annotation (74). For 37 cluster-specific gene markers, 22 had known functions and 15 encoded
399 hypothetical proteins with unknown functions, while 11 sporadic EIEC lineages specific gene
400 markers were identified with known functions and 43 were hypothetical proteins with
401 unknown functions. The location and functions of specific gene markers are listed in Table S3.
402
403 **Validation of cluster-specific gene markers**
404 The ability of cluster-specific gene markers to correctly assign *Shigella*/EIEC isolates was
405 evaluate with 15,501 *Shigella*/EIEC isolates in the validation dataset, 12,743 isolates from non-
406 enteroinvasive *E. coli* control database.
407
408 Using cluster-specific gene markers, 15,443 of the 15,501 (99.63%) *Shigella*/EIEC isolates
409 were correctly assigned to clusters which included 15,337 *Shigella* isolates, 102 EIEC isolates,
410 4 sporadic EIEC isolates, and 38 (0.24%) isolates with more than one clusters. Twenty of the
411 15,501 (0.13%) *Shigella*/EIEC isolates were not assigned to any of identified clusters.
412
413 To confirm the assignment of cluster-specific gene markers, we constructed a "validation"
414 phylogenetic tree (Fig. S4) using 1,159 representative isolates from the validation dataset and a
415 subset of 485 isolates from each cluster from the identification dataset. Isolates that grouped
416 with known cluster isolates (from identification dataset) with strong bootstrap support were
417 assigned to that cluster. All 1,159 isolates were grouped into known clusters on the validation
418 phylogenetic tree. The cluster-specific gene markers assignments were entirely consistent with
419 cluster assignments by phylogenetic tree.
420

421 We tested cluster-specific gene markers with the 12,743 non-enteroinvasive *E. coli* isolates.
422 The *Shigella*/EIEC cluster-specific gene markers were highly specific with specificity varying
423 from 98.8% to 100% for cluster-specific genes and 97.02% to 100% for sporadic EIEC specific
424 genes. Details are listed in Table S4.
425
426 **Development an automated pipeline for molecular serotyping of *Shigella*/EIEC**
427 Above results showed that cluster-specific gene markers were sensitive and specific and can
428 distinguish *Shigella* and EIEC isolates. We therefore used these genes combined with
429 established *Shigella*/EIEC serotype specific O antigen and H antigen genes to develop an
430 automated pipeline for *in silico* serotyping of *Shigella*/EIEC (Fig. 2).
431
432 The pipeline is named *Shigella* EIEC Cluster Enhanced Serotype Finder (ShigEiFinder).
433 ShigEiFinder can process either paired end Illumina sequencing reads or assembled genomes
434 (https://github.com/LanLab/ShigEiFinder). ShigEiFinder classifies isolates into Non-
435 *Shigella*/EIEC*, Shigella* or EIEC clusters based on the presence of *ipaH* gene, number of
436 virulence genes, cluster specific genes. The "Not *Shigella*/EIEC" assignment was determined
437 by the absence of *ipaH* gene, virulence genes (>25/38) and cluster-specific gene markers. The
438 "*Shigella* or EIEC clusters" assignments were made based on the presence of *ipaH* gene,
439 and/or more than 25 virulence genes together with the presence of any of cluster-specific gene
440 markers or marker set, whereas the presence of *ipaH* gene and/or more than 25 virulence genes
441 with absence of any of cluster-specific gene markers were assigned as "*Shigella*/EIEC
442 unclustered".
443
444 *Shigella* and EIEC isolates were differentiated and serotypes were assigned after cluster
445 assignment. ShigEiFinder predicts a serotype through examining the presence of any of
446 established *Shigella* serotype specific O antigen and modification genes and E. coli O and H
447 antigen genes that differentiate the serotypes as ShigaTyper and SerotypeFinder (10, 63). A
448 "novel serotype" is assigned if no match to known serotypes.
449
450 Two pairs of *Shigella* serotypes, SB1/SB20 and SB6/SB10, are known to be difficult to
451 differentiate as they share identical O antigen genes (10, 46, 75). ShigaTyper used a heparinase
452 gene for the differentiation of SB20 from SB1 and *wbaM* gene for the separation of SB6 from
453 SB10. We found that fragments of the heparinase and *wbaM* genes may be present in other
454 serotypes and cannot accurately differentiate SB1/SB20 and SB6/SB10. We found a SB20

455 specific gene which encoded hypothetical proteins with unknown functions and located on a
456 plasmid by comparative genomic analysis of all isolates in C1 accessory genome. The SB20
457 specific gene can reliably differentiate SB20 from SB1and also one SNP each in *wzx* and *wzy*
458 genes that can differentiate SB6 from SB10. We used these differences (Data S1) in
459 ShigEiFinder for the prediction of these serotypes.
460
461 **The accuracy and specificity of ShigEiFinder in cluster typing**
462 The accuracy of ShigEiFinder was tested with 1,969 isolates (1,969 assembled genomes and
463 1,951 Illumina reads [note no reads available for 18 EIEC isolates from NCBI) from the
464 identification dataset and 15,501 isolates from the validation dataset. The results are listed in
465 Table 3.
466
467 ShigEiFinder was able to assign 99.54% and 99.28% of the isolates in the identification dataset
468 to clusters for assembled genomes and reads mapping respectively. The accuracy was 99.70%
469 and 99.81% for assembled genomes and reads mapping respectively when applied to the
470 validation dataset. Discrepancies were observed between assembled genomes and reads
471 mapping (Table 3). There were more isolates assigned to "*Shigella*/EIEC unclustered" in reads
472 mapping, in contrast there were more isolates assigned to multiple clusters in genome
473 assemblies. The specificity of ShigEiFinder was 99.40% for assembled genomes and 99.38%
474 for reads mapping when evaluated with 12,743 non-*Shigella*/EIEC *E. coli* isolates. An
475 additional 2 isolates were detected as sporadic EIEC lineages by reads mapping.
476
477 **Comparison of ShigEiFinder and ShigaTyper**
478 To demonstrate ShigEiFinder for differentiation of *Shigella* from EIEC and enhancement of
479 cluster based serotyping, the comparison of reads mapping results between ShigEiFinder and
480 the existing *in silico Shigella* identification pipeline ShigaTyper (10) was performed with 488
481 isolates used in ShigaTyper and 15,501 isolates from *Shigella*/EIEC validation dataset used in
482 the present study.
483
484 The 488 isolates used in ShigaTyper consisted of 23 other species, 45 *E. coli* isolates and 420
485 *Shigella* isolates. ShigEiFinder identified 23 other species isolates and 453 out of 465 *E. coli*
486 and *Shigella* isolates correctly, in agreement with ShigaTyper assignment. ShigEiFinder also
487 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).

522 Our WGS-based phylogeny provided high resolution for assigning *Shigella* and EIEC isolates
523 to clusters. Several serotypes that are currently increasing in frequency (SB19, SB20, SD14,
524 SD15, SD provisional serotype 96-626) (76-79) were assigned to clusters and five new
525 clusters/outliers were identified. SB13 isolates in this study formed two known lineages. One
526 lineage was located outside of *Shigella*/EIEC clusters and represented the outlier SB13 which
527 is in fact belonging to the newly defined species *E. albertii* (25, 26, 38, 39). The second lineage
528 was with *E. coli*, and was defined as atypical SB13 previously (39). The newly identified
529 *Shigella* outlier SB12 was previously grouped into C3 based on housekeeping gene trees (25,
530 38) but was seen as outliers in two other studies (28, 56).
531
532 Newly identified clusters C8 (EIEC O96:H19) and C9 (EIEC O8:H19) represented the
533 emergence of novel EIEC serotypes. A recent study revealed that EIEC serotype O96:H19 (C8)
534 could be the result of a recent acquisition of the invasion plasmid by commensal E. coli (80).
535 The EIEC serotype O8:H19 (C9) had not been reported previously.
536
537 Apart from the 17 major and outlier clusters of *Shigella* and EIEC, the presence of 53 sporadic
538 EIEC lineages indicated greater genetic diversity than has been observed previously. Isolates
539 belonging to these sporadic EIEC groups were more closely related to non-enteroinvasive E.
540 coli isolates than to major *Shigella*/EIEC lineages. However, 41 of the isolates representing 38
541 sporadic EIEC lineages that carried pINV. *Shigella* and EIEC both carry the *Shigella* virulence
542 plasmid pINV which is vital for virulence and distinguishes *Shigella*/EIEC from other *E. coli*
543 (24, 33, 68). Therefore, these isolates may represent recently formed EIEC lineages through
544 acquisition of the pINV. The remaining 18 isolates contained the *ipaH* gene but may or may
545 not carry pINV. It is possible that these strains carried very low copy number of pINV or the
546 pINV plasmid was lost during culture.
547
548 **Highly sensitive and specific cluster-specific gene markers for differentiation of *Shigella***
549 **and EIEC isolates**
550 Several studies have identified phylogenetic related genomic markers for discrimination of
551 *Shigella* and EIEC and between *Shigella* species (23, 27, 28, 41, 55, 56). However, these
552 phylogenetic analyses were performed only with a small number of genomes (23, 28, 55). In
553 addition, non-invasive E. coli isolates were included in some of the phylogenetic clusters
554 identified (28) which led to non-invasive E. coli isolates being identified by the markers.
555

556 We identified cluster-specific gene markers for each respectively clusters which were only
557 composed of *Shigella* or EIEC isolates. Sets of cluster-specific gene markers were identified
558 for those clusters where no single suitable marker is present. The combination of genes
559 enhances the specificity of cluster-specific gene markers as demonstrated by the 100%
560 sensitivity and very high specificity in this analysis (Table 2). Genes specific to each of the 53
561 sporadic EIEC lineages were also identified and they were sensitive and specific, although it
562 should be noted that these values are based on very small sample sizes.
563
564 The cluster-specific gene markers or marker sets can be used to differentiate *Shigella*/EIEC
565 from non-enteroinvasive *E. coli* independent of *ipaH* gene. The *ipaH* gene as a molecular
566 target has been used to differentiate *Shigella* and EIEC from non-enteroinvasive E. coli (24,
567 43-45). In our study, the cluster-specific gene markers were specific to *Shigella*/EIEC with
568 98.8% to 100% specificity when evaluated on non-enteroinvasive *E. coli* control database,
569 giving us the confidence that the cluster-specific genes or sets are robust markers to identify
570 *Shigella*/EIEC. 53 sporadic EIEC lineage specific gene markers also have very high specificity
571 (97.02% to 100%) against non-enteroinvasive *E. coli* control database.
572
573 The cluster-specific gene markers or marker sets are able to assign *Shigella*/EIEC isolates
574 correctly in 99.63% of cases and can clearly distinguish *Shigella* isolates from EIEC when
575 applied to the validation dataset. While ShigaTyper assigned 46.95% isolates to *Shigella* and
576 51.45% isolates to EIEC in the same dataset we tested, leading to a large proportion of isolates
577 incorrectly assigned. The majority of the isolates predicted as EIEC by ShigaTyper were SS or
578 SD1 as they belonged to SS and SD1 specific STs and were positive to a set of SS or SD1
579 specific gene markers and grouped into SS or SD1 cluster on our phylogenetic tree. The genes
580 used in ShigaTyper were SS specific marker Ss_methylase gene (81, 82) together with SS O
581 antigen wzx gene. However, SS specific marker Ss_methylase gene was found in other
582 *Shigella* serotypes and EIEC (10) and SS O antigen wzx gene were located on a plasmid which
583 is frequently lost (83). Similarly, the SD1 O antigen genes used in ShigaTyper were plasmid-
584 borne which may also lead to inconsistent detection (84, 85). A previous study identified 6 loci
585 to distinguish EIEC from *Shigella* (23). We searched the 6 loci against our *Shigella*/EIEC
586 database and found that some *Shigella* isolates were misidentified as EIEC isolates, such as
587 SD8 isolates incorrectly identified as EIEC subtype 13. Our cluster-specific genes can
588 differentiate SD8 isolates from EIEC with 100% accuracy. Therefore, the cluster-specific gene
589 markers marker sets provided nearly perfect differentiation of *Shigella* from EIEC.

590

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

599

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

606

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

618

619 The genetic determinants used in ShigaTyper for differentiation of *Shigella* from EIEC and
620 identification of SS were *lacY*, *cadA*, *Ss_methylase,* SS and SD1 O antigen *wzx* genes (10). As
621 discussed above some of these genes were found to be non-specific in this study. Compared
622 with ShigaTyper, the cluster-specific gene markers used in ShigEiFinder for identification of
623 *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.

691 7. Khalil IA, Troeger C, Blacker BF, Rao PC, Brown A, Atherly DE, et al. Morbidity and
692 mortality due to *shigella* and enterotoxigenic *Escherichia coli* diarrhoea: the Global Burden of
693 Disease Study 1990–2016. *The Lancet Infectious Diseases*. 2018;18(11):1229-40.
694 8. van den Beld MJ, Warmelink E, Friedrich AW, Reubsaet FA, Schipper M, de Boer RF,
695 et al. Incidence, clinical implications and impact on public health of infections with *Shigella*
696 *spp*. and entero-invasive *Escherichia coli* (EIEC): results of a multicenter cross-sectional study
697 in the Netherlands during 2016–2017. *BMC Infectious Diseases*. 2019;19(1):1037.
698 9. Edwards PR, Ewing WH. Identification of enterobacteriaceae. Identification of
699 Enterobacteriaceae.. 1972(Third edition).
700 10. Wu Y, Lau HK, Lee T, Lau DK, Payne J. *In silico* serotyping based on Whole-Genome
701 sequencing improves the accuracy of *shigella* identification. *Appl Environ Microbiol*.
702 2019;85(7):e00165-19.
703 11. The HC, Thanh DP, Holt KE, Thomson NR, Baker SJNRM. The genomic signatures of
704 *Shigella* evolution, adaptation and geographical spread. *Nature Reviews Microbiology*.
705 2016;14(4):235.
706 12. Livio S, Strockbine NA, Panchalingam S, Tennant SM, Barry EM, Marohn ME, et al.
707 *Shigella* isolates from the global enteric multicenter study inform vaccine development.
708 *Clinical Infectious Diseases*. 2014;59(7):933-41.
709 13. Group OW. Monitoring the incidence and causes of diseases potentially transmitted by
710 food in Australia: Annual report of the OzFoodNet network, 2011. *Communicable diseases*
711 *intelligence quarterly report*. 2015;39(2):E236.
712 14. Connor TR, Barker CR, Baker KS, Weill F-X, Talukder KA, Smith AM, et al. Species-
713 wide whole genome sequencing reveals historical global spread and recent local persistence in
714 *Shigella flexneri*. *Elife*. 2015;4:e07335.
715 15. Taylor D, Echeverria P, Sethabutr O, Pitarangsi C, Leksomboon U, Blacklow N, et al.
716 Clinical and microbiologic features of *Shigella* and enteroinvasive *Escherichia coli* infections
717 detected by DNA hybridization. *Journal of clinical microbiology*. 1988;26(7):1362-6.
718 16. Levine MM. *Escherichia coli* that cause diarrhea: enterotoxigenic, enteropathogenic,
719 enteroinvasive, enterohemorrhagic, and enteroadherent. *Journal of infectious Diseases*. 1987
720 Mar 1;155(3):377-89.
721 17. Tai AY, Easton M, Encena J, Rotty J, Valcanis M, Howden BP, et al. A review of the
722 public health management of shigellosis in Australia in the era of culture-independent
723 diagnostic testing. *Australian New Zealand journal of public health*. 2016;40(6):588-91.

724 18. Gomes TA, Elias WP, Scaletsky IC, Guth BE, Rodrigues JF, Piazza RM, et al.
725 Diarrheagenic *Escherichia coli*. *brazilian journal of microbiology* 2016;47:3-30.
726 19. Pasqua M, Michelacci V, Di Martino ML, Tozzoli R, Grossi M, Colonna B, et al. The
727 Intriguing Evolutionary Journey of Enteroinvasive *E. coli* (EIEC) toward Pathogenicity.
728 *Frontiers in microbiology*. 2017;8:2390.
729 20. Herzig CT, Fleischauer AT, Lackey B, Lee N, Lawson T, Moore ZS, et al. Notes from
730 the Field: Enteroinvasive *Escherichia coli* Outbreak Associated with a Potluck Party—North
731 Carolina, June–July 2018. *Morbidity and Mortality Weekly Report*. 2019;68(7):183.
732 21. Pettengill EA, Hoffmann M, Binet R, Roberts RJ, Payne J, Allard M, et al. Complete
733 genome sequence of enteroinvasive *Escherichia coli* O96: H19 associated with a severe
734 foodborne outbreak. *Genome Announcements*. 2015;3(4):e00883-15.
735 22. Escher M, Scavia G, Morabito S, Tozzoli R, Maugliani A, Cantoni S, et al. A severe
736 foodborne outbreak of diarrhoea linked to a canteen in Italy caused by enteroinvasive
737 *Escherichia coli*, an uncommon agent. *Epidemiology and infecttion*. 2014;142(12):2559-66.
738 23. Dhakal R, Wang Q, Lan R, Howard P, Sintchenko VJJomm. Novel multiplex PCR
739 assay for identification and subtyping of enteroinvasive *Escherichia coli* and differentiation
740 from *Shigella* based on target genes selected by comparative genomics. *Journal of medical*
741 *microbiology*. 2018;67(9):1257-64.
742 24. Van den Beld M, Reubsaet FJEjocm, diseases i. Differentiation between *Shigella*,
743 enteroinvasive *Escherichia coli* (EIEC) and noninvasive *Escherichia coli*. *European Journal of*
744 *Clinical Microbiology & Infectious Diseases*. 2012;31(6):899-904.
745 25. Pupo GM, Lan R, Reeves PRJPotNAoS. Multiple independent origins of *Shigella*
746 clones of *Escherichia coli* and convergent evolution of many of their characteristics.
747 *Proceedings of the National Academy of Sciences* 2000;97(19):10567-72.
748 26. Lan R, Alles MC, Donohoe K, Martinez MB, Reeves PRJI, immunity. Molecular
749 evolutionary relationships of enteroinvasive *Escherichia coli* and *Shigella spp*. *Infection and*
750 *immunity*. 2004;72(9):5080-8.
751 27. Sahl JW, Morris CR, Emberger J, Fraser CM, Ochieng JB, Juma J, et al. Defining the
752 phylogenomics of *Shigella* species: a pathway to diagnostics. *Journal of clinical microbiology*.
753 2015;53(3):951-60.
754 28. Pettengill EA, Pettengill JB, Binet RJFim. Phylogenetic analyses of *Shigella* and
755 enteroinvasive *Escherichia coli* for the identification of molecular epidemiological markers:
756 whole-genome comparative analysis does not support distinct genera designation. *Frontiers in*
757 *microbiology*. 2016;6:1573.

758 29. Cheasty T, Rowe BJJocm. Antigenic relationships between the enteroinvasive
759 *Escherichia coli* O antigens O28ac, O112ac, O124, O136, O143, O144, O152, and O164 and
760 *Shigella* O antigens. *Journal of clinical microbiology*. 1983;17(4):681-4.
761 30. Landersjö C, Weintraub A, Widmalm GJCr. Structure determination of the O-antigen
762 polysaccharide from the enteroinvasive *Escherichia coli* (EIEC) O143 by component analysis
763 and NMR spectroscopy. *Carbohydrate research*. 1996;291:209-16.
764 31. Linnerborg M, Weintraub A, Widmalm GJEjob. Structural studies of the O-antigen
765 polysaccharide from the enteroinvasive *Escherichia coli* O164 cross-reacting with *Shigella*
766 dysenteriae type 3. *European journal of biochemistry*. 1999;266(2):460-6.
767 32. Sansonetti P, d'Hauteville H, Écobichon Ct, Pourcel C, editors. Molecular comparison
768 of virulence plasmids in *Shigella* and enteroinvasive *Escherichia coli*. Annales de l'Institut
769 Pasteur/Microbiologie; 1983: Elsevier.
770 33. Lan R, Lumb B, Ryan D, Reeves PRJI, immunity. Molecular Evolution of Large
771 Virulence Plasmid in *Shigella* Clones and Enteroinvasive *Escherichia coli*. *Infection and*
772 *immunity*. 2001;69(10):6303-9.
773 34. Venkatesan MM, Buysse JM, Kopecko DJJJoCM. Use of *Shigella flexneri* ipaC and
774 ipaH gene sequences for the general identification of *Shigella spp*. and enteroinvasive
775 *Escherichia coli*. *Journal of Clinical Microbiology*. 1989;27(12):2687-91.
776 35. Hale TLJM, Reviews MB. Genetic basis of virulence in *Shigella* species. Microbiology
777 *Microbiology and Molecular Biology Reviews*. 1991;55(2):206-24.
778 36. Jin Q, Yuan Z, Xu J, Wang Y, Shen Y, Lu W, et al. Genome sequence of *Shigella*
779 *flexneri* 2a: insights into pathogenicity through comparison with genomes of *Escherichia coli*
780 K12 and O157. *Nucleic acids research*. 2002;30(20):4432-41.
781 37. Yang F, Yang J, Zhang X, Chen L, Jiang Y, Yan Y, et al. Genome dynamics and
782 diversity of *Shigella* species, the etiologic agents of bacillary dysentery. *Nucleic acids*
783 *research*. 2005;33(19):6445-58.
784 38. Yang J, Nie H, Chen L, Zhang X, Yang F, Xu X, et al. Revisiting the molecular
785 evolutionary history of *Shigella spp*. *Journal of molecular evolution*. 2007;64(1):71-9.
786 39. Hyma KE, Lacher DW, Nelson AM, Bumbaugh AC, Janda JM, Strockbine NA, et al.
787 Evolutionary genetics of a new pathogenic *Escherichia* species: *Escherichia albertii* and
788 related *Shigella boydii* strains. *Journal of bacteriology*. 2005;187(2):619-28.
789 40. Walters LL, Raterman EL, Grys TE, Welch RA. Atypical *Shigella boydii* 13 encodes
790 virulence factors seen in attaching and effacing *Escherichia coli*. *FEMS Microbiol Lett*.
791 2012;328(1):20-5.

792 41. Hazen TH, Leonard SR, Lampel KA, Lacher DW, Maurelli AT, Rasko DAJI, et al.
793 Investigating the relatedness of enteroinvasive *Escherichia coli* to other *E. coli* and *Shigella*
794 isolates by using comparative genomics. *Infection and immunity*. 2016;84(8):2362-71.
795 42. Silva RM, Toledo M, Trabulsi LRJJoCM. Biochemical and cultural characteristics of
796 invasive *Escherichia coli*. *Journal of Clinical Microbiology*. 1980;11(5):441-4.
797 43. van den Beld MJ, Friedrich AW, van Zanten E, Reubsaet FA, Kooistra-Smid MA,
798 Rossen JWJJomm. Multicenter evaluation of molecular and culture-dependent diagnostics for
799 *Shigella* species and Entero-invasive *Escherichia coli* in the Netherlands. *Journal of*
800 *microbiological methods*. 2016;131:10-5.
801 44. Van Lint P, De Witte E, Ursi J, Van Herendael B, Van Schaeren JJDm, disease i. A
802 screening algorithm for diagnosing bacterial gastroenteritis by real-time PCR in combination
803 with guided culture. *Diagnostic microbiology*. 2016;85(2):255-9.
804 45. De Boer RF, Ott A, Kesztyüs B, Kooistra-Smid AMJJocm. Improved detection of five
805 major gastrointestinal pathogens by use of a molecular screening approach. *Journal of clinical*
806 *microbiology*. 2010;48(11):4140-6.
807 46. Liu B, Knirel YA, Feng L, Perepelov AV, Senchenkova SyN, Wang Q, et al. Structure
808 and genetics of *Shigella* O antigens. *FEMS microbiology reviews*. 2008;32(4):627-53.
809 47. Wattiau P, Boland C, Bertrand S. Methodologies for *Salmonella enterica ssp enterica*
810 subtyping: gold standards and alternatives. *Applied and environmental microbiology*.
811 2011:AEM. 05527-11.
812 48. Cai H, Lu L, Muckle C, Prescott J, Chen S. Development of a novel protein microarray
813 method for serotyping *Salmonella enterica* strains. *Journal of clinical microbiology*.
814 2005;43(7):3427-30.
815 49. van der Ploeg CA, Rogé AD, Bordagorría XL, de Urquiza MT, Castillo ABC, Bruno
816 SB. Design of Two Multiplex PCR Assays for Serotyping *Shigella flexneri*. *Foodborne*
817 *pathogens and disease*. 2018;15(1):33-8.
818 50. Sun Q, Lan R, Wang Y, Zhao A, Zhang S, Wang J, et al. Development of a multiplex
819 PCR assay targeting O-antigen modification genes for molecular serotyping of *Shigella*
820 *flexneri*. *Journal of clinical microbiology*. 2011;49(11):3766-70.
821 51. Li Y, Cao B, Liu B, Liu D, Gao Q, Peng X, et al. Molecular detection of all 34 distinct
822 O-antigen forms of *Shigella*. J Med Microbiol. 2009;58(Pt 1):69-81.
823 52. van den Beld MJ, de Boer RF, Reubsaet FA, Rossen JW, Zhou K, Kuiling S, et al.
824 Evaluation of a Culture-Dependent Algorithm and a Molecular Algorithm for Identification of

825 *Shigella spp*., *Escherichia coli*, and Enteroinvasive *E. coli*. *Journal of clinical microbiology*.
826 2018;56(10):e00510-18.
827 53. Pavlovic M, Luze A, Konrad R, Berger A, Sing A, Busch U, et al. Development of a
828 duplex real-time PCR for differentiation between *E. coli* and *Shigella spp*. *Journal of applied*
829 *microbiology*. 2011;110(5):1245-51.
830 54. Løbersli I, Wester AL, Kristiansen Å, Brandal LTJEJoM, Immunology. Molecular
831 differentiation of *Shigella spp*. from enteroinvasive *E. coli*. *European Journal of Microbiology*.
832 2016;6(3):197-205.
833 55. Kim H-J, Ryu J-O, Song J-Y, Kim H-YJFp, disease. Multiplex polymerase chain
834 reaction for identification of shigellae and four *Shigella* species using novel genetic markers
835 screened by comparative genomics. *Foodborne pathogens*. 2017;14(7):400-6.
836 56. Chattaway MA, Schaefer U, Tewolde R, Dallman TJ, Jenkins CJJocm. Identification of
837 *Escherichia coli* and *Shigella* species from whole-genome sequences. *Journal of clinical*
838 *microbiology*. 2017;55(2):616-23.
839 57. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-
840 aware quantification of transcript expression. *Nature Methods*. 2017;14(4):417-9.
841 58. Wood DE, Salzberg SLJGb. Kraken: ultrafast metagenomic sequence classification
842 using exact alignments. *Genome biology*. 2014;15(3):R46.
843 59. Alikhan N-F, Zhou Z, Sergeant MJ, Achtman M. A genomic overview of the
844 population structure of *Salmonella*. *PLoS genetics*. 2018;14(4):e1007261.
845 60. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al.
846 SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing.
847 *Journal of computational biology*. 2012;19(5):455-77.
848 61. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for
849 genome assemblies. *Bioinformatics*. 2013;29(8):1072-5.
850 62. Jolley KA, Maiden MCJ. BIGSdb: Scalable analysis of bacterial genome variation at
851 the population level. *BMC Bioinformatics*. 2010;11(1):595.
852 63. Joensen KG, Tetzschner AM, Iguchi A, Aarestrup FM, Scheutz FJJocm. Rapid and
853 easy *in silico* serotyping of *Escherichia coli* isolates by use of whole-genome sequencing data.
854 *Journal of clinical microbiology*. 2015;53(8):2410-26.
855 64. Hu D, Liu B, Wang L, Reeves PR. Living Trees: High-Quality Reproducible and
856 Reusable Construction of Bacterial Phylogenetic Trees. *Molecular Biology and Evolution*.
857 2019;37(2):563-75.

858 65. Letunic I, Bork PJNar. Interactive Tree Of Life (iTOL) v4: recent updates and new
859 developments. *Nucleic acids research*. 2019;47(W1):W256-W9.
860 66. Zhou Z, Alikhan N-F, Sergeant MJ, Luhmann N, Vaz C, Francisco AP, et al.
861 GrapeTree: visualization of core genomic relationships among 100,000 bacterial pathogens.
862 *Genome research*. 2018;28(9):1395-404.
863 67. Li HJapa. Aligning sequence reads, clone sequences and assembly contigs with BWA-
864 MEM. *arXiv preprint arXiv*. 2013.
865 68. Buchrieser C, Glaser P, Rusniok C, Nedjari H, d'Hauteville H, Kunst F, et al. The
866 virulence plasmid pWR100 and the repertoire of proteins secreted by the type III secretion
867 apparatus of *Shigella flexneri*. *Molecular microbiology*. 2000;38(4):760-71.
868 69. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence
869 Alignment/Map format and SAMtools. *Bioinformatics* 2009;25(16):2078-9.
870 70. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic
871 features. *Bioinformatics*. 2010;26(6):841-2.
872 71. Seemann T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics*.
873 2014;30(14):2068-9.
874 72. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MT, et al. Roary: rapid
875 large-scale prokaryote pan genome analysis. *Bioinformatics*. 2015;31(22):3691-3.
876 73. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al.
877 BLAST+: architecture and applications. *BMC bioinformatics*. 2009;10(1):421.
878 74. Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, et al. The RAST
879 Server: rapid annotations using subsystems technology. *BMC genomics*. 2008;9(1):75.
880 75. Sof'ya NS, Feng L, Yang J, Shashkov AS, Cheng J, Liu D, et al. Structural and genetic
881 characterization of the *Shigella boydi*i type 10 and type 6 O antigens. *Journal of bacteriology*.
882 2005;187(7):2551-4.
883 76. Ansaruzzaman M, Kibriya A, Rahman A, Neogi P, Faruque A, Rowe B, et al. Detection
884 of provisional serovars of *Shigella* dysenteriae and designation as *S*. dysenteriae serotypes 14
885 and 15. *Journal of clinical microbiology*. 1995;33(5):1423-5.
886 77. Balows AJDM, Disease I. Manual of clinical microbiology 8th edition: PR Murray, EJ
887 Baron, JH Jorgenson, MA Pfaller, and RH Yolken, eds., ASM Press, 2003, 2113 pages, 2 vol,
888 2003+ subject & author indices, ISBN: 1-555810255-4, US $189.95. *Diagnostic Microbiology*.
889 2003;47(4):625.

890 78. Woodward DL, Clark CG, Caldeira RA, Ahmed R, Soule G, Bryden L, et al.
891 Identification and characterization of *Shigella boydii* 20 serovar nov., a new and emerging
892 *Shigella* serotype. *J Med Microbiol*. 2005;54(8):741-8.
893 79. Kim J, Lindsey RL, Garcia-Toledo L, Loparev VN, Rowe LA, Batra D, et al. High-
894 quality whole-genome sequences for 59 historical *Shigella* strains generated with PacBio
895 sequencing. *Genome Announcements*. 2018;6(15).
896 80. Michelacci V, Prosseda G, Maugliani A, Tozzoli R, Sanchez S, Herrera-León S, et al.
897 Characterization of an emergent clone of enteroinvasive *Escherichia coli* circulating in Europe.
898 *Clinical Microbiology*. 2016;22(3):287. e11-. e19.
899 81. Liu J, Platts-Mills JA, Juma J, Kabir F, Nkeze J, Okoi C, et al. Use of quantitative
900 molecular diagnostic methods to identify causes of diarrhoea in children: a reanalysis of the
901 GEMS case-control study. *The Lancet*. 2016;388(10051):1291-301.
902 82. Cho MS, Ahn T-Y, Joh K, Kwon O-S, Jheong W-H, Park DSJJMB. A novel marker for
903 the species-specific detection and quantitation of *Shigella sonnei* by targeting a methylase
904 gene. *J Microbiol Biotechnol*. 2012;22(8):1113-7.
905 83. Sansonetti PJ, Kopecko DJ, Formal SBJI, immunity. *Shigella sonnei* plasmids:
906 evidence that a large plasmid is necessary for virulence. *Infection and immunity*.
907 1981;34(1):75-83.
908 84. Feng L, Perepelov AV, Zhao G, Shevelev SD, Wang Q, Sof'ya NS, et al. Structural and
909 genetic evidence that the *Escherichia coli* O148 O antigen is the precursor of the *Shigella*
910 *dysenteriae* type 1 O antigen and identification of a glucosyltransferase gene. *Microbiology*.
911 2007;153(1):139-47.
912 85. Göhmann S, Manning P, Alpert C-A, Walker M, Timmis KJMp. Lipopolysaccharide
913 O-antigen biosynthesis in *Shigella* dysenteriae serotype 1: analysis of the plasmid-carried rfp
914 determinant. *Microbial pathogenesis*. 1994;16(1):53-64.
915
916

917 **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.

924 **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

925

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

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

929

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

931 932

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%

933 934 935 936 937

938

939 [a]: Identification dataset has 90 non-*Shigella*/EIEC strains including 72 ECOR strains and 18 *E.albertii* strains. 1,969 assembled genomes and
940 1,951 reads (reads not available for 18 EIEC isolates downloaded from NCBI) in identification dataset. One of *E.albertii* strain was assigned as
941 SB13 which was grouped into SB13 cluster on the phylogenetic tree.

942

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

944 945 946 947 948 949 950 951

952

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

954

955

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

956 957 958 959 960 961 962 963 964 965 966 967

968

969 **Figure legends:**

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

971 Representative isolates from the identification dataset were used to construct the phylogenetic
972 tree by Quicktree v1.3 (64) to identify *Shigella* and EIEC clusters and visualised by
973 Grapetree's interactive mode. The dendrogram tree shows the phylogenetic relationships of
974 1879 *Shigella* and EIEC isolates represented in the identification dataset. Branch lengths are
975 log scale for clarity. The tree scales indicated the 0.2 substitutions per locus. *Shigella* and EIEC
976 clusters are coloured. Numbers in square brackets indicate the number of isolates of each
977 identified cluster. CSP is sporadic EIEC lineages.

978

979 **Figure 2: *in silico* serotyping pipeline workflow**

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

983

984 **Supplementary Material**

985 **Figure S1**: **Identification phylogenetic tree**

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

992

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

994 A confirmation phylogenetic tree was constructed by Quicktree v1.3 (64) based on 2375
995 isolates and visualised by Grapetree's interactive mode. The tree shows the phylogenetic
996 relationships between identified *Shigella*/EIEC clusters in identification dataset and non-
997 enteroinvasive E. coli isolates. Branch lengths are log scale for clarity. The tree scales
998 indicated the 0.1 substitutions per locus. Known *Shigella* and EIEC clusters from identification
999 dataset are colored. Numbers in square brackets indicate the number of isolates of each
1000 identified cluster. CSP is sporadic EIEC lineages.

1001

1002 **Figure S2-B**: **Confirmation phylogenetic tree**

1003 A confirmation phylogenetic tree constructed by Quicktree v1.3 (64) and visualised by ITOL
1004 v5 shows the phylogenetic relationships between identified *Shigella*/EIEC clusters in
1005 identification dataset and non-enteroinvasive E. coli isolates. The tree scales indicated the 0.01
1006 substitutions per locus. *Shigella* and EIEC clusters are colored. The internal branches are
1007 colored to represent the bootstrap values. Green color indicates the maximum bootstrap value
1008 (1). The red color shows the minimum bootstrap value (0). Each of cluster is well supported by
1009 bootstrap value. CSP is sporadic EIEC lineages.

1010

1011 **Figure S3**: **Distribution of mapped 38 virulence genes in 58 sporadic isolates**

1012 The presence of *Shigella* virulence plasmid pINV in 58 sporadic isolates in identification
1013 dataset was determined by the mapped 38 virulence genes. Detailed genes were described in
1014 Results "**Investigation of *Shigella* virulence plasmid pINV in 58 sporadic isolates"**. Three
1015 categories were defined based on the number of virulence genes mapped to isolate. Virulence
1016 plasmid positive: > 25 genes mapped to isolate; Intermediate: 13 to 25 genes mapped to isolate;
1017 Virulence plasmid negative: less than 13 genes mapped to isolate.

1018

1019 **Figure S4 (A)**: **Validation phylogenetic tree**

1020 A validation tree was generated by Quicktree v1.3 (64) and visualised by Grapetree's
1021 interactive mode to assign representative isolates in validation dataset to clusters. Branch
1022 lengths are log scale for clarity. The tree scales indicated the 0.2 substitutions per locus.
1023 Known *Shigella* and EIEC clusters from identification dataset are colored. Numbers in square
1024 brackets indicate the number of isolates of each identified cluster. Isolates in validation dataset
1025 are colored white. The isolates are assigned to clusters if they grouped into known cluster
1026 isolates. CSP is sporadic EIEC lineages.

1027

1028 **Figure S4 (B)**: **Validation phylogenetic tree**

1029 A validation phylogenetic tree was constructed by Quicktree v1.3 (64) and visualised by ITOL
1030 v5 to assign representative isolates in validation dataset to clusters. The tree scales indicated
1031 the 0.01 substitutions per locus. *Shigella* and EIEC clusters are colored. The internal branches
1032 are colored to represent the bootstrap values. Green color indicates the maximum bootstrap
1033 value (1). The red color shows the minimum bootstrap value (0). Each of cluster is well
1034 supported by bootstrap value. Isolates that grouped with known cluster isolates (from
1035 identification dataset) with strong bootstrap support are categorised into that cluster. CSP is
1036 sporadic EIEC lineages.

1037

1038 **Table S1**: 1,969 isolates used in identification dataset

1039 **Table S2**: 15,501 isolates used in validation dataset

1040 **Table S3**: The location and function of cluster-specific genes

1041 **Table S4**: The results of cluster-specific gene markers tested with 12,743 non-enteroinvasive
1042 E. coli isolates

1043

1044 **Data S1**: Algorithms incorporated into the ShigEiFinder

1045 **Data S2**: Genetic signature O and H genes from ShigaTyper and SerotypeFinder

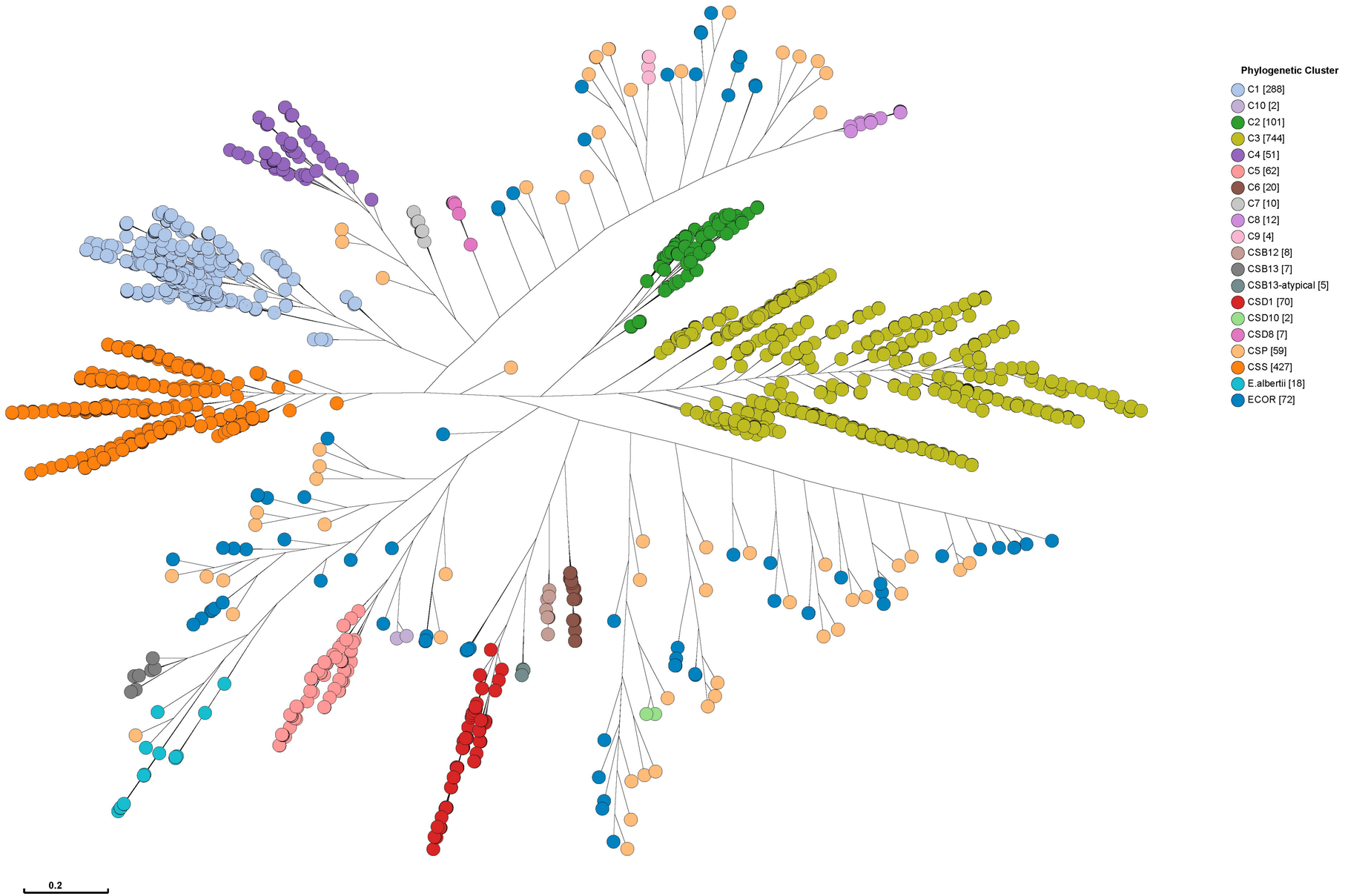
1046

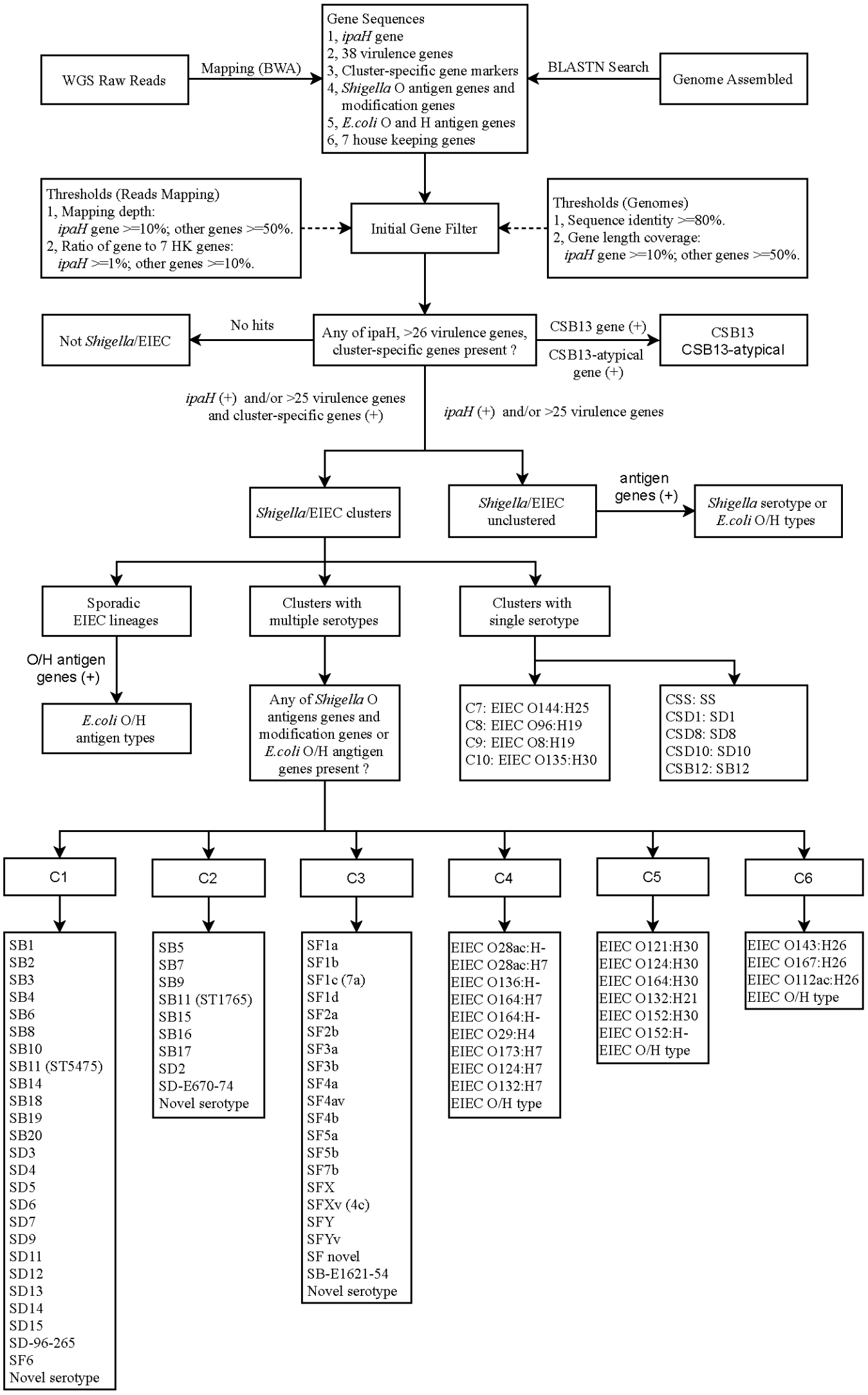
1047 **Data Availability Statement**

1048

1049 Custom python scripts used in this study are available from the authors on request.

1050


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