Dataset S1: Additional scripts information:

Usage and further details can be found in the scripts folder at github.com/LanLab/ShigEiFinder.

clade_specific_gene_combinations.py :

This script was used to identify specific gene sets for each cluster from the pan genomes of the identification dataset. The script ran on one cluster at a time. The script takes in 4 inputs, a roary presence absence file, a genome cluster assignment file, the genomes of all isolates, the annotated genes in all genomes (as used in roary). The script first identified individual candidate genes that were present in all isolates of the target cluster (true positives) and were present only in a percentage of non-target cluster isolates (false positives). For the list of candidates each combination of genes was tested to see whether all are found in the same false positive strain. If a set of genes are never all found together then that set of genes is reported as a result. The size of the gene combinations starts at 1 for the whole list and increases progressively. At each size, successful sets of genes were reported until the total number of reported sets equals the maximum specified in the settings. Additionally, if a successful set of 2 genes (for example) was found within a subsequent set of 3 genes that three gene set was excluded because the additional gene provides no benefit.

extract_gene_sequences_from_roary.py:

This script extracts specific gene set sequences for sets produced by *clade_specific_gene_combinations.py.* The script accepted 4 inputs: the presence absence roary output csv, the annotated genes in all genomes (as used in roary), a list of cluster specific genes sets and their corresponding cluster, a list of genome ids and their corresponding cluster. An output prefix is also required. The script will:

- select a representative genome from each cluster
- identify the roary orthologue group that contains a given specific gene
- retrieve the gene ID for that orthologue group and the representative genome
- extract the gene from the genes fasta file for that genome
- save the specific gene to an output file (output prefix)
- produce a summary file of genes retrieved (output prefix)

The selection of cluster/lineage-specific gene markers after initial screening

- Obtain the list of genes for each set (Specific_genes_groupID.txt) from the output file after running script: clade_specific_gene_combinations-fnfp.py.
- Extract the sequences of genes using script: prokka_genome_gene_from_roary.py.
- Run blastn against identification dataset with the sequence identity of 80% to check for truncated orthologues which are not evaluated in roary.
- Gene length filtering for blastn output: >=50% length coverage.
- Check the number of FN and FP for each cluster/lineage-specific gene set (the output file from running clade_specific_gene_combinations-fnfp.py), combined with the blastn results, the gene set with the lowest FN and FPs was selected.

Dataset S2: Algorithms incorporated into the ShigEiFinder

ShigEiFinder stands for *Shigella* EIEC Cluster Enhanced Serotype Finder and is a cluster-specific gene marker based *in silico* pipeline developed for differentiation of *Shigella* and enteroinvasive *E. coli* (EIEC) and serotyping of *Shigella* and EIEC. ShigEiFinder is available as a web tool (https://mgtdb.unsw.edu.au/ShigEiFinder/) and on github (https://github.com/LanLab/ShigEiFinder).

Note that for brevity, in all references to *Shigella* serotypes below, *Shigella sonnei*, *Shigella flexneri*, *Shigella boydii* and *Shigella 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. *Shigella dysenteriae* serotype 1 is abbreviated as SD1.

Typing reference sequences used in ShigEiFinder

The typing reference sequences consisted of cluster-specific gene markers and sporadic EIEC lineages specific gene markers from this study, *ipaH* gene, 38 virulence genes, *Shigella* serotype specific O antigen genes collected from ShigatTyper (1), *E.coli* O antigen genes and *fliC* genes collected from SerotypeFinder (2) and 7 House Keeping (HK) genes from the MLST (3) scheme.

The cluster-specific gene marker sets and sporadic EIEC lineages specific gene markers are listed as supplementary material with file name in Table S3. The 38 virulence genes are listed in "Analysis of the 59 sporadic EIEC isolates" section in the main text. *Shigella* and *E.coli* O and H antigen genes are listed as supplementary material with file name in Data S3.

All sequences are listed in fasta format available at https://github.com/LanLab/ShigEiFinder.

ShigEiFinder input

Either paired end Illumina sequencing reads or assembled genomes are acceptable.

ShigEiFinder output

ShigEiFinder output included the sample, presence of *ipaH* gene, number of virulence genes, cluster assignment, serotype, *E. coli* O and H antigen present and any further notes for the result in a tabular format.

Runtime and memory requirements

The average run time is approximately 0.89s per genome in which the average size of a genome was approximately 4.4 MB on a machine with 4 threads and 32Gb RAM.
Average script runtime for WGS reads was approximately 1.5 minutes on a machine with 4 threads and 32 Gb RAM.

Determination of presence or absence of genes

The presence or absence of genes were determined by the cutoff value of gene length coverage for assembled genomes and the mapping length percentage and the ratio of mean mapping depth to the average mean mapping depth of 7 HK genes (Table 1). For example, the *ipaH* gene was defined as present if mapping length coverage was over 10% together with the ratio of mean depth to the average mean depth of 7 HK genes was over 1% from reads mapping.

Table 1: Thresholds used for determination of genes present or absent

Typing reference genes	Genomes	Reads mapping	
	Gene length coverage	Mapping length coverage	Ratio to 7 HK
ipaH gene	10%	10%	1%
Virulence genes	50%	50%	10%
Cluster-specific gene markers	50%	50%	10%
O antigen and H antigen genes	50%	50%	10%

Algorithms for cluster assignment and serotyping

The *Shigella* or EIEC cluster assignment was determined by the presence of cluster-specific gene marker set that was only found within a single *Shigella* or EIEC cluster. Where marker set was used to identify a cluster, all genes must be present for a cluster to be called. ShigEiFinder also used 38 virulence genes from the pINV invasive plasmid to determine whether the plasmid was present in the isolate. When more than 25 of these genes were present, the isolate was considered to be pINV positive.

The presence of cluster-specific gene marker sets combined with the presence of *ipaH* gene and/or virulence genes the isolate was assigned to *Shigella* or EIEC cluster (Table 2).

The isolate assigned as *Shigella* or EIEC unclustered could be any new cluster that cannot be detected by any of cluster-specific gene marker set. Unclustered *Shigella* or EIEC isolate could also be those that all genes in the markers set were present but one or more of the genes from the markers set have mapping ratio between 1% and 10% and do not meet the cutoff for presence and therefore are classified as unclustered (11 isolates of 15,501 isolates in validation dataset were in that category).

Table 2: The cluster-specific gene markers based cluster assignment

Cluster assignment	*ipaH* gene	>=26 virulence genes	cluster-specific gene/set
Shigella or EIEC clusters	+	+/-	+
Shigella or EIEC unclustered	+	+/-	-
SB13 or SB13-atypical	-	-	+
Not *Shigella*/EIEC	-	-	-

"+": gene presence; "+/-": can be present or absent; "-": gene absence.

The serotype is then assigned based on the presence of *Shigella* serotype specific O antigen genes and modification genes or *E. coli* O and H antigen genes. A "novel serotype" is assigned if there is no match to known serotypes.

Low level contamination check and notes for unclustered *Shigella* or EIEC isolates

The gene markers with mapping ratio between 1% and 10% demonstrated that the genes in the genomes may not be sequenced very well or a potential contamination. In such cases ShigEiFinder will write out a note "Possible contamination by *Shigella* or EEIC strain or low cluster-specific gene mapping depth to HK genes in cluster [cluster name]".

The genes may have mapping ratio between 1% and 10% are listed in Table 3.

Table 3: Gene markers with mapping ratio between 1% and 10%

Gene markers	Number of isolates of 15,501 isolates
C1_gene_2	8
C1_gene_4	1
C5_gene_1	1
CSS_gene_3	1

Additional subsets of gene markers used for *Shigella* or EIEC clusters assignment

To increase the accuracy of typing, we added additional subsets of genes to eliminate the known false presences for cluster-specific gene markers. For example, the combination of C1 specific markers set and CSB12 specific gene marker can distinguish CSB12 from C1, if both cluster specific genes are present, the isolate is assigned CSB12 while if CSB12 specific gene is absent, the isolate is assigned as C1. There are 6 subsets of combined genes incorporated into the ShigEiFinder for elimination of false cluster assignment (Table 4).

Table 4: Subsets of combined gene markers for elimination of false cluster assignment

Subset 1	C1 markers set	CSB12 gene	Cluster Assignment
Isolate	+	+	CSB12
Isolate	+	-	C1
Isolate	-	+	CSB12
Subset 2	C1 markers set	CSD1 markers set	Cluster Assignment
Isolate	+	+	CSD1
Isolate	+	-	C1
Isolate	-	+	CSD1
Subset 3	C1 markers set	C2 markers set	Cluster Assignment
Isolate	+	+	C2
Isolate	+	-	C1
Isolate	-	+	C2
Subset 4	C3 markers set	C5 markers set	Cluster Assignment
Isolate	+	+	C3
Isolate	+	-	C3
Isolate	-	+	C5
Subset 5	C5 markers set	C8 markers set	Cluster Assignment
Isolate	+	+	C8
Isolate	+	-	C5
Isolate	-	+	C8
Subset 6	C2 markers set	CSS markers set	Cluster Assignment
Isolate	+	+	C2
Isolate	+	-	C2
Isolate	-	+	CSS

"+": gene presence; "-": gene absence.

Serotyping SB1 and SB20 within C1

SB1 and SB20 share identical O antigen genes. For better differentiation of SB1 from SB20, we analysed C1 subbranch on the identification tree (Fig.1 in main text). The 21 isolates with presence of SB1 wzx and wzy genes were grouped into one subbranch which consisted of 2 lineages, lineage I and lineage II as Table 5.

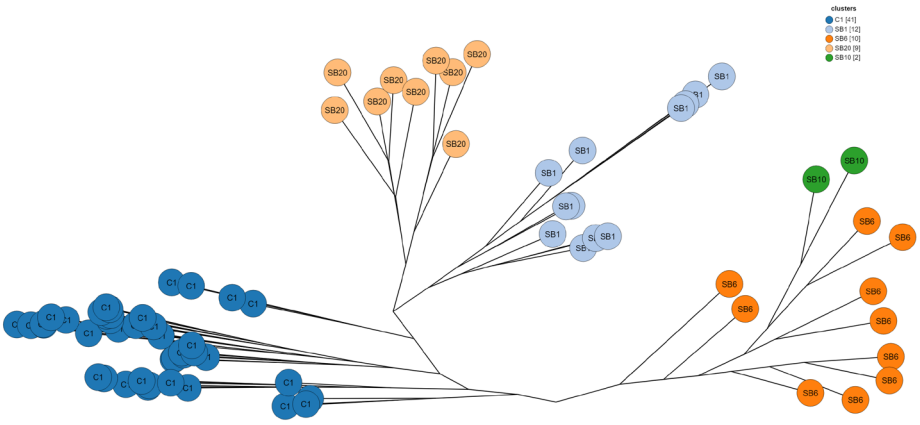
Table 5: The distribution of SB1 and SB20 isolates in two lineages

Lineages	ShigaTyper assignment				
	SB1	SB20	EIEC	Untypeable	Total
Lineage I	11	0	1	2[a]	14
Lineage II	4	2	2	1	9[b]

[a]: One isolate with the presence of heparinase gene which was used in ShigaTyper to separate SB20 form SB1.
[b]: All 9 isolates had heparinase gene either full length or fragments by BLASTN search.

HierBAPS (4) analysis was further performed to confirm the 2 lineages. Lineage I was defined as potential SB1 lineage and lineage II was defined as SB20 lineage (Figure below). Based on phylogenetic analysis, we identified an SB20 specific gene by comparing 288 accessory genomes in C1 from the identification dataset. The gene was validated with *Shigella* and EIEC validation dataset C1 isolates. The isolate was assigned as SB20 with the presence of SB20 specific gene and SB1 wzx/wzy genes, otherwise the isolate was SB1 with the only SB1 wzx/wzy genes present.

Figure: Subbranch of C1 on identification tree



Serotyping SB6 and SB10 within C1

SB6 and SB10 share identical O antigen genes but there are SNP differences in the O antigen gene clusters. The SNP in SB10 wzx and SB10 wzy genes at positions 904 and 141 respectively were used to separate SB6 from SB10. For assembled genomes, we first checked

the SNP positions that were covered in the blast search with 100% identify for SB10. The isolate was classified as SB10 if the SB10 SNPs were present. Otherwise, the isolate was assigned as SB6. Samtools mpileups was used to gather the nucleotide base at the SNP positions for reads mapping. The isolate was SB10 if the SB10-SNPs was found. The absence of the SNP was assigned as SB6.

Serotyping EIEC O164/O124

The *E.coli* O164 and O124 O antigen genes are near identical with > 99.4% identity (5). There was a 2-base indel (a frame shift mutation (6)) at positions 429 and 430 in *wfeP* gene of O164 in comparison to O124. We used this indel to differentiate O164 from O124. The isolate was assigned as O164 if the indel was found.

Multiple variants of H antigens

There are multiple variants for one type of H antigen. To assign an H type when multiple H variants are present, the highest match was chosen as the H antigen present.

SF serotyping within C3

C3 contains all SF serotypes except for SF6 which is grouped into C1. We used the established scheme of SF O antigen genes and modification genes including *gtr*, *oac* and *opt* genes to type SF within C3 (7-19) (Table 6). ShigEiFInder assigned all possibilities when there was a multiple match of combinations of modification genes. The isolate was classified as SFY if there was only backbone O antigen genes present. While the isolate was assigned as SF novel serotype if no match to known serotypes and the note was given with the presence or absence of genes.

Table 6: The combination of O antigen genes and modification genes used for SF serotyping

	wzx$_{1-5}$	*wzx*$_6$	*gtrI*	*gtrIC*	*gtrII*	*gtrIV*	*gtrV*	*gtrX*	*oac*	*oac1b*	*oacB*	*oacC*	*oacD*	*optII*	*optIII*
SF1a	+	-	+	-	-	-	-	-	-	-	+/-	-	-	-	-
SF1b	+	-	+	-	-	-	-	-	-	+	+/-	-	-	-	-
SF1c(7a)	+	-	+	+	-	-	-	-	-	-	-	-	-	-	-
SF1d	+	-	+	-	-	-	-	+	-	-	-	-	-	-	-
SF2a	+	-	-	-	+	-	-	-	-	-	+/-	-	+/-	-	-
SF2b	+	-	-	-	+	-	-	+	-	-	-	-	+/-	-	-

SF3a	+	-	-	-	-	-	-	+	+/-	-	-	-	+/-	-	-
SF3b	+	-	-	-	-	-	-	-	+	-	-	-	-	-	-
SF4a	+	-	-	-	-	+	-	-	-	-	-	-	-	-	-
SF4av	+	-	-	-	-	+	-	-	-	-	-	-	-	-	+
SF4b	+	-	-	-	-	+	-	-	+	-	-	-	-	-	-
SF5a	+	-	-	-	-	-	+	-	-	-	+	-	-	-	-
SF5b	+	-	-	-	-	-	+	+	-	-	-	-	-	-	-
SF7b	+	-	-	+	-	-	-	-	+	-	-	-	-	-	-
SFX	+	-	-	-	-	-	-	+	-	-	-	-	+/-	-	-
SFXv(4c)	+	-	-	-	-	-	-	+	-	-	-	-	+/-	+	-
SFY	+	-	-	-	-	-	-	-	-	-	+/-	-	+/-	-	-
SFYv	+	-	-	-	-	-	-	-	-	-	-	-	+	+	+
SF6	-	+	-	-	-	-	-	-	-	-	-	+/-	-	-	-

"+": gene presence and highlighted in pink color.

"+/-": can be present or absent.

"-": gene absence.

Reference:

1. Wu Y, Lau HK, Lee T, Lau DK, Payne J. *in Silico* Serotyping Based on Whole-Genome Sequencing Improves the Accuracy of *Shigella* Identification. *Applied and environmental microbiology*. 2019;85(7).
2. Joensen KG, Tetzschner AM, Iguchi A, Aarestrup FM, Scheutz F. 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.
3. Jolley KA, Maiden MC. BIGSdb: Scalable analysis of bacterial genome variation at the population level. *BMC Bioinformatics*. 2010;11:595.
4. Cheng L, Connor TR, Sirén J, Aanensen DM, Corander J. Hierarchical and spatially explicit clustering of DNA sequences with BAPS software. *Molecular biology and evolution*. 2013;30(5):1224-8.
5. Liu B, Furevi A, Perepelov AV, Guo X, Cao H, Wang Q, et al. Structure and genetics of *Escherichia coli* O antigens. *FEMS microbiology reviews*. 2020;44(6):655-83.
6. Liu B, Knirel YA, Feng L, Perepelov AV, Senchenkova SN, Wang Q, et al. Structure and genetics of *Shigella* O antigens. *FEMS microbiology reviews*. 2008;32(4):627-53.
7. Adhikari P, Allison G, Whittle B, Verma NK. Serotype 1a O-antigen modification: molecular characterization of the genes involved and their novel organization in the *Shigella flexneri* chromosome. *Journal of bacteriology*. 1999;181(15):4711-8.
8. Stagg RM, Tang SS, Carlin NI, Talukder KA, Cam PD, Verma NK. A novel glucosyltransferase involved in O-antigen modification of *Shigella flexneri* serotype 1c. *Journal of bacteriology*. 2009;191(21):6612-7.
9. Mavris M, Manning PA, Morona R. Mechanism of bacteriophage SfII-mediated serotype conversion in *Shigella flexneri*. *Molecular microbiology*. 1997;26(5):939-50.
10. Adams MM, Allison GE, Verma NK. Type IV O antigen modification genes in the genome of *Shigella flexneri* NCTC 8296. Microbiology (Reading, England). 2001;147(Pt 4):851-60.
11. Huan PT, Bastin DA, Whittle BL, Lindberg AA, Verma NK. Molecular characterization of the genes involved in O-antigen modification, attachment, integration and excision in *Shigella flexneri* bacteriophage SfV. Gene. 1997;195(2):217-27.

12. Verma NK, Verma DJ, Huan PT, Lindberg AA. Cloning and sequencing of the glucosyl transferase-encoding gene from converting bacteriophage X (SFX) of *Shigella flexneri*. Gene. 1993;129(1):99-101.
13. Clark CA, Beltrame J, Manning PA. The *oac* gene encoding a lipopolysaccharide O-antigen acetylase maps adjacent to the integrase-encoding gene on the genome of *Shigella flexneri* bacteriophage Sf6. Gene. 1991;107(1):43-52.
14. Sun Q, Lan R, Wang Y, Wang J, Xia S, Wang Y, et al. Identification of a divergent O-acetyltransferase gene oac 1b from *Shigella flexneri* serotype 1b strains. Emerging microbes & infections. 2012;1(9):e21.
15. Wang J, Knirel YA, Lan R, Senchenkova SN, Luo X, Perepelov AV, et al. Identification of an O-acyltransferase gene (*oacB*) that mediates 3- and 4-O-acetylation of rhamnose III in *Shigella flexneri* O antigens. *Journal of bacteriology*. 2014;196(8):1525-31.
16. Knirel YA, Wang J, Luo X, Senchenkova SN, Lan R, Shpirt AM, et al. Genetic and structural identification of an O-acyltransferase gene (*oacC*) responsible for the 3/4-O-acetylation on rhamnose III in *Shigella flexneri* serotype 6. BMC microbiology. 2014;14:266.
17. Sun Q, Knirel YA, Wang J, Luo X, Senchenkova SN, Lan R, et al. Serotype-converting bacteriophage SfII encodes an acyltransferase protein that mediates 6-O-acetylation of GlcNAc in *Shigella flexneri* O-antigens, conferring on the host a novel O-antigen epitope. *Journal of bacteriology*. 2014;196(20):3656-66.
18. Sun Q, Knirel YA, Lan R, Wang J, Senchenkova SN, Jin D, et al. A novel plasmid-encoded serotype conversion mechanism through addition of phosphoethanolamine to the O-antigen of *Shigella flexneri*. PloS one. 2012;7(9):e46095.
19. Knirel YA, Lan R, Senchenkova SN, Wang J, Shashkov AS, Wang Y, et al. O-antigen structure of *Shigella flexneri* serotype Yv and effect of the *lpt-O* gene variation on phosphoethanolamine modification of *S. flexneri* O-antigens. Glycobiology. 2013;23(4):475-85.

Dataset S3: *Shigella*/EIEC serotypes specific O and H antigens used in ShigEiFinder

Shigella serotype specific O antigen genes were collected from ShigaTyper (1). *E.coli* O antigen genes and *fliC* genes were collected from SerotypeFinder (2).

Sequences	Accession number
SD1_*wzx*, SD1_*wzy*	L07293
SD1-rfp	CP000640
SD2_*wzx*, SD2_*wzy*	EU296404
SD3_*wzx*, SD3_*wzy*	EU296415
SD4_*wzx*, SD4_*wzy*	EU296402
SD5_*wzx*, SD5_*wzy*	EU294174
SD6_*wzx*, SD6_*wzy*	EU296414
SD7_*wzx*, SD7_*wzy*	AY380835
SD8_*wzx*, SD8_*wzy*	EU294166
SD9_*wzx*, SD9_*wzy*	EU296416
SD10_*wzx*, SD10_*wzy*	EU294178
SD11_*wzx*, SD11_*wzy*	EU294172
SD12_*wzx*, SD12_*wzy*	EU294169
SD13_*wzx*, SD13_*wzy*	EU294167
SD14_*wzx*, SD14_*wzy*	CP026832
SD15_*wzx*, SD15_*wzy*	CP026834
SDP 96-265_*wzx*, *wzy*	CP026819
SDP E670-74_*wzx*,*wzy*	CP027027
SB1_*wzx*, SB1_*wzy*	AY630255
SB2_*wzx*, SB2_*wzy*	EU296418
SB3_*wzx*, SB3_*wzy*	EU296407
SB4_*wzx*, SB4_*wzy*	AF402312
SB5_*wzx*, SB5_*wzy*	AF402313
SB6_*wzx*, SB6_*wzy*	AF402314
SB7_*wzx*, SB7_*wzy*	EU296411
SB8_*wzx*, SB8_*wzy*	EU294163

SB9_*wzx*, SB9_*wzy*	AF402315
SB10_*wzx*, SB10_*wzy*	AY693427
WbaM	AY693427
SB11_*wzx*, SB11_*wzy*	AY529126
SB12_*wzx*, SB12_*wzy*	EU296406
SB13_*wzx*, SB13_*wzy*	AY369140
SB14_*wzx*, SB14_*wzy*	EU296409
SB15_*wzx*, SB15_*wzy*	EU296412
SB16_*wzx*, SB16_*wzy*	DQ371800
SB17_*wzx*, SB17_*wzy*	DQ875941
SB18_*wzx*, SB18_*wzy*	AY948196
SB19_*wzx*, SB19_*wzy*	CP026814
Heparinase	CP016036
SBP E1621-54_*wzx*,*wzy*	CP026810
SF $wzx_{1\text{-}5}$ gene	AE005674
SF6 *wzx* gene	EU294165
SF *gtrI*	AF139596
SF *gtrIC*	FJ905303
SF *gtrII*	AF021347
SF *gtrIV*	AF288197
SF *gtrV*	U82619
SF *gtrX*	L05001
SF *oacA*	AF547987
SF *oac1b*	JN377795
SF *oacB*	NC_004337 (SF0315)
SF *oacC*	AKMW01000058
SF *oacD*	NC_004337 (SF0309)
SF *optIII*	KC020049
SF Xv *optII*	CP001385 (SFxv_5135)
SS_*wzx*, SS_*wzy*	AF285971
O1_*wzx*, O1_*wzy*	GU299791

O2_*wzx*, O2_*wzy*	EU549863
O4_*wzx*, O4_*wzy*	AY568960
O6_*wzx*, O6_*wzy*	AJ426045
O7_*wzx*, O7_*wzy*	AF125322
O8_*wzx*, O8_*wzy*	AF013583
O8_*wzm*, O8_*wzt*	AB010150
O12_*wzx*, O12_*wzy*	AB811600
O13_*wzx*, O13_*wzy*	EU296422
O16_*wzx*, O16_*wzy*	AB811601
O17_*wzx*, O17_*wzy*	AB812084
O18ac_*wzx*, O18ac_*wzy*	AB811603
O21_*wzx*, O21_*wzy*	EU694098
O22_*wzx*, O22_*wzy*	AB811606
O25_*wzx*, O25_*wzy*	GU014554
O26_*wzx*, O26_*wzy*	AF529080
O28ac_*wzx*, O28ac_*wzy*	DQ462205
O29_*wzx*, O29_*wzy*	EU294173
O32_*wzx*, O32_*wzy*	EU296410
O36_*wzx*, O36_*wzy*	AB811613
O39_*wzx*, O39_*wzy*	AB811616
O40_*wzx*, O40_*wzy*	EU296417
O50_*wzx*, O50_*wzy*	AB811624
O53_*wzx*, O53_*wzy*	EU289392
O71_*wzx*, O71_*wzy*	GU445927
O77_*wzx*, O77_*wzy*	AB972416
O79_*wzx*, O79_*wzy*	EU294162
O86_*wzx*, O86_*wzy*	AY220982
O89_*wzx*, O89_*wzy*	AB812038
O92_*wzx*, O92_*wzy*	AB812040
O93_*wzx*, O93_*wzy*	AB812041
O96_*wzx*, O96_*wzy*	AB812043
O102_*wzx*, O102_*wzy*	JX087966

O105_*wzx*, O105_*wzy*	EU294171
O110_*wzx*, O110_*wzy*	AB812049
O111_*wzx*, O111_*wzy*	JN887675
O112ab_*wzx*, O112ab _*wzy*	EU296413
O112ac_*wzx*, O112ac_*wzy*	EU296405
O117_*wzx*, O117_*wzy*	EU694096
O118_*wzx*, O118_*wzy*	HM204927
O121_*wzx*, O121_*wzy*	JN859209
O124_*wzx*, O124_*wzy*	EU296419
O129_*wzx*, O129_*wzy*	EU296424
O130_*wzx*, O130_*wzy*	EU296421
O132_*wzx*, O132_*wzy*	AB812056
O135_*wzx*, O135_*wzy*	EU296423
O136_*wzx*, O136_*wzy*	AB812059
O143_*wzx*, O143_*wzy*	EU294164
O144_*wzx*, O144_*wzy*	AB812062
O147_*wzx*, O147_*wzy*	DQ868766
O148_*wzx*, O148_*wzy*	DQ167407
O149_*wzx*, O149_*wzy*	DQ868764
O151_*wzx*, O151_*wzy*	HM204926
O152_*wzx*, O152_*wzy*	EU294170
O155_*wzx*, O155_*wzy*	AY657020
O162_*wzx*, O162_*wzy*	AB812067
O164_*wzx*, O164_*wzy*	EU296420
O167_*wzx*, O167_*wzy*	EU296408
O173_*wzx*, O173_*wzy*	GU068046
O180_*wzx*, O180_*wzy*	JQ751058
O183_*wzx*, O183_*wzy*	AB627352
H1_*fliC*	AB028471
H2_*fliC*	AIHA01000023
H4_*fliC*	AJ605764
H4_*fliC*	AJ605765

H4_*fliC*	AJ536600
H5_*fliC*	AY249990
H5_*fliC*	AY337469
H6_*fliC*	AIEY01000041
H7_*fliC*	AY337468
H7_*fliC*	AKML01000326
H7_*fliC*	ANLT01000257
H7_*fliC*	ANLJ01000383
H7_*fliC*	AOES01000098
H7_*fliC*	AMVH01000352
H7_*fliC*	AF228487
H7_*fliC*	AF228496
H7_*fliC*	AF228495
H7_*fliC*	AB334575
H7_*fliC*	AB334574
H7_*fliC*	AF228494
H7_*fliC*	AF228491
H7_*fliC*	AF228492
H7_*fliC*	AB028474
H8_*fliC*	AJ865465
H9_*fliC*	AY249994
H10_*fliC*	AF169320
H11_*fliC*	AY337472
H12_*fliC*	AY337471
H14_*fliC*	AY249998
H16_*fliC*	AB128919
H16_*fliC*	JH954529
H16_*fliC*	JH953794
H16_*fliC*	AY337476
H16_*fliC*	AY337477
H16_*fliC*	AY337475 AY2500001
H16_*fliC*	AY250000

H18_*fliC*	AY250001
H19_*fliC*	AY337479
H19_*fliC*	AY250002
H20_*fliC*	AY250003
H21_*fliC*	AIHL01000060
H24_*fliC*	K72 (H25w)
H25_*fliC*	AGSG01000116
H26_*fliC*	AY250008
H26_*fliC*	AY337483
H27_*fliC*	AY250009
H28_*fliC*	AY250010
H30_*fliC*	AY250011
H30_*fliC*	AY337483
H31_*fliC*	AY250013
H33_*fliC*	AY250015
H40_*fliC*	AJ884568
H42_*fliC*	AY250021
H45_*fliC*	AY250023
H48_*fliC*	AY250025
H49_*fliC*	AY250026
H51_*fliC*	AY250027

Reference:

1. Wu Y, Lau HK, Lee T, Lau DK, Payne J. *in Silico* Serotyping Based on Whole-Genome Sequencing Improves the Accuracy of *Shigella* Identification. *Applied and environmental microbiology*. 2019;85(7).
2. Joensen KG, Tetzschner AM, Iguchi A, Aarestrup FM, Scheutz F. 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.