AMERICAN SOCIETY FOR MICROBIOLOGY Applied and Environmental Microbiology®

PUBLIC AND ENVIRONMENTAL HEALTH MICROBIOLOGY

In Silico Serotyping Based on Whole-Genome Sequencing Improves the Accuracy of *Shigella* Identification

Yun Wu,[a,b] Henry K. Lau,[b] Teresa Lee,[b] David K. Lau,[b] Justin Payne[c]

[a]U.S. Food and Drug Administration, Office of Commissioner, Commissioner's Fellowship Program, Silver Spring, Maryland, USA
[b]U.S. Food and Drug Administration, Office of Regulatory Affairs, San Francisco Laboratory, Alameda, California, USA
[c]U.S. Food and Drug Administration, Center for Food Safety and Applied Nutrition, College Park, Maryland, USA

ABSTRACT Bacteria of the genus *Shigella*, consisting of 4 species and >50 serotypes, cause shigellosis, a foodborne disease of significant morbidity, mortality, and economic loss worldwide. Classical *Shigella* identification based on selective media and serology is tedious, time-consuming, expensive, and not always accurate. A molecular diagnostic assay does not distinguish *Shigella* at the species level or from enteroinvasive *Escherichia coli* (EIEC). We inspected genomic sequences from 221 *Shigella* isolates and observed low concordance rates between conventional designation and molecular serotyping: 86.4% and 80.5% at the species and serotype levels, respectively. Serotype determinants for 6 additional serotypes were identified. Examination of differentiation gene markers commonly perceived as characteristic hallmarks in *Shigella* showed high variability among different serotypes. Using this information, we developed ShigaTyper, an automated workflow that utilizes limited computational resources to accurately and rapidly determine 59 *Shigella* serotypes using Illumina paired-end whole-genome sequencing (WGS) reads. *Shigella* serotype determinants and species-specific diagnostic markers were first identified through read alignment to an in-house curated reference sequence database. Relying on sequence hits that passed a threshold level of coverage and accuracy, serotype could be unambiguously predicted within 1 min for an average-size WGS sample of ~500 MB. Validation with WGS data from 380 isolates showed an accuracy rate of 98.2%. This pipeline is the first step toward building a comprehensive WGS-based analysis pipeline of *Shigella* spp. in a field laboratory setting, where speed is essential and resources need to be more cost-effectively dedicated.

IMPORTANCE *Shigella* causes diarrheal disease with serious public health implications. However, conventional *Shigella* identification methods are laborious and time-consuming and can be erroneous due to the high similarity between *Shigella* and enteroinvasive *Escherichia coli* (EIEC) and cross-reactivity between serotyping antisera. Further, serotype interpretation is complicated for inexperienced users. To develop an easier method with higher accuracy based on whole-genome sequencing (WGS) for *Shigella* serotyping, we systematically examined genomic information of *Shigella* isolates from 53 serotypes to define rules for differentiation and serotyping. We created ShigaTyper, an automated pipeline that accurately and rapidly excludes non-*Shigella* isolates and identifies 59 *Shigella* serotypes using Illumina paired-end WGS reads. A serotype can be unambiguously predicted at a data processing speed of 538 MB/min with 98.2% accuracy from a regular laptop. Once it is installed, training in bioinformatics analysis and *Shigella* genetics is not required. This pipeline is particularly useful to general microbiologists in field laboratories.

KEYWORDS *Shigella*, *in silico*, serotying, whole-genome sequencing

Citation Wu Y, Lau HK, Lee T, Lau DK, Payne J. 2019. *In silico* serotyping based on whole-genome sequencing improves the accuracy of *Shigella* identification. Appl Environ Microbiol 85:e00165-19. https://doi.org/10.1128/AEM.00165-19.

Editor Edward G. Dudley, The Pennsylvania State University

Address correspondence to Yun Wu, Yun.Wu@fda.hhs.gov, or Henry K. Lau, Henry.Lau@fda.hhs.gov.

Received 22 January 2019
Accepted 25 January 2019
Accepted manuscript posted online 1 February 2019
Published 22 March 2019

Bacteria of the genus *Shigella* cause bacillary dysentery (shigellosis), one of the leading diarrheal diseases worldwide, disproportionately affecting children under 5 years of age from low- and middle-income countries (1–3). *Shigella* is transmitted through the fecal-oral route at an extremely low infectious dose (4) and manifests clinical symptoms, including fever, abdominal pain, watery or bloody diarrhea, vomiting, and potentially death (5). Although primarily a disease of the poor, shigellosis is still a public health concern in developed countries. An estimated 450,000 cases occur annually in the United States (6), bringing an economic loss of $257 million (7). The actual impact may be higher, as culture-based diagnosis underestimates shigellosis ~2-fold (2), and a substantial increase in culture-confirmed shigellosis cases was reported in recent years (8). There is no licensed shigellosis vaccine, and >90% of *Shigella* isolates are antimicrobial resistant (9), leaving those exposed at risk.

Shigella consists of 4 species (serogroups) and >54 serotypes: *Shigella dysenteriae* (15 serotypes), *S. flexneri* (18 serotypes), *S. boydii* (20 serotypes), and *S. sonnei* (1 serotype). These serotypes are distinguished solely through the somatic (O) antigen, or lipopolysaccharide, expressed on the cell surface. *Shigella* is believed to acquire the O antigen from commensal *Escherichia coli* strains (10). *Shigella* serodiversity is further expanded through acquiring genes from other enteric bacteria and mobile genetic elements to either lose or replace the O-antigen biosynthetic genes or modify the O antigen (10–15).

It is important to determine the distribution of *Shigella* species and serotypes in time and space for disease burden tracking and outbreak investigation and to inform and evaluate policies aimed for disease reduction and vaccine development (16). Conventional *Shigella* identification relies on a combination of biochemical and serological assessment. Biochemical assays are conducted to distinguish *Shigella* from *E. coli*, the results of which roughly identify *Shigella* to the species level. Serological testing (slide agglutination) follows to determine the serotype. Serological differentiation is essential but is laborious, time-consuming, and expensive and can be erroneous. Intra- and interspecies cross-reactivity is common, and commercial antisera are at best 91% accurate (17). Rough strains that do not express O antigen and newly emerged *Shigella* serotypes without antisera that recognize them are nontypeable, accounting for 6 to 10% of annual *Shigella* isolates in the United States (8).

Molecular typing has been in development to replace conventional *Shigella* identification. The multilocus virulence gene *ipaH*, which has been employed by many institutions as a molecular target for *Shigella* (18, 20), does not differentiate *Shigella* and enteroinvasive *E. coli* (EIEC), a virulence clade of *E. coli* that shares many biochemical properties and virulence genes with *Shigella* (21, 22). Serotypes of the same *Shigella* species are not necessarily genetically closer than those from another species. Consequently, methods relying on genetic relatedness often cannot successfully place *Shigella* into clearly segregated clades by species (23–27). Although multilocus sequence typing (MLST) showed promise in *Shigella* classification (28), some MLST sequence types (STs) consist of multiple serotypes, which can lead to loss of critical information for vaccine development, as immunity against *Shigella* O-antigen is associated with protection from shigellosis (29). Molecular assays directly targeting the O-antigen-specific biosynthetic genes, such as PCR-restriction fragment length polymorphism (RFLP) (30), multiplex PCR (31, 32), and microarray (33), have been developed. However, these methods require additional biochemical assays to differentiate *Shigella* from *E. coli*, as many *Shigella* serotypes share identical surface O antigens with commensal *E. coli* (10).

Whole-genome sequencing (WGS) is a promising technology to replace conventional assays for microbial typing. With the cost of WGS decreasing precipitously (34), it is increasingly used in clinical diagnosis and disease surveillance. The bottleneck for adopting WGS, however, resides in WGS analysis, a skill not often possessed by analysts trained as general microbiologists. Furthermore, interpretation of *Shigella* serotypes is complicated, as it is determined by the combination of O-antigen synthesis and

TABLE 1 Summary of *Shigella* WGS development set used in this study[a]

Strain designation	No. of strains	No. (%) with concordant species designation	No. of serotypes included	No. of strains with serotype designation	No. (%) with concordant serotype designation
S. boydii	97	79 (81.4)	21	87	68 (78.2)
S. dysenteriae	55	48 (87.3)	15	37	31 (83.8)
S. flexneri	49	47 (95.9)	13	42	33 (78.6)[b]
S. sonnei	19	17 (89.5)	1	19	17 (89.5)
Shigella sp.	1				
EIEC	13	13 (100)			
Non-*Shigella*/EIEC	25	25 (100)			
Shigella only	221	191 (86.4)	50	185	149 (80.5)
Total	259	229 (88.4)			

[a] Strains were sequenced from an in-house collection ($n = 58$) or their WGSs were downloaded from the NCBI ($n = 201$).
[b] Partial agreement between designation and *in silico* serotyping for *S. flexneri* was considered concordance (e.g., serotype 5 versus 5a).

modification enzymes. An easy, simple serotyping pipeline with a user-friendly interface is needed for a WGS-based *Shigella* surveillance program.

Genoserotyping of *Shigella* requires information on both genetic determinants for serotype and those that differentiate from *E. coli*, particularly EIEC. Nevertheless, *Shigella*-specific genetic markers were often only studied in common serotypes but not rare serotypes. As *Shigella* underwent convergent evolution to arrive at similar phenotypes (23, 27, 35), conclusions drawn from type strains cannot reflect all serotypes. Here we report a comprehensive examination of *Shigella* genomic data covering 53 different serotypes, from which we derived results for the development of an *in silico* serotyping pipeline, ShigaTyper, that can make a direct prediction for 59 *Shigella* serotypes. ShigaTyper was specifically designed to meet the need of general microbiologists in field laboratories, where resources for *Shigella* identification are often limited. Such a pipeline is especially useful when species and serotype information is essential in quickly identifying organisms in outbreak situations.

RESULTS

We examined genetic determinants from a development set of 48 genome assemblies and the raw reads of 221 *Shigella* isolates, 56 of which were generated in-house and 165 were downloaded from the National Center for Biotechnology Institute (NCBI), collectively representing 53 different serotypes. There were 97 isolates designated *S. boydii* (including 6, 4, 4, 2, 2, 5, 2, 4, 7, 7, 5, 3, 3, 3, 3, 2, 3, 6, 4, 7, and 2 isolates typed to serotypes 1 to 20 and E1621-54, respectively, and 10 untyped isolates), 55 isolates designated *S. dysenteriae* (including 4, 5, 2, 3, 3, 1, 1, 2, 2, 2, 2, 3, 3, 2, and 2 isolates typed to serotypes 1 to 15, respectively, and 18 untyped isolates), 49 isolates designated *S. flexneri* (including 2, 1, 3, 2, 2, 7, 3, 3, 2, 2, 3, 1, 1, 3, 1, 2, and 3 isolates typed to serotypes Y, X, 1a, 1b, 1c [7a], 2a, 2b, 3a, 3b, 3, 4a, 4bv, 4, 5b, 5, 6, and provisional, respectively, and 7 untyped isolates), and 19 isolates designated *S. sonnei*. Additionally, 38 isolates of 14 Gram-positive and -negative foodborne bacteria were used as an exclusion group, including 13 EIEC isolates, 8 non-EIEC *E. coli* isolates (including 1 Shiga toxin-producing *E. coli* [STEC] isolate), 2 enterobacterial species that share O antigen with *Shigella* (*Escherichia albertii* and *Plesiomonas shigelloides*), 8 other enterobacteria (*Salmonella enterica*, *Klebsiella pneumoniae*, *Enterobacter cloacae*, and *Yersinia enterocolitica*), 3 nonenterobacterial Gram-negative diarrheagenic bacteria (*Vibrio parahaemolyticus* and *Campylobacter jejuni*), and 4 Gram-positive pathogens (*Staphylococcus aureus*, *Listeria monocytogenes*, and *Enterococcus faecium*). The distribution of species and serotypes of these strains (the development set) is described in Table 1 and in Tables S1 and S2.

Genetic determinants for *Shigella* serotypes. Genetic determinants for most *Shigella* O antigens have been characterized. Within the O-antigen biosynthetic gene cluster (*rfb*), genes encoding O-antigen flippase, *wzx*, and polymerase, *wzy*, are serotype specific; their sequences were obtained from published reports (10, 12–15, 36–53). For serotypes without O-antigen information, *rfb* sequence located between the conserved *galF* and *gnd* genes was first extracted from assembled genomes (54). Sequences of *wzx* and *wzy* on *rfb* were determined by gene annotation, BLAST search, and protein secondary structure analysis. *S. dysenteriae* 14 and *S. boydii* 19 each possessed a unique *rfb* sequence with no close homolog in another bacterial species. The *rfb* sequences from *S. dysenteriae* 15 and provisional serotypes 96-265, E670-74, and *S. boydii* E1621-54 were nearly identical to those of enterotoxigenic *E. coli* (ETEC) OgN15 strain E819, *E. albertii* strain SP140152, *E. coli* O170, and *E. coli* O7. *S. boydii* 20 shares identical *rfb* sequences with *S. boydii* 1. A chromosomally carried *rfb* gene was not found for *S. dysenteriae* 93-119 and 204-96.

To further differentiate between serotypes, we collected the sequences of O-antigen modification enzymes of *S. flexneri* (12, 15, 52, 53), a chromosomally encoded *S. sonnei*-specific putative methylase (this sequence is hereafter referred to as *Ss_methylase*) (55), *wbaM* of *S. boydii* 10 (48), and the plasmid-borne *rfp* of *S. dysenteriae* 1 (14, 38). Genome comparison of *S. boydii* 1 and 20 revealed a unique, nonchromosomal heparinase gene in all *S. boydii* 20 isolates ($n = 7$) but not *S. boydii* 1 ($n = 6$), which we tentatively included as the *S. boydii* 20 marker.

Source, coordinates, and references for sequences included in the reference sequence database are listed in Table 2.

Genetic determinants to differentiate *Shigella* and EIEC. Differentiation of *Shigella* from *E. coli* is an indispensable part of *Shigella* identification. The highly conserved 3′ end of *ipaH* genes (*ipaH_C*) was employed as an indicator for *Shigella*/EIEC (56–62). Most *Shigella* organisms are impaired for lactose fermentation and lysine decarboxylation. Therefore, sequences of *E. coli lacY* (*EclacY*) and lysine decarboxylase (*cadA*) were used as differentiation markers for *Shigella* from EIEC. These sequences were also included in the reference sequence database (Table 2).

We hypothesized that genetic markers *ipaH_C*, *EclacY*, *cadA*, and *Ss_methylase* can be used for EIEC differentiation and *Shigella* identification. As the same defective phenotype in *Shigella* can be caused by different types of mutations and in different genes, we sought to identify exceptions in the *Shigella* and EIEC genomes, summarized in Table 3.

(i) *ipaH*. We detected *ipaH_C* in all *Shigella* and EIEC strains as expected, except *S. boydii* 13 ($n = 3$). This is consistent with previous findings that *S. boydii* 13 is noninvasive and more closely related to *E. albertii* than to *Shigella* (35, 63).

(ii) *lacY*. *EclacY* was reported to be absent from *Shigella* organisms except *S. sonnei* and *S. dysenteriae* 1 while present in most EIEC isolates (64). A remnant from the 5′ end of *EclacY* was detected in 21 of 22 *S. sonnei* (107 to 270 bp) and 8 of 8 *S. dysenteriae* 1 (361 to 475 bp) genomes, respectively. We observed full-length *EclacY* in 4 *S. boydii* serotype 9 isolates ($n = 7$) and 373 bp of 5′-*EclacY* in the other 3. *S. boydii* serotype 15 ($n = 3$) carried nearly full-length *EclacY* except a 72-bp deletion at the 5′ end. Eleven EIEC isolates carried full-length *EclacY* ($n = 14$).

(iii) *cadA*. It was reported that *cadA* was deleted in most *Shigella* isolates but present in the genomes of 70% of EIEC isolates (65). We observed that all *S. sonnei* ($n = 22$) and *S. dysenteriae* ($n = 8$) isolates carry full-length *cadA* as previously reported (66, 67). *S. dysenteriae* 8 ($n = 5$) also harbored full-length *cadA*. *S. dysenteriae* 10 ($n = 3$) carried a 3′ remnant of *cadA*. Among *S. boydii* 11 isolates ($n = 5$), 4 harbored full-length *cadA* and 1 carried a 258-bp remnant at the 3′ end. Ten EIEC isolates harbored full-length *cadA* ($n = 14$).

(iv) *Ss_methylase*. We detected *Ss_methylase* in all 22 *S. sonnei* genomes. However, *Ss_methylase* was also detected in all *S. dysenteriae* 10 ($n = 3$) and 2 EIEC ($n = 14$) isolates that we examined.

Virulence factors. Virulence of *Shigella*/EIEC is attributed to pINV (56), an invasion plasmid that carries genes allowing enteroinvasion. Sequence of *ipaB*, an essential gene for invasion, was included as a marker for pINV. Shiga toxin expressed from *S. dysenteriae* 1 is associated with hemolytic-uremic syndrome. Both type 1 (stx_1) (57–60) and type 2 (stx_2) (61) Shiga toxins have been reported for other *Shigella* serotypes. Therefore, we included the reference sequences of stx_1 (stx/stx_{1a}) and stx_2 (stx_{2a}) (62). In

TABLE 2 Sources of sequences included in the reference sequence database for *Shigella* serotyping[a]

Sequence	Source	Accession no.	Gene identifier(s)	Beginning position	Ending position	Complement	Length (bp)	Reference(s)
ipaH_C	*S. flexneri* 2a 301 chromosome	AE005674.2	*ipaH_7*	2686919	2687698		780	91
ipaB	*S. sonnei* pSS_046	CP000039.1	*ipaB*	84468	86210	x	1,743	92
Ss_methylase	*S. sonnei* SS046 chromosome	CP000038.1	SSON_1583	1663456	1665288		1,833	55
Ss_wzx	*S. sonnei* pSS_046	CP000039.1	*wzx*	197086	198354		1,269	13, 92
Ss_wzy	*S. sonnei* pSS_046	CP000039.1	*wzy*	198423	199595		1,173	13, 92
Sf_wzx	*S. flexneri* 2a 301 chromosome	AE005674.2	*rfbE*	2113945	2115201	x	1,257	36, 91
Sf_wzy	*S. flexneri* 2a 301 chromosome	AE005674.2	*rfc*	2110959	2112107	x	1,149	36, 91
Sf6_wzx	*S. flexneri* 6 NCTC9779 O-Ag cluster	EU118169	*wzx*	4509	5741		1,233	10
Sf6_wzy	*S. flexneri* 6 NCTC9779 O-Ag cluster	EU118169	*wzy*	5797	6984		1,188	10
GtrI	*S. flexneri* 1a 0439 chromosome	CP020342.1	BS647_00210	24836	25787	x	952	39, 42
GtrII	*S. flexneri* 2a 301 chromosome	AE005674.2	*gtrII*	319152	320612		1,461	91
GtrIV	*Shigella* phage SfIV	NC_022749	V416_gp25	20408	21721	x	1,314	93
GtrV	*S. flexneri* 5 8401 chromosome	CP000266	*gtrV*	274667	275920	x	1,254	40, 94
GtrX	*S. flexneri* 4c 1205 chromosome	CP012140	AD871_01870	346525	347790	x	1,266	95
GtrIC	*S. flexneri* 7b 3007	CP024473	Unannotated	4289665	4291260		1,596	96
Oac	Enterobacterial phage Sf6	AF547987	15	15624	16625	x	1,002	37, 97
Oac1b	Partial sequence from *S. flexneri* 1b	JF450728.1	*oac*	2673	3674		1,002	53
Xv	pSFxv_2 from strain 2002017	NC_017320.1	SFXV_RS26800	1523	3043	x	1,521	15, 98
Sd1_wzx	*S. dysenteriae* 1 Sd197 chromosome	CP000034	*rfbX*	2010259	2011449		1,191	14, 99
Sd1_wzy	*S. dysenteriae* 1 Sd197 chromosome	CP000034	*rfc*	2011446	2012588		1,143	14, 99
Sd1_rfp	*S. dysenteriae* 1 pSD197_spA	CP000640	*rfpB*	2979	4112	x	1,134	14, 38, 99
Sd2_wzx	*S. dysenteriae* 2 O-Ag cluster	EU296404	*wzx*	5775	7202		1,428	10
Sd2_wzy	*S. dysenteriae* 2 O-Ag cluster	EU296404	*wzy*	1867	2952		1,086	10
Sd3_wzx	*S. dysenteriae* 3 O-Ag cluster	EU296415	*wzx*	1111	2625		1,515	10
Sd3_wzy	*S. dysenteriae* 3 O-Ag cluster	EU296415	*wzy*	7014	8117		1,104	10
Sd4_wzx	*S. dysenteriae* 4 O-Ag cluster	EU296402	*wzx*	1081	2352		1,272	10
Sd4_wzy	*S. dysenteriae* 4 O-Ag cluster	EU296402	*wzy*	4769	6001		1,233	10
Sd5_wzx	*S. dysenteriae* 5 O-Ag cluster	EU294174	*wzx*	4728	5915		1,188	10
Sd5_wzy	*S. dysenteriae* 5 O-Ag cluster	EU294174	*wzy*	9803	11134		1,332	10
Sd6_wzx	*S. dysenteriae* 6 O-Ag gene cluster	EU296414	*wzx*	340	1614		1,275	10
Sd6_wzy	*S. dysenteriae* 6 O-Ag gene cluster	EU296414	*wffH_5* (including the fusion junction between *wzy* and *wffH*)	3592	4722		1,131	10
Sd7_wzx	*S. dysenteriae* 7 O-Ag gene cluster, strain M1354	AY380835	*wzx*	7707	9194		1,488	45
Sd7_wzy	*S. dysenteriae* 7 O-Ag gene cluster, strain M1354	AY380835	*wzy*	10646	11836		1,191	45
Sd8_wzx	*S. dysenteriae* 8 O-Ag gene cluster	EU294166	*wzx*	3180	4349		1,170	10
Sd8_wzy	*S. dysenteriae* 8 O-Ag gene cluster	EU294166	*wzy*	1899	3125		1,227	10
Sd9_wzx	*S. dysenteriae* 9 O-Ag gene cluster	EU296416	*wzx*	7775	9055		1,281	10
Sd9_wzy	*S. dysenteriae* 9 O-Ag gene cluster	EU296416	*wzy*	5373	6578		1,206	10
Sd10_wzx	*S. dysenteriae* 10 O-Ag gene cluster	EU294178	*wzx*	4706	5866		1,161	10
Sd10_wzy	*S. dysenteriae* 10 O-Ag gene cluster	EU294178	*wzy*	6933	8354		1,422	10
Sd11_wzx	*S. dysenteriae* 11 O-Ag gene cluster	EU294172	*wzx*	3741	5141		1,401	10
Sd11_wzy	*S. dysenteriae* 11 O-Ag gene cluster	EU294172	*wzy*	5138	6301		1,164	10
Sd12_wzx	*S. dysenteriae* 12 O-Ag gene cluster	EU294169	*wzx*	5663	6874		1,212	10
Sd12_wzy	*S. dysenteriae* 12 O-Ag gene cluster	EU294169	*wzy*	7882	8955		1,074	10
Sd13_wzx	*S. dysenteriae* 13 O-Ag cluster	EU294167	*wzx*	5631	7133		1,503	10
Sd13_wzy	*S. dysenteriae* 13 O-Ag cluster	EU294167	*wzy*	11011	12108		1,098	10
Sd14_wzx	*S. dysenteriae* 14 ATCC 49346 chromosome	CP026832	Unannotated	2829534	2830799	x	1,266	54, this study

(Continued on next page)

TABLE 2 (Continued)

Sequence	Source	Accession no.	Gene identifier(s)	Beginning position	Ending position	Complement	Length (bp)	Reference(s)
Sd14_wzy	*S. dysenteriae* 14 ATCC 49346 chromosome	CP026832	Unannotated	2831775	2832977	x	1,203	54, this study
Sd15_wzx	*S. dysenteriae* 15 ATCC 49347 chromosome	CP026834.1	Unannotated	965436	966623	x	1,188	54, this study
Sd15_wzy	*S. dysenteriae* 15 ATCC 49347 chromosome	CP026834.1	Unannotated	966613	967854	x	1,242	54, this study
SdProv_wzx	*S. dysenteriae* provisional 96-265 chromosome	CP026819.1	Unannotated	2893541	2894860		1,320	54, this study
SdProv_wzy	*S. dysenteriae* provisional 96-265 chromosome	CP026819.1	Unannotated	2899151	2900362		1,212	54, this study
SdProvE_wzx	*S. dysenteriae* provisional E670-74 chromosome	CP027027.1	Unannotated	3878013	3879260	x	1,248	54, this study
SdProvE_wzy	*S. dysenteriae* provisional E670-74 chromosome	CP027027.2	Unannotated	3874004	3875176	x	1,173	54, this study
Sb1_wzx	*S. boydii* 1 O-Ag cluster	AY630255	*wzx*	4570	5754		1,185	49
Sb1_wzy	*S. boydii* 1 O-Ag cluster	AY630255	*wzy*	6605	7669		1,065	49
Sb2_wzx	*S. boydii* 2 O-Ag cluster	EU296418	*wzx*	4106	5353		1,248	10
Sb2_wzy	*S. boydii* 2 O-Ag cluster	EU296418	*wzy*	6127	7314		1,188	10
Sb3_wzx	*S. boydii* 3 O-Ag cluster	EU296407	*wzx*	1774	3039		1,266	10
Sb3_wzy	*S. boydii* 3 O-Ag cluster	EU296407	*wzy*	7165	8415		1,251	10
Sb4_wzx	*S. boydii* Sb227 chromosome	CP000036	*wzx*	877026	878255	x	1,230	43, 99
Sb4_wzy	*S. boydii* Sb227 chromosome	CP000036	*wzy*	874635	875927	x	1,293	43, 99
Sb5_wzx	*S. boydii* 5 O-Ag cluster	AF402313	*wzx*	7263	8411		1,149	43
Sb5_wzy	*S. boydii* 5 O-Ag cluster	AF402313	*wzy*	5379	6470		1,092	43
Sb6_wzx	*S. boydii* 6 O-Ag cluster	AF402314	*wzx*	11771	13183	x	1,413	43
Sb6_wzy	*S. boydii* 6 O-Ag cluster	AF402314	*wzy*	3195	4247		1,053	43
WbaM	*S. boydii* 10 O-Ag cluster	AY693427	*wbaM*	9468	10424		957	48
Sb7_wzx	*S. boydii* 7 O-Ag cluster	EU296411	*wzx*	6714	7928		1,215	10
Sb7_wzy	*S. boydii* 7 O-Ag cluster	EU296411	*wzy*	8889	10184		1,296	10
Sb8_wzx	*S. boydii* 8 O-Ag cluster	EU294163	*wzx*	4664	5869		1,206	10
Sb8_wzy	*S. boydii* 8 O-Ag cluster	EU294163	*wzy*	7853	8935		1,083	10
Sb9_wzx	*S. boydii* 9 O-Ag cluster	AF402315	*wzx*	6944	8191		1,248	43
Sb9_wzy	*S. boydii* 9 O-Ag cluster	AF402315	*wzy*	4922	6139		1,218	43
Sb11_wzx	*S. boydii* 11 O-Ag cluster	AY529126	*wzx*	10367	11794	x	1,428	46
Sb11_wzy	*S. boydii* 11 O-Ag cluster	AY529126	*wzy*	5432	6460		1,029	46
Sb12_wzx	*S. boydii* 12 O-Ag cluster	EU296406	*wzx*	6264	7580		1,317	10
Sb12_wzy	*S. boydii* 12 O-Ag cluster	EU296406	*wzy*	7561	8700		1,140	10
Sb13_wzx	*S. boydii* 13 O-Ag cluster	AY369140	*wzx*	1713	3032		1,320	44
Sb13_wzy	*S. boydii* 13 O-Ag cluster	AY369140	*wzy*	3019	4197		1,179	44
Sb14_wzx	*S. boydii* 14 O-Ag cluster	EU296409	*wzx*	1054	2511		1,458	10
Sb14_wzy	*S. boydii* 14 O-Ag cluster	EU296409	*wzy*	2515	3726		1,212	10
Sb15_wzx	*S. boydii* 15 O-Ag cluster	EU296412	*wzx*	1042	2433		1,392	10
Sb15_wzy	*S. boydii* 15 O-Ag cluster	EU296412	*wzy*	2453	3631		1,179	10
Sb16_wzx	*S. boydii* 16 O-Ag cluster	DQ371800	*wzx*	2385	3785		1,401	50
Sb16_wzy	*S. boydii* 16 O-Ag cluster	DQ371800	*wzy*	3796	4953		1,158	50
Sb17_wzx	*S. boydii* 17 O-Ag cluster	DQ875941	*wzx*	926	2191		1,266	51
Sb17_wzy	*S. boydii* 17 O-Ag cluster	DQ875941	*wzy*	4681	5892		1,212	51
Sb18_wzx	*S. boydii* CDC 3083-94 chromosome	CP001063	*wzx*	1103246	1104463		1,218	47, 54
Sb18_wzy	*S. boydii* CDC 3083-94 chromosome	CP001063	*wzy*	1105612	1106757		1,146	47, 54
Sb19_wzx	*S. boydii* 83-578 chromosome	CP026814.1	Unannotated	3586585	3587835		1,251	54, this study
Sb19_wzx	*S. boydii* 83-578 chromosome	CP026814.1	Unannotated	3588633	3589772		1,140	54, this study
SbProv_wzx	*S. boydii* provisional 54-1621 chromosome	CP026810	Unannotated	1495194	1496618		1,425	54, this This study
SbProv_wzy	*S. boydii* provisional 54-1621 chromosome	CP026810	Unannotated	1499881	1501056		1,176	54, this This study
Heparinase	*E. coli* isolate Co6114 plasmid pCo6114_2	CP016036	Heparinase II/III-like protein	2803	4428	x	1,626	This study
EclacY	*E. coli* DH1 (ME8569) chromosome	AP012030	*lacY*	362350	363603		1,254	100

(Continued on next page)

TABLE 2 (Continued)

Sequence	Source	Accession no.	Gene identifier(s)	Beginning position	Ending position	Complement	Length (bp)	Reference(s)
cadA	*S. dysenteriae* 1 Sd197 chromosome	CP000034	*cadA*	4179987	4182129		2,143	99
Stx1	*S. dysenteriae* 1 3818T Shiga toxin sequence	M19437	*stxA*, *stxB*	161	1387		1,227	62
Stx2	*E. coli* O157:H7 EDL933 Stx2 sequence	X07865	*sltIIA*, *sltIIB*	239	1479		1,241	62
ShET1	*S. flexneri* 2a 301 chromosome	AE005674.2	*set1B*, *set1A*	3069555	3070277		723	91, 101
ShET2	*S. sonnei* pSS_046	CP000039.1	*sen*/*ospD2*	5491	7200	x	1,710	92, 102
Sat_N	*S. dysenteriae* 10 ATCC 12039 chromosome	CP026831.1	*espC*	4214643	4217642		3,000	54, 103

[a] Annotated reference genome sequences or O-antigen gene clusters based on which original characterization was published were chosen unless there was a mistake in the sequence (for example, EU118169 was chosen over EU294165, because the EU294165 differs from the rest of *S. flexneri* 6 O-antigen gene cluster sequences by 5 nucleotides). When annotation was not available, sequences were annotated by RAST (87–89) and serotype-specific determinants were identified. Publications from which the sequence was generated and the O-antigen gene cluster was characterized are included as references. Accession numbers refer to GenBank or SRA.

TABLE 3 Exceptions to the inclusion/exclusion criteria used for *Shigella* serotyping[a]

Gene marker	Function and use	Serotype	Description of exception	No. of strains with exception (accession number of strains)	No. of strains examined
ipaH_C	Conserved virulence gene *Shigella*/EIEC inclusion marker	*S. boydii* 13	Absent	3 (SRR4178425, SRR4181329, SRR4181518)	3
EclacY	Lactose permease *Shigella* exclusion marker	*S. boydii* 9	Full-length *lacY*	4 (CP026836.1, SRR4180506, SRR6760302, SRR8186698)	7
			373-bp 5′ end of *lacY*	3 (SRR4176997, SRR4180898, SRR4181342)	
		S. boydii 15	*lacY* with 72-bp deletion at 5′ end	3 (ASM296813v1 [GCF_002968135.1], SRR4179879, SRR8186662)	3
		S. dysenteriae 1	366- to 475-bp 5′ end of *lacY*	8 (CP000034, CP006736, DRR015930, SRR1811629, SRR5330538, SRR6373753, SRR8186696, SRR8186588)	8
		S. sonnei	270-bp 5′ end of *lacY*	21 (CP000038.1, CP023645.1, ERR1762061, ERR1762062, SRR4180904, SRR6927290, SRR6954223, SRR6982834, SRR7013788, SRR7013790, SRR7013792, SRR7013793, SRR7013794, SRR7013797, SRR7013799, SRR8186598, SRR8186617, SRR8186670, SRR8186671, SRR8186733, SRR8186738)	22
			No *lacY*	1 (SRR6927273)	
cadA	Lysine decarboxylase, *Shigella* exclusion marker	*S. boydii* 11	Full-length *cadA*	4 (CP026846.1, SRR4176974, SRR4180810, SRR4180822)	5
			258-bp 3′ end of *cadA*	1 (DRR015925)	
		S. dysenteriae 1	Full-length *cadA*[b]	8 (see above for accession numbers)	8
		S. dysenteriae 8	Full-length *cadA*	6 (CP026827.1, DRR015992, SRR2994193, SRR8186616, SRR8186618, SRR8186667)	6
		S. dysenteriae 10	127- to 171-bp 3′ end of *cadA*	3 (CP026831.1, DRR015994, SRR8186726)	3
		S. sonnei	Full-length *cadA*[c]	22 (see above for accession numbers)	22
Ss_methylase	Putative (DNA) methylase, *S. sonnei*-specific marker	*S. dysenteriae* 10	Full-length Ss_methylase	3 (see above for accession numbers)	3
		EIEC	Full-length Ss_methylase	2 (DRR015801, SRR5330536)	14

[a] Gene markers *ipaH*, *EclacY*, *cadA*, and *Ss_methylase* were examined from the genomes of 53 *Shigella* serotypes and EIEC. Serotypes displaying a genotype(s) that is an exception to the rule for *Shigella*/EIEC inclusion/exclusion, description of the exception, number of strains with the exception, accession numbers of the isolates, and total number of strains examined are listed.

[b] The full-length *cadA* in the *S. dysenteriae* 1 genome has a 4-bp deletion and the gene product is prematurely truncated.

[c] The full-length *cadA* in the *S. sonnei* genome is disrupted by two insertional elements and the gene product is prematurely truncated.

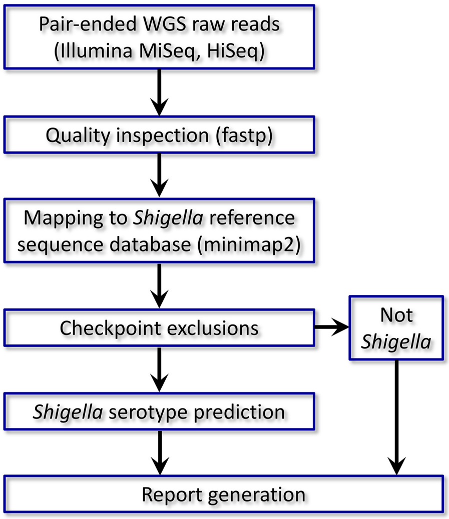

Pair-ended WGS raw reads
(Illumina MiSeq, HiSeq)
Quality inspection (fastp)
Mapping to *Shigella* reference sequence database (minimap2)
Checkpoint exclusions
Not *Shigella*
Shigella serotype prediction
Report generation

FIG 1 Summary of workflow for ShigaTyper. A detailed description can be found in Results ("Development of an automated *in silico Shigella* serotyping pipeline").

addition, sequences encoding the *Shigella* enterotoxins 1 and 2 (ShET1 and ShET2) and the N terminus of autotransporter toxin Sat (*sat*_N) were included (Table 2).

Comparison of results from conventional and molecular *Shigella* serotyping. We manually inspected WGS reads of 259 isolates in the development set to molecularly determine their species and serotype based on their O-antigen synthesis and modification genes, *Shigella*/EIEC differentiation markers, and MLST profile. Overall, a serotype can be assigned to 253 isolates (97.7%) based on their molecular profiles. Sequences of *wzx* gene for multiple serotypes were observed in 30 (11.6%) isolates. Nevertheless, reads mapped to minor *wzx* genes were usually <2% of that mapped to a dominant *wzx* gene, indicating low levels of contamination, and a serotype could be assigned in 25 cases. Five genomes had multiple *wzx* genes present at comparable levels and 1 had no recognizable serotype determinant genes. By molecular profiling, there were 83 *S. boydii* isolates, with 8, 5, 4, 6, 2, 1, 2, 4, 6, 7, 4, 1, 3, 3, 2, 2, 2, 5, 4, 9, and 3 isolates belonging to serotypes 1 to 20 and E1621-54, respectively (21 serotypes), 55 *S. dysenteriae* isolates, with 6, 10, 6, 2, 3, 2, 1, 5, 4, 2, 2, 3, 3, 3, 2, and 1 isolates belonging to serotypes 1 to 15 and 96-265, respectively (16 serotypes), 50 *S. flexneri* isolates, with 3, 1, 2, 3, 4, 11, 3, 7, 3, 1, 3, 1, 3, 6, 1 isolates belonging to serotypes Y, Xv (4c), 1a, 1b, 1c (7a), 2a, 2b, 3a, 3b, 4a, 4av, 4bv, 5a, 6, and 7b, respectively (15 serotypes), 20 *S. sonnei* isolates, and 14 EIEC isolates. Six isolates designated *Shigella* were *ipaH* negative and therefore considered non-*Shigella*/EIEC. No isolates carried *wzx* and *wzy* belonging to different O antigens as was reported for *E. coli* (68).

We compared *in silico* and conventional designations for these 259 strains (Table 1). All 24 non-*Shigella*/EIEC isolates were identified by the absence of *ipaH*_C. Twelve of the 13 EIEC isolates (92.3%) were identified as EIEC except 1 isolate that carried *S. dysenteriae* 3 *wzx* and *wzy* but lacked *EclacY* and *cadA*. As this strain was typed to ST270 by MLST, it is likely to be an EIEC isolate. For the 221 *Shigella* genomes, 191 (86.4%) were congruent at species level and 6, 1, and 17 were molecularly determined as not *Shigella*/EIEC, EIEC, and another *Shigella* species. Of the 185 isolates with a serotype designation, 149 (80.5%) had concordant serotype determinants.

Development of an automated *in silico Shigella* serotyping pipeline. Molecular serotyping of *Shigella* requires careful consideration of multiple gene determinants, which can be daunting for inexperienced analysts. To automate such a process in a rapid and efficient way, we developed ShigaTyper, an integrative workflow for *in silico Shigella* serotyping using Illumina paired-end WGS reads (Fig. 1). Jupyter Notebook was used as the user interface so that all command line tools could be prerecorded and

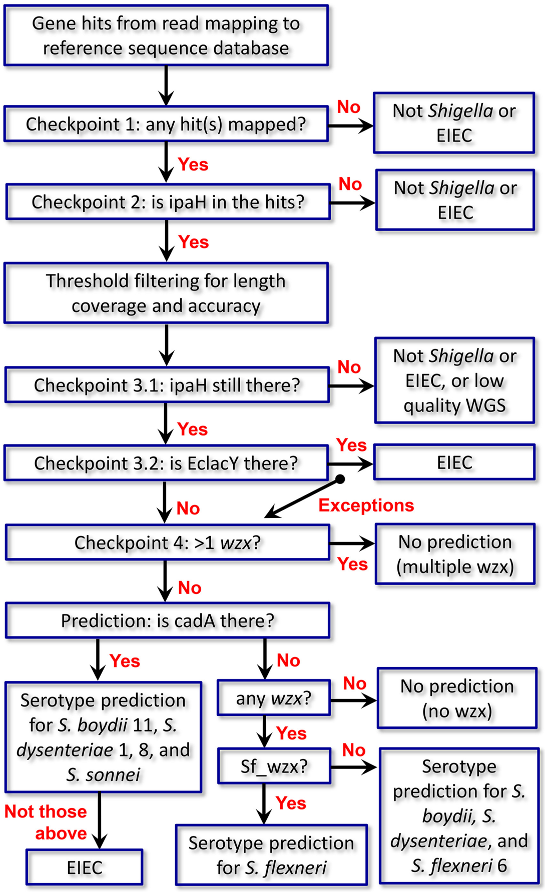
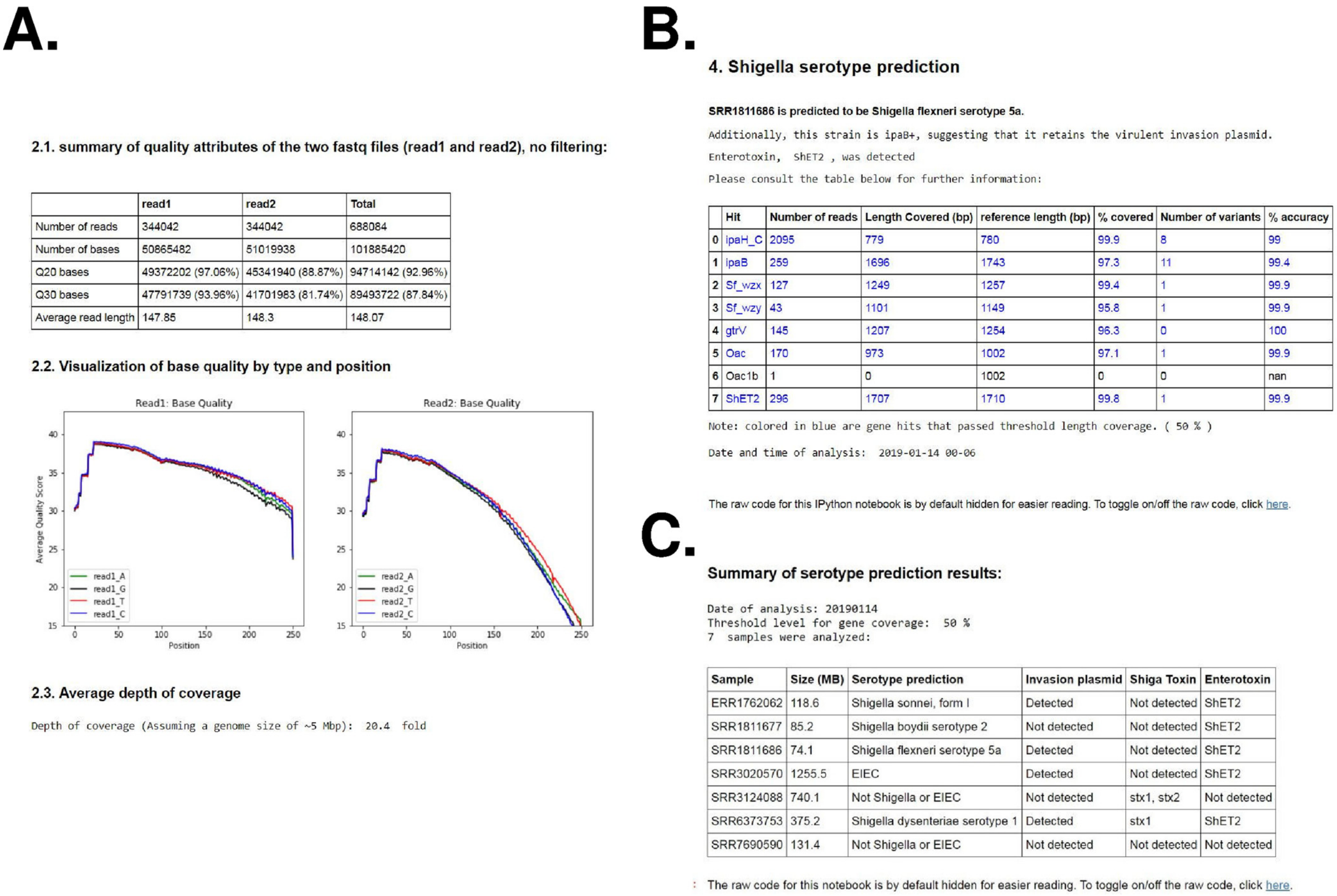

Gene hits from read mapping to reference sequence database
Checkpoint 1: any hit(s) mapped?
No
Not Shigella or EIEC
Yes
Checkpoint 2: is ipaH in the hits?
No
Not Shigella or EIEC
Yes
Threshold filtering for length coverage and accuracy
Checkpoint 3.1: ipaH still there?
No
Not Shigella or EIEC, or low quality WGS
Yes
Checkpoint 3.2: is EclacY there?
Yes
EIEC
No
Exceptions
Checkpoint 4: >1 wzx?
Yes
No prediction (multiple wzx)
No
Prediction: is cadA there?
Yes
No
Serotype prediction for S. boydii 11, S. dysenteriae 1, 8, and S. sonnei
any wzx?
No
No prediction (no wzx)
Yes
Sf_wzx?
No
Serotype prediction for S. boydii, S. dysenteriae, and S. flexneri 6
Yes
Not those above
EIEC
Serotype prediction for S. flexneri

FIG 2 Schematic illustration of a decision tree for *Shigella* differentiation before serotype prediction employed in ShigaTyper. ShigaTyper was designed to differentiate and exclude non-*Shigella* or contaminated isolates before predicting serotype for *Shigella* isolates. Distantly related non-*Shigella*/EIEC species (such as *Listeria*) usually have no read mapped to any of the genes in the reference sequence database and fail at checkpoint 1. Enterobacterial species (such as *Salmonella*) may have one or more hits but not *ipaH*_C and fail at checkpoint 2. Checkpoint 3 excludes EIEC based on the presence of full-length *EclacY* gene, with the exception of *S. boydii* 9 and 15. Last, if there are more than one *wzx* genes present in the WGS reads, it indicates multiple serotypes and fails checkpoint 4. Details on serotype prediction are provided in Results.

executed in one place with the click of a button and output directly printed below each step. A recently reported WGS reads preprocessing package, fastp, was used for quality inspection (69). Comparing results from 95 isolates with and without quality filtering and trimming showed 100% consistency in prediction outcomes. Therefore, quality filtering and trimming were omitted. WGS raw reads were directly aligned to the reference sequence database using minimap2 (70). *Shigella* differentiation was conducted through exclusion steps before serotype prediction (Fig. 2). Strains that did not carry *ipaH*_C were considered "not *Shigella*/EIEC" and eliminated. For *ipaH*$^+$ strains, length coverage and number of variants for each of the gene hits were determined using samtools (71) and bcftools. Threshold values were set to eliminate gene hits that did not achieve sufficient coverage and accuracy. Gene coverage and accuracy were defined as the fraction of gene length covered by WGS reads (breadth of coverage) and fraction of nucleotide identity to the reference sequence, respectively. We tested 80% and 50% for gene coverage and 80% for gene accuracy. The list of gene hits passing threshold filtering was screened for *Shigella*/EIEC differentiation markers. Strains deemed to belong to *Shigella* were then subjected to serotype prediction. In this method, a report is automatically generated for each sample, including name, quality attributes of the WGS reads, a serotype prediction, and a summary table for each of the



2.1. summary of quality attributes of the two fastq files (read1 and read2), no filtering:

	read1	read2	Total
Number of reads	344042	344042	688084
Number of bases	50865482	51019938	101885420
Q20 bases	49372202 (97.06%)	45341940 (88.87%)	94714142 (92.96%)
Q30 bases	47791739 (93.96%)	41701983 (81.74%)	89493722 (87.84%)
Average read length	147.85	148.3	148.07

2.2. Visualization of base quality by type and position

2.3. Average depth of coverage

```
Depth of coverage (Assuming a genome size of ~5 Mbp):  20.4  fold
```

4. Shigella serotype prediction

SRR1811686 is predicted to be Shigella flexneri serotype 5a.

```
Additionally, this strain is ipaB+, suggesting that it retains the virulent invasion plasmid.
Enterotoxin,  ShET2 , was detected
Please consult the table below for further information:
```

	Hit	Number of reads	Length Covered (bp)	reference length (bp)	% covered	Number of variants	% accuracy
0	ipaH_C	2095	779	780	99.9	8	99
1	ipaB	259	1696	1743	97.3	11	99.4
2	Sf_wzx	127	1249	1257	99.4	1	99.9
3	Sf_wzy	43	1101	1149	95.8	1	99.9
4	gtrV	145	1207	1254	96.3	0	100
5	Oac	170	973	1002	97.1	1	99.9
6	Oac1b	1	0	1002	0	0	nan
7	ShET2	296	1707	1710	99.8	1	99.9

```
Note: colored in blue are gene hits that passed threshold length coverage. ( 50 % )

Date and time of analysis:  2019-01-14 00-06
```

The raw code for this IPython notebook is by default hidden for easier reading. To toggle on/off the raw code, click here.

Summary of serotype prediction results:

```
Date of analysis: 20190114
Threshold level for gene coverage:  50 %
7  samples were analyzed:
```

Sample	Size (MB)	Serotype prediction	Invasion plasmid	Shiga Toxin	Enterotoxin
ERR1762062	118.6	Shigella sonnei, form I	Detected	Not detected	ShET2
SRR1811677	85.2	Shigella boydii serotype 2	Not detected	Not detected	ShET2
SRR1811686	74.1	Shigella flexneri serotype 5a	Detected	Not detected	ShET2
SRR3020570	1255.5	EIEC	Detected	Not detected	ShET2
SRR3124088	740.1	Not Shigella or EIEC	Not detected	stx1, stx2	Not detected
SRR6373753	375.2	Shigella dysenteriae serotype 1	Detected	stx1	ShET2
SRR7690590	131.4	Not Shigella or EIEC	Not detected	Not detected	Not detected

: The raw code for this notebook is by default hidden for easier reading. To toggle on/off the raw code, click here.

FIG 3 A representative output for ShigaTyper. (A) QC inspection of WGS reads. Quality inspection results were parsed from reports generated by fastp and are summarized in a table showing number of reads, number of bases, number of bases with >Q20 and >Q30 scores, and average read length. A visual representation of average quality score of each of the 4 bases over sequencing cycle and an estimated average depth for genome coverage are given below the table. (B) Serotype prediction for the sample. A direct serotype prediction is made by ShigaTyper based on threshold filter values passed by gene determinants as described in Results. A warning signal is given if sequence of the pINV-encoded virulence factor IpaB, a Shiga toxin, or an enterotoxin is detected in the WGS reads. The table summarizes characteristics of each of the genetic determinants identified from the WGS data. Those that passed the threshold filter values are shown in blue. All the codes are hidden from view for clarity of reporting but can be toggled to show for examination if needed. (C) Report of ShigaTyper batch processing. The summary table lists outcomes for serotype prediction, invasion plasmid, Shiga toxin, and enterotoxin.

identified gene hits for final review. In addition, when the virulence plasmid pINV or a toxin is detected, a warning message is also included. An example of report is shown in Fig. 3A and B. An additional summary table listing serotype prediction for each of the strains is listed in the batch processing notebook (Fig. 3C).

Serotyping prediction by ShigaTyper. Serotype prediction for ShigaTyper was made primarily through the serotype-specific *wzx* gene, as O-antigen expression is absolutely dependent on *wzx* but not *wzy* (10, 46, 72). Additionally, we observed better gene coverage for *wzx* than *wzy* (93.4% ± 8.9% versus 81.7% ± 19.1% for the 46 in-house samples under 1 GB), presumably because of the higher GC content of *wzx* than *wzy* (30.2% ± 1.95% versus 28.6% ± 1.52%), as the transposase-based library generation in the current MiSeq workflow disfavors AT-rich sequences (73, 74). For serotypes that cannot be predicted solely by *wzx*, additional criteria were applied as follows.

S. boydii 1 (Sb1) and 20 contain identical chromosomal *rfb* genes. For strains carrying *Sb1_wzx*, those that also carried a heparinase were assigned to *S. boydii* 20, while those that did not were designated *S. boydii* 1.

S. boydii 6 and 10 contain identical *rfb* genes; however, *wbaM* in *S. boydii* 6 is disrupted with an insertional element between positions 252 and 253 (48). Therefore, read alignment to wild-type *wbaM* is expected to be poor at the insertional junction for

S. boydii 6, and quality filtering should remove a significant fraction of these bases. Indeed, for the *S. boydii* 6 strain we examined, only 18.6% of bases passed quality filtering at the junction, while 57.4% of bases did for the overall *wbaM* gene. In contrast, in *S. boydii* 10 ($n = 3$), the percentage of bases passing quality filtering at the junction was comparable to that for the overall *wbaM* gene (50.0% ± 0.0% versus 51.7% ± 4.3%). A threshold ratio of percent passing filtering at the junction over the entire *wbaM* was therefore set at 0.5. An isolate with a ratio below the threshold was considered *S. boydii* 6, and an isolate with a ratio above 0.5 was considered *S. boydii* 10. We identified another 4 *S. boydii* isolates that were *wbaM*$^{+}$. Sequence alignment showed a contiguous, undisrupted *wbaM* gene consistent with *S. boydii* 10 for all 4 strains. All of them had a junction-to-overall ratio above 0.5 (0.955 ± 0.087), validating the use of *wbaM* junctional quality to distinguish *S. boydii* 6 and 10. There was only 1 *S. boydii* 6 isolate in our development set (even though 5 were designated *S. boydii* 6). However, this strategy later successfully distinguished *S. boydii* 6 from *S. boydii* 10 in our validation study.

All *S. flexneri* serotypes except *S. flexneri* 6 share the same *rfb* gene cluster but have different O-antigen modifications by enzymes encoded on bacteriophages or plasmids (12, 75). For strains containing the *S. flexneri* 1 to 5 *wzx* gene (*Sf_wzx*), the list of O-antigen modification genes identified was searched in a Python dictionary containing gene formulas of 19 *S. flexneri* serotypes. When a serotype had two or more possible gene formulas, all were included for interpretation. (For example, *S. flexneri* 5a is defined as *S. flexneri* modified by the glucosyltransferase, GtrV, regardless of the presence of the O-antigen acetylase, Oac. Both gene formulas "*gtrV*" and "*gtrV*, *oac*" were considered *S. flexneri* 5a.)

S. sonnei carries its *rfb* on pINV, which is lost at high frequency (76). Therefore, we used the chromosomal *Ss_methylase* as a diagnostic marker. To distinguish *S. sonnei* from *S. dysenteriae* 10 and EIEC, only strains positive for both *cadA* and *Ss_methylase* but negative for any *wzx* except *S. sonnei wzx* (*Ss_wzx*) were considered *S. sonnei*. A strain carrying *Ss_wzx* or pINV marker *ipaB* was assigned virulent *S. sonnei* form I. Otherwise, such a strain was considered form II. Only 5 (25%) *S. sonnei* strains were form I, consistent with the reported high plasmid instability.

Performance of ShigaTyper. We determined the prediction accuracy for ShigaTyper, excluding the 6 uninterpretable samples from the development set. When we used 80% as the threshold level for gene length coverage and accuracy, as previously reported for *S. flexneri* (77), we achieved 95.7% (242/253) and 94.5% (239/253) accuracies at the species and serotype levels. Sequence accuracy was >97% for all serotype determinants identified. In 9 out of the 11 isolates with inaccurate species designation, it could be attributed to low length coverage of one or more gene hits, leading to no prediction ($n = 7$) or misdesignation ($n = 2$). Prediction accuracies increased to 98.8% (250/253) and 98.0% (248/253) at the species and serotype levels, respectively, when the threshold gene coverage was reduced to 50%. Isolates that could not be serotyped at 50% gene coverage by ShigaTyper were manually examined. Two isolates had low-level contamination of the genetic determinant(s) from another serotype or EIEC that were >50% covered. One isolate did not have a *wzx* gene and therefore could not be typed. One isolate was predicted to be *S. flexneri* 5b because it carried O-antigen modification genes *gtrV*, *gtrX*, and *oac*. However, this strain was phenotypically *S. flexneri* 3a due to a 1-bp insertion in *gtrV*. Another isolate was a *S. flexneri* carrying unconventional gene formula not included in the prediction script and was designated "*S. flexneri* novel serotype."

The turnaround time for ShigaTyper was directly proportional to the size of the paired end fastq files irrespective of the prediction outcome (Fig. 4A). On average, the pipeline processed WGS raw reads at 538.1 MB/min, translating to a time to prediction of ~1 min for an average-size sample (509.9 ± 538.1 MB). Most of the time was spent on executing the three command line tools, fastp, minimap2 and samtools, accounting for 35.8% ± 4.9% and 45.7% ± 4.8% of the total time. As a result, 36.6% ± 4.9% and

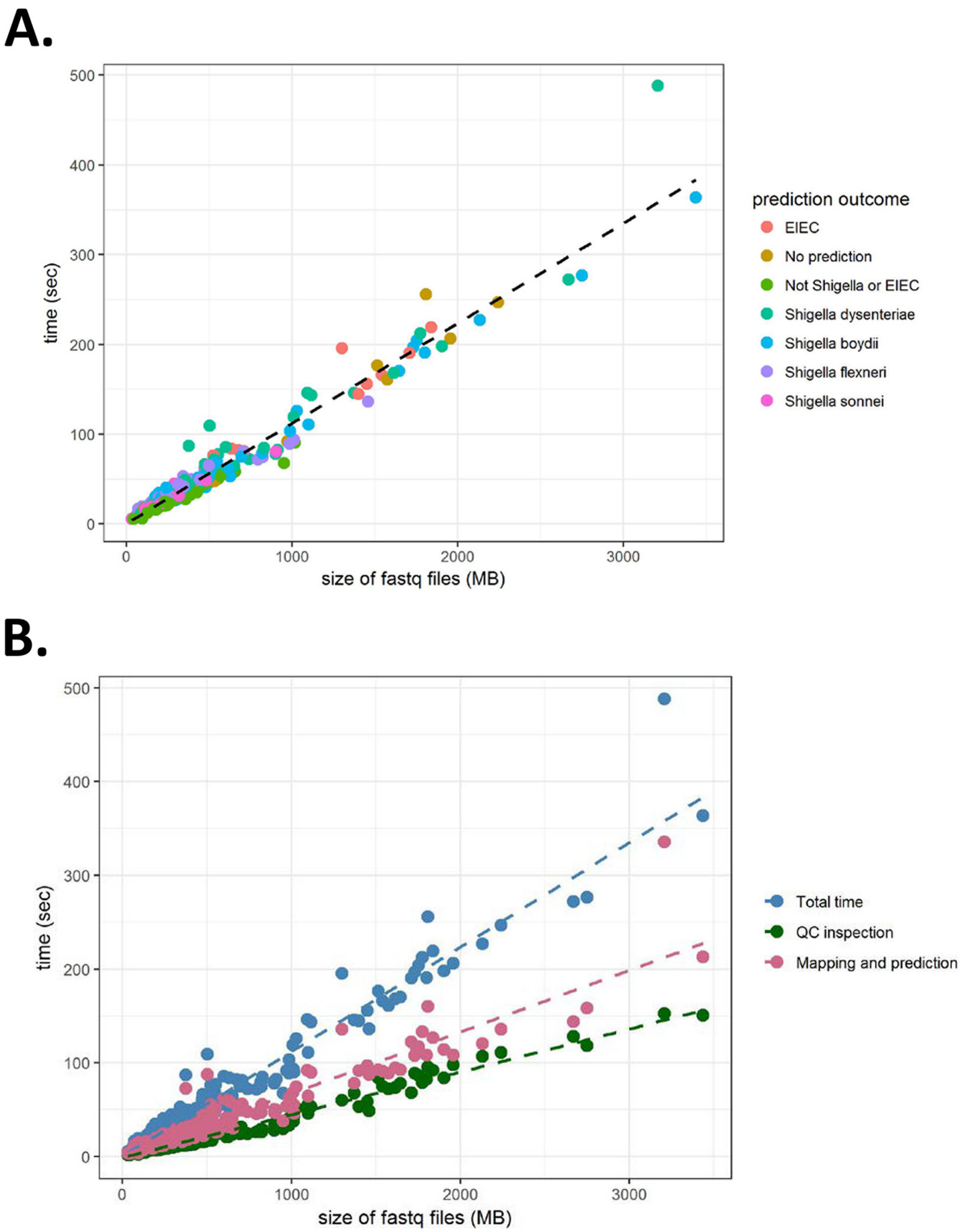


FIG 4 Speed for serotype prediction is directly proportional to the size of WGS files. (A) Total time spent for *Shigella* serotyping was plotted against the sum of size of the paired-end WGS reads in fastq.gz format. Outcomes of serotype prediction are indicated on the right. A linear regression line is shown in black. (B) Total time, time spent on quality (QC) inspection, and time spent on mapping and prediction are plotted against the sum of size of the paired-end WGS reads in fastq.gz format. Linear regression lines of the same color are also shown. The average size for the sum of the paired-end WGS reads was 509.9 ± 538.1 MB and ranged from 30.7 to 3,436.7 MB.

63.4% ± 4.9% of the total time were spent on quality inspection and mapping and prediction, respectively (Fig. 4B).

Validation of ShigaTyper. ShigaTyper was validated using a separate collection of *Shigella* strains, including 62 well-designated clinical isolates (16), 33 reference strains, and WGS reads of 287 isolates downloaded from the NCBI, comprising 49 designated serotypes and 32 non-*Shigella* isolates. Specifically, the validation set included 94 isolates designated *S. sonnei* and 87 isolates designated *S. boydii*, of which 10, 8, 2, 9,

TABLE 4 Summary of *Shigella* WGS validation sets used in this study[a]

			Prediction accuracy, no. of strains (%)			
			80% length coverage		50% length coverage	
Strain designation (molecular)	No. of strains	No. of serotypes included	Species	Serotype	Species	Serotype
S. boydii	90	19	79 (87.7)	76 (84.4)	87 (96.7)	87 (96.7)
S. dysenteriae	74	13	72 (97.3)	72 (97.3)	74 (100)	74 (100)
S. flexneri	87	15	87 (100)	82 (88.9)	87 (100)	84 (96.6)
S. sonnei	93	1	92 (98.9)	92 (98.9)	92 (98.9)	92 (98.9)
Shigella subtotal	344	49	330 (95.9)	322 (93.6)	340 (98.8)	337 (98.0)
EIEC	14		14 (100)		14 (100)	
Non-*Shigella*/EIEC	22		22 (100)		22 (100)	
Overall	380			366 (96.3)		373 (98.2)

[a] Strains were sequenced from an in-house collection of 62 clinical isolates and 33 reference strains ($n = 95$), or their WGSs were downloaded from the NCBI ($n = 255$).

2, 1, 1, 5, 0, 3, 2, 4, 3, 6, 3, 0, 0, 6, 4, 11, 3, 2, and 2 isolates typed to serotypes 1 to 20, E1621-54, E140634-99, and E25411-82 (20 serotypes), respectively. Eighty-four isolates designated *S. dysenteriae* included 3, 9, 8, 7, 2, 1, 1, 3, 2, 5, 1, 1, and 1 isolate typed to serotypes 1 to 4, 6 to 9, 11, 12, 14, E11207-96 (96-265), and E670-74 (13 serotypes), respectively, and 40 untyped isolates. Eighty-five isolates designated *S. flexneri* included 1, 1, 2, 9, 9, 17, 3, 7, 3, 2, 1, 1, 2, 7, and 1 isolate typed to serotypes Y, Xv (4c), 1a, 1b, 1c (7a), 2a, 2b, 3a, 3b, 4a, 4b, 4, 5a, 5b, 6, and 7 (15 serotypes), respectively, and 19 untyped isolates. Thirty-two isolates that were common diarrheagenic or foodborne bacteria were selected as the exclusion group, including 26 *E. coli* isolates (of which 10, 3, and 1 were designated EIEC, ETEC, and STEC), 1 *P. shigelloides* isolate, 1 *Salmonella enterica* isolate, 1 *K. pneumoniae* isolate, 2 *S. aureus* isolates, and 1 *L. monocytogenes* isolate. After correction by manual inspection, there were 90 *S. boydii* isolates, with 9, 9, 3, 11, 2, 1, 1, 4, 0, 4, 2, 2, 2, 7, 3, 1, 0, 5, 6, 13, and 5 isolates belonging to serotypes 1 to 20 and E1621-54, respectively, 74 *S. dysenteriae* isolates, with 5, 19, 12, 11, 0, 2, 1, 1, 7, 0, 2, 8, 1, 1, 0, 3, and 1 isolate belonging to serotypes 1 to 15, 96-265, and E670/74, respectively, 87 *S. flexneri* isolates, with 1, 2, 1, 17, 12, 27, 1, 3, 8, 4, 2, 1, 1, 6, and 1 isolate belonging to serotypes Y, Xv (4c), 1a, 1b, 1c (7a), 2a, 2av, 2b, 3a, 3b, 4av, 5a, 5b, 6, and 7b, respectively, 93 *S. sonnei* isolates, and 14 EIEC isolates. Two isolates designated *Shigella* were *ipaH* negative and therefore considered non-*Shigella*/EIEC. One isolate had gene determinants from multiple serotypes (*wzx* or *wzy*) and another had no gene determinants, and the two were deemed uninterpretable. This corresponded to 95.7% concordance rate at the species level and 90% at the serotype level. After correction and removal of the 2 uninterpretable isolates, the validation set contained 344 *Shigella* isolates of 49 serotypes, 14 EIEC isolates, and 22 non-*Shigella*/EIEC isolates (Table 4).

The validation set was subjected to automated serotype prediction by ShigaTyper. At the 80% gene coverage threshold, the accuracy rates for the 344 *Shigella* strains were 95.9% and 93.6% at species and serotype levels, respectively. At the 50% gene coverage level, the accuracy rates increased to 98.8% and 98.0% at the species and serotype levels, respectively. At both threshold levels ShigaTyper differentiated 14 out of 14 EIEC isolates and 22 out of 22 non-*Shigella*/EIEC isolates (100%). The overall accuracies for the 380 isolates were 96.3% and 98.2% at 80% and 50% gene coverage thresholds, respectively. Of the 7 *Shigella* isolates that could not be identified at the 50% gene coverage level, 4 had low-level contamination of another serotype, 1 was an *S. sonnei* isolate without a full-length *cadA*, 1 was an *S. flexneri* 3b isolate carrying an unconventional O-antigen acetylase gene (*oac1b* instead of *oac*), and 1 was an *S. flexneri* 4av carrying *Ss_methylase*, rendering the algorithm unable to make a correct prediction. The presence of *Ss_methylase* was not due to contamination with *S. sonnei*, *S. dysenteriae* 10, or EIEC, because no read was mapped to *EclacY*, *cadA*, *Sd10_wzx*, or *Sd10_wzy*.

Four of the 10 clinical *S. sonnei* isolates (40%) were completely devoid of *EclacY* sequence, while only 6 (5.8%) out of the remaining 103 *S. sonnei* isolates from the developed world did not carry an *EclacY* remnant. This allelic polymorphism did not affect the serotype prediction outcome.

The ability of *Shigella* to absorb Congo red to form red colonies is linked to its virulence plasmid pINV (56). We examined the predictive value of *ipaB* and ShET2 as a pINV-indicative marker in 83 in-house *Shigella* isolates. Sixty-four of the 83 isolates formed red or dark pink colonies (pINV$^+$) in the presence of Congo red, and 59 and 61 of them were positive for *ipaB* and ShET2, respectively. Sixty-one of them were positive for either *ipaB* or ShET2. Of the 19 isolates that formed light pink or white colonies (pINV$^-$), 8 and 10 were positive for *ipaB* and ShET2, respectively. Twelve of them were positive for either *ipaB* or ShET2. Both *ipaB* and ShET2 had an overall accuracy of 84.3%. However, *ipaB* has a slightly higher positive predictive value (88.1%) than ShET2 (85.9%) or *ipaB* or ShET2 (83.6%) in predicting pINV.

A subset of 68 *S. sonnei* isolates in the validation set were characterized for the presence of Shiga toxin-producing bacteriophage (78). ShigaTyper detected *stx*$_1$ in 42 out of the 42 Shiga toxin-positive isolates and did not detect *stx*$_1$ in 26 out of the 26 Shiga toxin-negative isolates, 100% consistent with the previous report.

Genetic variation due to bacteriophages. *Ss_methylase* was observed in the genomes of all *S. sonnei* (*n* = 115) and *S. dysenteriae* 10 (*n* = 3) isolates, 4 EIEC isolates (*n* = 28), and 1 *S. flexneri* 4av isolate (*n* = 5). Therefore, we investigated if this gene is associated with mobile genetic elements. *Ss_methylase* in *S. sonnei* was present within an ~9.7-kbp contig flanked by insertion sequence 1 (IS1) sequence between the *ynfF* and *ydf* loci, suggesting a transposon-mediated insertion event in the founding ancestor. In contrast, *Ss_methylase* in *S. dysenteriae* 10, 2 EIEC, and the *S. flexneri* 4av resided in an ~40-kbp lambdoid prophage integrated between the *potB* and *potC* loci. The prophage from another EIEC isolate (SRA accession number SRR6049563) was integrated between the *mtfA and zinT* loci. Prophages from the EIEC and *S. flexneri* 4av isolates shared gene organizations similar to the one from the *S. dysenteriae* 10 genome but were nonconserved in genes encoding structural phage proteins.

Shiga toxin has never been reported for *S. boydii*. We observed *stx*$_1$ in 3 *S. boydii* 19 (*n* = 10). Sequence comparison of the assembled genomes revealed that the *stx*$_1$ coding sequence resided within a prophage integrated between the *ynfG* and *ynfF* loci that is 99.9% identical to the POCJ13 phage, a lysogenic bacteriophage that infects and converts *S. dysenteriae* 4 and *S. flexneri* into Shiga toxin producers (58).

***Shigella* identification through MLST and biochemical analyses.** All isolates from the development and validation sets were screened for their MLST profiles (*n* = 637). Most *Shigella* serotypes belonged to STs previously reported (28), except that ST1753 was previously assigned to *S. flexneri*, while we observed that only *S. boydii* E1621-54 typed to this ST (*n* = 7). ST groups for some *Shigella* serotypes were not known. Isolates from some serotypes returned no or previously unreported STs. Overall, 78 of the 552 *Shigella* isolates (14.1%) could not be properly categorized by MLST (Table S4).

Fifty-one *Shigella* isolates from 42 different serotypes were selected for automated microbial identification through biochemical properties. Except *S. sonnei* isolates that could be identified to species level, most *Shigella* isolates were identified as "*Shigella* group" (non-*S. sonnei Shigella*). Five (9.8%) isolates from 3 serotypes were identified as *E. coli* and 2 isolates from 2 serotypes showed low confidence in discrimination between *Shigella* and *E. coli*. A control isolated identified as EIEC by molecular profiling was subjected to the same analysis and was identified as *E. coli* (Table S5).

DISCUSSION

Shigella is a serious threat to public health, despite the low number of cases in developed countries. However, it can be expensive to maintain a pathogen-specific surveillance program, factoring in the time required for analyst training, reagent preparation, and maintenance, especially for rare *Shigella* serotypes. Conventional *Shigella* identification method is labor-intensive, potentially subjective, and not sufficiently accurate. Using molecular profiling, we showed that conventional *Shigella* serotyping was at best 90% accurate, consistent with a previous report of an upper limit at 91% (17). Similarly, biochemical identification could erroneously assign some *Shigella* isolates as *E. coli*. A WGS-based identification method has a universal workflow for all pathogens and provides high-resolution data with better accuracy. Therefore, it is ideal to replace the conventional methods. The same sequencing data can be used in screenings for virulence genes and antimicrobial resistance (AMR), MLST, and single nucleotide polymorphism (SNP) analysis, further reducing the cost in pathogen characterization and outbreak investigation.

In this study, we conducted a comprehensive examination of genome information from 56 *Shigella* serotypes. By identifying gene determinants for novel *Shigella* serotypes, as well as setting criteria for *E. coli* and EIEC differentiation, we provided information enabling *in silico Shigella* serotyping. We further demonstrated the feasibility of this approach with a proof-of-concept WGS serotyping pipeline, ShigaTyper, using bioinformatic and programming tools freely available online. ShigaTyper provides a prototype for simple and rapid identification of clinical *Shigella* isolates with high accuracy.

The lack of lactose fermentation and the lack of lysine decarboxylation represent two hallmark traits of *Shigella*. Nevertheless, we observed considerable variation in the gene structure of the *EclacY* and *cadA*, confirming the previous observation that *Shigella* is not a homogeneous group and the seemingly identical phenotype was often caused by different inactivation mechanisms (67, 79). Variability was even present within serotype (Table 3), suggesting further genome rearrangement post-speciation. This high genome variability, together with the observation that the putative *S. sonnei*-specific marker, *Ss_methylase* (2, 55), was present in multiple *Shigella* and EIEC serotypes associated with bacteriophages, suggests that no single genetic marker alone should be used for *Shigella* identification at the species or serotype level. Rather, a combination of genes should be taken into consideration for proper EIEC differentiation and *Shigella* designation. WGS is an especially valuable tool for this purpose, as it provides abundant information and the data can always be reevaluated with additional gene makers. For example, serotype determinants of EIEC can be included for better differentiation.

Shigella is a highly dynamic group of bacteria. Annually, 6 to 10% of *Shigella* isolates are untypeable (8), suggesting that novel serotypes are constantly emerging and *Shigella* evolution is an ongoing process. We identified serotype determinants from six previously uncharacterized *Shigella* serotypes. Four of them had nearly identical O-antigen genes of another enterobacterium and likely arose from horizontal transfer. The absence of *rfb* in some serotypes (*S. dysenteriae* 93-119 and 204-96) and the presence of additional serotype determinants on mobile elements (*S. boydii* 20) indicate that there are multiple mechanisms at work for *Shigella* serotype diversification.

Direct target mapping using WGS reads has been successfully employed for predicting bacterial serotypes for *E. coli*, *S. flexneri*, and *Salmonella* (68, 77, 80) and for inferring AMR (81, 82). We developed a similar pipeline for *Shigella* using a short-read mapping approach that has been used for microbial MLST (83) and *Salmonella* serotyping (80). The assembly-free approach reduced analysis time and is less computation intensive, enabling resource-limited field labs to perform *in silico* serotyping on a regular office laptop. The average fastq file for validation was 446.7 ± 296.7 MB, and took 49.8 ± 33.1 s to prediction, or 31.6 ± 21.0 s had quality inspection been omitted. We attribute the higher speed than SeqSero for *Salmonella* (80) to the fact that *Shigella* serotype determinants are unique enough and did not require subsequent rounds of alignment/BLAST to discriminate multiple probable alleles. Because the sequence aligner in ShigaTyper, minimap2, is capable of aligning long DNA sequences efficiently (70), assembled genomes in fasta format can be used for serotype prediction in a similar fashion. We included codes processing genome assemblies in ShigaTyper. However, it might not be as time-effective, as genome assembly usually takes more than 1 min to complete.

ShigaTyper is particularly suitable for general microbiologists. The use of Jupyter Notebook consolidated all codes in one place. Once installed, there will be no need for bioinformatic and programming skills. A direct serotype prediction is made without operator interpretation, reducing user subjectivity and ensuring reproducibility. This pipeline is highly flexible. For example, by including sequences of Shiga toxins in the reference sequence database, we identified *S. boydii* 19 as another Shiga toxin-producing serotype. Detection of novel serotypes or additional virulence genes can be easily achieved by updating the reference sequence database. We determined the current threshold setting optimal at 50% gene length coverage, as this setting captured most of the serotype gene determinants but allowed tolerance for low-level contamination. Additional filter settings can be adopted to meet regulatory requirement as necessary.

The scope of our work was limited by the availability of well-designated *Shigella* WGS raw reads. Of the 59 serotypes that ShigaTyper was designed to identify, we were unable to obtain WGS reads of *S. flexneri* Yv, X, and 4b for examination. As our cohort was small and some serotypes were represented by only a few strains, larger-scale confirmation and validation are needed for the implementation of *in silico Shigella* serotyping. Nevertheless, our work contributed to the transition of public health surveillance into molecular technologies and can be integrated with other WGS-based tools for detection and investigation of enteric pathogens.

MATERIALS AND METHODS

Strains. In-house strains used in this study are described in Table S1. Most strains used for ShigaTyper development were provided by the FDA Pacific Regional Laboratory Southwest. *Shigella* strains used for validation were generously provided by the Global Enteric Multicenter Study consortium (16) and California Department of Public Health. Strains were propagated in brain heart infusion or on tryptic soy agar (TSA) with 0.1 mg/ml of Congo red. All strains were screened by PCR for the presence of *ipaH* and *Ss_methylase*. Selected isolates were examined using serotype-specific PCR primers. Strains corresponding to the sequences downloaded from the National Center for Biotechnology Institute (NCBI) are described in Tables S2 (genomes or sequence assemblies) and 3 (WGS reads).

Whole-genome sequencing. Genomic DNA was extracted from 1 to 2 ml of overnight culture using QIAamp DNA minikit on a QiaCube (Qiagen, Hilden, Germany) and fragmented and indexed using Nextera XT DNA sample preparation and DNA index kits (Illumina, San Diego, CA). DNA concentration was determined using a Qubit dsDNA BR Assay system (Thermo Fisher, Waltham, MA). Libraries were normalized and pooled for sequencing on an Illumina MiSeq system using 500 V.2 reagent cartridges.

Bioinformatic analyses. Local computational analyses were conducted on a Dell laptop (Intel core i7-6600U CPU, 16 GB of memory) with a Windows 7 host and an Ubuntu 18.04 guest addition (4 processor cores, 4.3 GB of memory) on a VMware Player 14.1.1. Bioinformatic packages were installed and managed by Anaconda 4.4.11 with Python 3.6.5 through Bioconda, including fastp 0.12.2 (69), minimap2 2.13 (70), and htslib/samtools/bcftools 1.7 (71, 84). All command line and python codes were maintained in Jupyter Notebooks and run on Jupyter 1.0.0 and nbconvert 5.3.1. Papermill 0.14.2 was used for batch processing of samples. MLST of scheme "ecoli1" was determined using stringMLST 0.5.1 (85) with 12 GB memory allocation. When needed, *de novo* genome assembly, gene annotation, and *E. coli* serotyping were performed using Spades 3.11.1 (86) on GalaxyTrakr, RAST (87–89), and SerotypeFinder (68). Mauve (90) 2015-02-26 was used for genome comparison.

Biochemical identification. A Vitek 2 Compact automated system with GN ID card (bioMérieux, Marcy-l'Étoile, France) was used for microbial identification per manufacturer's instruction.

Data availability. Sequences generated in this study have been deposited in the NCBI Sequence Read Archive under the BioProject number PRJNA490540; accession numbers for each of the strains are listed in Table S1. The stand-alone *Shigella* serotyping pipeline, together with instructions for system setup and running, is available by request or at https://github.com/CFSAN-Biostatistics/shigatyper. An online version will be made available soon on GalaxyTrakr (https://galaxytrakr.org).

SUPPLEMENTAL MATERIAL

Supplemental material for this article may be found at https://doi.org/10.1128/AEM.00165-19.

SUPPLEMENTAL FILE 1, PDF file, 0.2 MB.

SUPPLEMENTAL FILE 2, XLSX file, 0.02 MB.

SUPPLEMENTAL FILE 3, XLSX file, 0.01 MB.

SUPPLEMENTAL FILE 4, XLSX file, 0.04 MB.

ACKNOWLEDGMENTS

We thank Sharon Tennant of University of Maryland School of Medicine, Baltimore, MD, Edward Desmond and Gregory B. Inami of California Department of Public Health, Kai-Shun Chen of FDA Winchester Engineering and Analytical Center, Wen Lin of FDA

Center for Tobacco Products, and Chong-Ming Cheng, a former member of FDA Pacific Southwest Laboratory, for providing *Shigella* strains. We are grateful to Rachel Binet of FDA CFSAN, Selen Stromgren of FDA Office of Regulatory Affairs, and Christopher J. Webb of Dupont for critical reading of the manuscript. We thank Okumu Kaluoch, Elaine Yeh, and Jonathan Yenovkian of FDA San Francisco Laboratory for technical assistance and discussion.

This work was supported by FDA Commissioner's Fellowship to Y.W. and FDA CARTS project number IR01650 to H.K.L.

The contents of this article do not reflect any position of the U.S. government or the FDA. Reference to any commercial materials, equipment, or process does not in any way constitute approval, endorsement, or recommendation by the FDA.

REFERENCES

1. Pires SM, Fischer-Walker CL, Lanata CF, Devleesschauwer B, Hall AJ, Kirk MD, Duarte AS, Black RE, Angulo FJ. 2015. Aetiology-specific estimates of the global and regional incidence and mortality of diarrhoeal diseases commonly transmitted through food. PLoS One 10:e0142927. https://doi.org/10.1371/journal.pone.0142927.
2. Liu J, Platts-Mills JA, Juma J, Kabir F, Nkeze J, Okoi C, Operario DJ, Uddin J, Ahmed S, Alonso PL, Antonio M, Becker SM, Blackwelder WC, Breiman RF, Faruque AS, Fields B, Gratz J, Haque R, Hossain A, Hossain MJ, Jarju S, Qamar F, Iqbal NT, Kwambana B, Mandomando I, McMurry TL, Ochieng C, Ochieng JB, Ochieng M, Onyango C, Panchalingam S, Kalam A, Aziz F, Qureshi S, Ramamurthy T, Roberts JH, Saha D, Sow SO, Stroup SE, Sur D, Tamboura B, Taniuchi M, Tennant SM, Toema D, Wu Y, Zaidi A, Nataro JP, Kotloff KL, Levine MM, Houpt ER. 2016. Use of quantitative molecular diagnostic methods to identify causes of diarrhoea in children: a reanalysis of the GEMS case-control study. Lancet 388: 1291–1301. https://doi.org/10.1016/S0140-6736(16)31529-X.
3. GBD Diarrhoeal Diseases Collaborators. 2017. Estimates of global, regional, and national morbidity, mortality, and aetiologies of diarrhoeal diseases: a systematic analysis for the Global Burden of Disease Study 2015. Lancet Infect Dis 17:909–948. https://doi.org/10.1016/S1473-3099(17)30276-1.
4. DuPont HL, Levine MM, Hornick RB, Formal SB. 1989. Inoculum size in shigellosis and implications for expected mode of transmission. J Infect Dis 159:1126–1128. https://doi.org/10.1093/infdis/159.6.1126.
5. Kotloff KL, Riddle MS, Platts-Mills JA, Pavlinac P, Zaidi AKM. 2017. Shigellosis. Lancet 391:801–812. https://doi.org/10.1016/S0140-6736(17)33296-8.
6. Scallan E, Hoekstra RM, Angulo FJ, Tauxe RV, Widdowson MA, Roy SL, Jones JL, Griffin PM. 2011. Foodborne illness acquired in the United States—major pathogens. Emerg Infect Dis 17:7–15. https://doi.org/10.3201/eid1701.091101p1.
7. Scharff RL. 2012. Economic burden from health losses due to foodborne illness in the United States. J Food Prot 75:123–131. https://doi.org/10.4315/0362-028X.JFP-11-058.
8. CDC. 2018. National enteric disease surveillance: Shigella annual report, 2016. https://www.cdc.gov/nationalsurveillance/pdfs/LEDS-Shig-2016-REPORT-508.pdf.
9. CDC. 2014. 2014 annual human isolates report. https://wwwcdcgov/narms/pdf/2014-Annual-Report-narms-508cpdf.
10. Liu B, Knirel YA, Feng L, Perepelov AV, Senchenkova SN, Wang Q, Reeves PR, Wang L. 2008. Structure and genetics of Shigella O antigens. FEMS Microbiol Rev 32:627–653. https://doi.org/10.1111/j.1574-6976.2008.00114.x.
11. Lai V, Wang L, Reeves PR. 1998. Escherichia coli clone Sonnei (Shigella sonnei) had a chromosomal O-antigen gene cluster prior to gaining its current plasmid-borne O-antigen genes. J Bacteriol 180:2983–2986.
12. Allison GE, Verma NK. 2000. Serotype-converting bacteriophages and O-antigen modification in Shigella flexneri. Trends Microbiol 8:17–23. https://doi.org/10.1016/S0966-842X(99)01646-7.
13. Shepherd JG, Wang L, Reeves PR. 2000. Comparison of O-antigen gene clusters of Escherichia coli (Shigella) sonnei and Plesiomonas shigelloides O17: sonnei gained its current plasmid-borne O-antigen genes from P. shigelloides in a recent event. Infect Immun 68:6056–6061. https://doi.org/10.1128/IAI.68.10.6056-6061.2000.
14. Feng L, Perepelov AV, Zhao G, Shevelev SD, Wang Q, Senchenkova SN, Shashkov AS, Geng Y, Reeves PR, Knirel YA, Wang L. 2007. Structural and genetic evidence that the Escherichia coli O148 O antigen is the precursor of the Shigella dysenteriae type 1 O antigen and identification of a glucosyltransferase gene. Microbiology 153:139–147. https://doi.org/10.1099/mic.0.2006/001107-0.
15. Sun Q, Knirel YA, Lan R, Wang J, Senchenkova SN, Jin D, Shashkov AS, Xia S, Perepelov AV, Chen Q, Wang Y, Wang H, Xu J. 2012. A novel plasmid-encoded serotype conversion mechanism through addition of phosphoethanolamine to the O-antigen of Shigella flexneri. PLoS One 7:e46095. https://doi.org/10.1371/journal.pone.0046095.
16. Livio S, Strockbine NA, Panchalingam S, Tennant SM, Barry EM, Marohn ME, Antonio M, Hossain A, Mandomando I, Ochieng JB, Oundo JO, Qureshi S, Ramamurthy T, Tamboura B, Adegbola RA, Hossain MJ, Saha D, Sen S, Faruque AS, Alonso PL, Breiman RF, Zaidi AK, Sur D, Sow SO, Berkeley LY, O'Reilly CE, Mintz ED, Biswas K, Cohen D, Farag TH, Nasrin D, Wu Y, Blackwelder WC, Kotloff KL, Nataro JP, Levine MM. 2014. Shigella isolates from the global enteric multicenter study inform vaccine development. Clin Infect Dis 59:933–941. https://doi.org/10.1093/cid/ciu468.
17. Lefebvre J, Gosselin F, Ismail J, Lorange M, Lior H, Woodward D. 1995. Evaluation of commercial antisera for Shigella serogrouping. J Clin Microbiol 33:1997–2001.
18. Lindsay B, Ochieng JB, Ikumapayi UN, Toure A, Ahmed D, Li S, Panchalingam S, Levine MM, Kotloff K, Rasko DA, Morris CR, Juma J, Fields BS, Dione M, Malle D, Becker SM, Houpt ER, Nataro JP, Sommerfelt H, Pop M, Oundo J, Antonio M, Hossain A, Tamboura B, Stine OC. 2013. Quantitative PCR for detection of Shigella improves ascertainment of Shigella burden in children with moderate-to-severe diarrhea in low-income countries. J Clin Microbiol 51:1740–1746. https://doi.org/10.1128/JCM.02713-12.
19. Reference deleted.
20. Venkatesan MM, Buysse JM, Kopecko DJ. 1989. Use of Shigella flexneri ipaC and ipaH gene sequences for the general identification of Shigella spp. and enteroinvasive Escherichia coli. J Clin Microbiol 27:2687–2691.
21. Pasqua M, Michelacci V, Di Martino ML, Tozzoli R, Grossi M, Colonna B, Morabito S, Prosseda G. 2017. The intriguing evolutionary journey of enteroinvasive E. coli (EIEC) toward pathogenicity. Front Microbiol 8:2390. https://doi.org/10.3389/fmicb.2017.02390.
22. van den Beld MJ, Reubsaet FA. 2012. Differentiation between Shigella, enteroinvasive Escherichia coli (EIEC) and noninvasive Escherichia coli. Eur J Clin Microbiol Infect Dis 31:899–904. https://doi.org/10.1007/s10096-011-1395-7.
23. Pupo GM, Lan R, Reeves PR. 2000. Multiple independent origins of Shigella clones of Escherichia coli and convergent evolution of many of their characteristics. Proc Natl Acad Sci U S A 97:10567–10572. https://doi.org/10.1073/pnas.180094797.
24. Sahl JW, Morris CR, Emberger J, Fraser CM, Ochieng JB, Juma J, Fields B, Breiman RF, Gilmour M, Nataro JP, Rasko DA. 2015. Defining the phylogenomics of Shigella species: a pathway to diagnostics. J Clin Microbiol 53:951–960. https://doi.org/10.1128/JCM.03527-14.
25. Prosseda G, Di Martino ML, Campilongo R, Fioravanti R, Micheli G, Casalino M, Colonna B. 2012. Shedding of genes that interfere with the pathogenic lifestyle: the Shigella model. Res Microbiol 163:399–406. https://doi.org/10.1016/j.resmic.2012.07.004.
26. Lan R, Alles MC, Donohoe K, Martinez MB, Reeves PR. 2004. Molecular evolutionary relationships of enteroinvasive Escherichia coli and Shi-

gella spp. Infect Immun 72:5080–5088. https://doi.org/10.1128/IAI.72.9.5080-5088.2004.
27. Pupo GM, Karaolis DK, Lan R, Reeves PR. 1997. Evolutionary relationships among pathogenic and nonpathogenic Escherichia coli strains inferred from multilocus enzyme electrophoresis and mdh sequence studies. Infect Immun 65:2685–2692.
28. Chattaway MA, Schaefer U, Tewolde R, Dallman TJ, Jenkins C. 2017. Identification of Escherichia coli and Shigella species from whole-genome sequences. J Clin Microbiol 55:616–623. https://doi.org/10.1128/JCM.01790-16.
29. McArthur MA, Maciel M, Jr, Pasetti MF. 2017. Human immune responses against Shigella and enterotoxigenic E. coli: current advances and the path forward. Vaccine 35:6803–6806. https://doi.org/10.1016/j.vaccine.2017.05.034.
30. Coimbra RS, Grimont F, Grimont PA. 1999. Identification of Shigella serotypes by restriction of amplified O-antigen gene cluster. Res Microbiol 150:543–553. https://doi.org/10.1016/S0923-2508(99)00103-5.
31. Sun Q, Lan R, Wang Y, Zhao A, Zhang S, Wang J, Wang Y, Xia S, Jin D, Cui Z, Zhao H, Li Z, Ye C, Zhang S, Jing H, Xu J. 2011. Development of a multiplex PCR assay targeting O-antigen modification genes for molecular serotyping of Shigella flexneri. J Clin Microbiol 49:3766–3770. https://doi.org/10.1128/JCM.01259-11.
32. van der Ploeg CA, Roge AD, Bordagorria XL, de Urquiza MT, Celi Castillo AB, Bruno SB. 2017. Design of two multiplex PCR assays for serotyping Shigella flexneri. Foodborne Pathog Dis 15:33–38. https://doi.org/10.1089/fpd.2017.2328.
33. Li Y, Cao B, Liu B, Liu D, Gao Q, Peng X, Wu J, Bastin DA, Feng L, Wang L. 2009. Molecular detection of all 34 distinct O-antigen forms of Shigella. J Med Microbiol 58:69–81. https://doi.org/10.1099/jmm.0.000794-0.
34. Wetterstrand K. 2017. DNA sequencing costs: data from the NHGRI Genome Sequencing Program (GSP). https://www.genome.gov/27541954/dna-sequencing-costs-data/.
35. Hyma KE, Lacher DW, Nelson AM, Bumbaugh AC, Janda JM, Strockbine NA, Young VB, Whittam TS. 2005. Evolutionary genetics of a new pathogenic Escherichia species: Escherichia albertii and related Shigella boydii strains. J Bacteriol 187:619–628. https://doi.org/10.1128/JB.187.2.619-628.2005.
36. Simmons DA, Romanowska E. 1987. Structure and biology of Shigella flexneri O antigens. J Med Microbiol 23:289–302. https://doi.org/10.1099/00222615-23-4-289.
37. Clark CA, Beltrame J, Manning PA. 1991. 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 107:43–52. https://doi.org/10.1016/0378-1119(91)90295-M.
38. Gohmann S, Manning PA, Alpert CA, Walker MJ, Timmis KN. 1994. Lipopolysaccharide O-antigen biosynthesis in Shigella dysenteriae serotype 1: analysis of the plasmid-carried rfp determinant. Microb Pathog 16:53–64. https://doi.org/10.1006/mpat.1994.1005.
39. Bastin DA, Lord A, Verma NK. 1997. Cloning and analysis of the glucosyl transferase gene encoding type I antigen in Shigella flexneri. FEMS Microbiol Lett 156:133–139. https://doi.org/10.1111/j.1574-6968.1997.tb12718.x.
40. Huan PT, Bastin DA, Whittle BL, Lindberg AA, Verma NK. 1997. Molecular characterization of the genes involved in O-antigen modification, attachment, integration and excision in Shigella flexneri bacteriophage SfV. Gene 195:217–227. https://doi.org/10.1016/S0378-1119(97)00143-1.
41. Guan S, Verma NK. 1998. Serotype conversion of a Shigella flexneri candidate vaccine strain via a novel site-specific chromosome-integration system. FEMS Microbiol Lett 166:79–87. https://doi.org/10.1111/j.1574-6968.1998.tb13186.x.
42. Adhikari P, Allison G, Whittle B, Verma NK. 1999. Serotype 1a O-antigen modification: molecular characterization of the genes involved and their novel organization in the Shigella flexneri chromosome. J Bacteriol 181:4711–4718.
43. Wang L, Qu W, Reeves PR. 2001. Sequence analysis of four Shigella boydii O-antigen loci: implication for Escherichia coli and Shigella relationships. Infect Immun 69:6923–6930. https://doi.org/10.1128/IAI.69.11.6923-6930.2001.
44. Feng L, Senchenkova SN, Yang J, Shashkov AS, Tao J, Guo H, Zhao G, Knirel YA, Reeves P, Wang L. 2004. Structural and genetic characterization of the Shigella boydii type 13 O antigen. J Bacteriol 186:383–392. https://doi.org/10.1128/JB.186.2.383-392.2004.
45. Feng L, Tao J, Guo H, Xu J, Li Y, Rezwan F, Reeves P, Wang L. 2004. Structure of the Shigella dysenteriae 7 O antigen gene cluster and identification of its antigen specific genes. Microb Pathog 36:109–115. https://doi.org/10.1016/j.micpath.2003.10.003.
46. Tao J, Feng L, Guo H, Li Y, Wang L. 2004. The O-antigen gene cluster of Shigella boydii O11 and functional identification of its wzy gene. FEMS Microbiol Lett 234:125–132. https://doi.org/10.1016/j.femsle.2004.03.021.
47. Feng L, Senchenkova SN, Wang W, Shashkov AS, Liu B, Shevelev SD, Liu D, Knirel YA, Wang L. 2005. Structural and genetic characterization of the Shigella boydii type 18 O antigen. Gene 355:79–86. https://doi.org/10.1016/j.gene.2005.06.001.
48. Senchenkova SN, Feng L, Yang J, Shashkov AS, Cheng J, Liu D, Knirel YA, Reeves PR, Jin Q, Ye Q, Wang L. 2005. Structural and genetic characterization of the Shigella boydii type 10 and type 6 O antigens. J Bacteriol 187:2551–2554. https://doi.org/10.1128/JB.187.7.2551-2554.2005.
49. Tao J, Wang L, Liu D, Li Y, Bastin DA, Geng Y, Feng L. 2005. Molecular analysis of Shigella boydii O1 O-antigen gene cluster and its PCR typing. Can J Microbiol 51:387–392. https://doi.org/10.1139/w05-015.
50. Liu B, Senchenkova SN, Feng L, Perepelov AV, Xu T, Shevelev SD, Zhu Y, Shashkov AS, Zou M, Knirel YA, Wang L. 2006. Structural and molecular characterization of Shigella boydii type 16 O antigen. Gene 380:46–53. https://doi.org/10.1016/j.gene.2006.05.024.
51. Senchenkova SN, Feng L, Wang Q, Perepelov AV, Qin D, Shevelev SD, Ren Y, Shashkov AS, Knirel YA, Wang L. 2006. Structural and genetic characterization of Shigella boydii type 17 O antigen and confirmation of two new genes involved in the synthesis of glucolactilic acid. Biochem Biophys Res Commun 349:289–295. https://doi.org/10.1016/j.bbrc.2006.08.040.
52. Stagg RM, Tang SS, Carlin NI, Talukder KA, Cam PD, Verma NK. 2009. A novel glucosyltransferase involved in O-antigen modification of Shigella flexneri serotype 1c. J Bacteriol 191:6612–6617. https://doi.org/10.1128/JB.00628-09.
53. Sun Q, Lan R, Wang Y, Wang J, Xia S, Wang Y, Zhang J, Yu D, Li Z, Jing H, Xu J. 2012. Identification of a divergent O-acetyltransferase gene oac 1b from Shigella flexneri serotype 1b strains. Emerg Microbes Infect 1:e21. https://doi.org/10.1038/emi.2012.22.
54. Kim J, Lindsey RL, Garcia-Toledo L, Loparev VN, Rowe LA, Batra D, Juieng P, Stoneburg D, Martin H, Knipe K, Smith P, Strockbine N. 2018. High-quality whole-genome sequences for 59 historical Shigella strains generated with PacBio sequencing. Genome Announc 6:e00282-18. https://doi.org/10.1128/genomeA.00282-18.
55. Cho MS, Ahn TY, Joh K, Kwon OS, Jheong WH, Park DS. 2012. A novel marker for the species-specific detection and quantitation of Shigella sonnei by targeting a methylase gene. J Microbiol Biotechnol 22:1113–1117. https://doi.org/10.4014/jmb.1111.11006.
56. Maurelli AT, Blackmon B, Curtiss R, III. 1984. Loss of pigmentation in Shigella flexneri 2a is correlated with loss of virulence and virulence-associated plasmid. Infect Immun 43:397–401.
57. Beutin L, Strauch E, Fischer I. 1999. Isolation of Shigella sonnei lysogenic for a bacteriophage encoding gene for production of Shiga toxin. Lancet 353:1498. https://doi.org/10.1016/S0140-6736(99)00961-7.
58. Gray MD, Lampel KA, Strockbine NA, Fernandez RE, Melton-Celsa AR, Maurelli AT. 2014. Clinical isolates of Shiga toxin 1a-producing Shigella flexneri with an epidemiological link to recent travel to Hispaniola. Emerg Infect Dis 20:1669–1677. https://doi.org/10.3201/eid2010.140292.
59. Gupta SK, Strockbine N, Omondi M, Hise K, Fair MA, Mintz E. 2007. Emergence of Shiga toxin 1 genes within Shigella dysenteriae type 4 isolates from travelers returning from the island of Hispanola. Am J Trop Med Hyg 76:1163–1165. https://doi.org/10.4269/ajtmh.2007.76.1163.
60. Lamba K, Nelson JA, Kimura AC, Poe A, Collins J, Kao AS, Cruz L, Inami G, Vaishampayan J, Garza A, Chaturvedi V, Vugia DJ. 2016. Shiga toxin 1-producing Shigella sonnei infections, California, United States, 2014-2015. Emerg Infect Dis 22:679–686. https://doi.org/10.3201/eid2204.151825.
61. Nyholm O, Lienemann T, Halkilahti J, Mero S, Rimhanen-Finne R, Lehtinen V, Salmenlinna S, Siitonen A. 2015. Characterization of Shigella sonnei isolate carrying Shiga toxin 2-producing gene. Emerg Infect Dis 21:891–892. https://doi.org/10.3201/eid2105.140621.
62. Scheutz F, Teel LD, Beutin L, Pierard D, Buvens G, Karch H, Mellmann A, Caprioli A, Tozzoli R, Morabito S, Strockbine NA, Melton-Celsa AR, Sanchez M, Persson S, O'Brien AD. 2012. Multicenter evaluation of a sequence-based protocol for subtyping Shiga toxins and standardizing

Stx nomenclature. J Clin Microbiol 50:2951–2963. https://doi.org/10.1128/JCM.00860-12.

63. Walters LL, Raterman EL, Grys TE, Welch RA. 2012. Atypical Shigella boydii 13 encodes virulence factors seen in attaching and effacing Escherichia coli. FEMS Microbiol Lett 328:20–25. https://doi.org/10.1111/j.1574-6968.2011.02469.x.

64. Lobersli I, Wester AL, Kristiansen A, Brandal LT. 2016. Molecular differentiation of Shigella spp. from enteroinvasive E. coli. Eur J Microbiol Immunol 6:197–205. https://doi.org/10.1556/1886.2016.00004.

65. Hazen TH, Leonard SR, Lampel KA, Lacher DW, Maurelli AT, Rasko DA. 2016. Investigating the relatedness of enteroinvasive Escherichia coli to other E. coli and Shigella isolates by using comparative genomics. Infect Immun 84:2362–2371. https://doi.org/10.1128/IAI.00350-16.

66. Casalino M, Latella MC, Prosseda G, Ceccarini P, Grimont F, Colonna B. 2005. Molecular evolution of the lysine decarboxylase-defective phenotype in Shigella sonnei. Int J Med Microbiol 294:503–512. https://doi.org/10.1016/j.ijmm.2004.11.001.

67. Day WA, Jr, Fernandez RE, Maurelli AT. 2001. Pathoadaptive mutations that enhance virulence: genetic organization of the cadA regions of Shigella spp. Infect Immun 69:7471–7480. https://doi.org/10.1128/IAI.69.12.7471-7480.2001.

68. Joensen KG, Tetzschner AM, Iguchi A, Aarestrup FM, Scheutz F. 2015. Rapid and easy in silico serotyping of Escherichia coli isolates by use of whole-genome sequencing data. J Clin Microbiol 53:2410–2426. https://doi.org/10.1128/JCM.00008-15.

69. Chen S, Zhou Y, Chen Y, Gu J. 2018. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics 34:i884–i890. https://doi.org/10.1093/bioinformatics/bty560.

70. Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34:3094–3100. https://doi.org/10.1093/bioinformatics/bty191.

71. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078–2079. https://doi.org/10.1093/bioinformatics/btp352.

72. Grozdanov L, Zahringer U, Blum-Oehler G, Brade L, Henne A, Knirel YA, Schombel U, Schulze J, Sonnenborn U, Gottschalk G, Hacker J, Rietschel ET, Dobrindt U. 2002. A single nucleotide exchange in the wzy gene is responsible for the semirough O6 lipopolysaccharide phenotype and serum sensitivity of Escherichia coli strain Nissle 1917. J Bacteriol 184:5912–5925. https://doi.org/10.1128/JB.184.21.5912-5925.2002.

73. Kia A, Gloeckner C, Osothprarop T, Gormley N, Bomati E, Stephenson M, Goryshin I, He MM. 2017. Improved genome sequencing using an engineered transposase. BMC Biotechnol 17:6. https://doi.org/10.1186/s12896-016-0326-1.

74. Green B, Bouchier C, Fairhead C, Craig NL, Cormack BP. 2012. Insertion site preference of Mu, Tn5, and Tn7 transposons. Mob DNA 3:3. https://doi.org/10.1186/1759-8753-3-3.

75. Knirel YA, Sun Q, Senchenkova SN, Perepelov AV, Shashkov AS, Xu J. 2015. O-antigen modifications providing antigenic diversity of Shigella flexneri and underlying genetic mechanisms. Biochemistry 80:901–914. https://doi.org/10.1134/S0006297915070093.

76. Sansonetti PJ, Kopecko DJ, Formal SB. 1981. Shigella sonnei plasmids: evidence that a large plasmid is necessary for virulence. Infect Immun 34:75–83.

77. Chattaway MA, Greig DR, Gentle A, Hartman HB, Dallman TJ, Jenkins C. 2017. Whole-genome sequencing for national surveillance of Shigella flexneri. Front Microbiol 8:1700. https://doi.org/10.3389/fmicb.2017.01700.

78. Svab D, Balint B, Vasarhelyi B, Maroti G, Toth I. 2017. Comparative genomic and phylogenetic analysis of a Shiga toxin producing Shigella sonnei (STSS) strain. Front Cell Infect Microbiol 7:229. https://doi.org/10.3389/fcimb.2017.00229.

79. Di Martino ML, Fioravanti R, Barbabella G, Prosseda G, Colonna B, Casalino M. 2013. Molecular evolution of the nicotinic acid requirement within the Shigella/EIEC pathotype. Int J Med Microbiol 303:651–661. https://doi.org/10.1016/j.ijmm.2013.09.007.

80. Zhang S, Yin Y, Jones MB, Zhang Z, Deatherage Kaiser BL, Dinsmore BA, Fitzgerald C, Fields PI, Deng X. 2015. Salmonella serotype determination utilizing high-throughput genome sequencing data. J Clin Microbiol 53:1685–1692. https://doi.org/10.1128/JCM.00323-15.

81. Zankari E, Hasman H, Cosentino S, Vestergaard M, Rasmussen S, Lund O, Aarestrup FM, Larsen MV. 2012. Identification of acquired antimicrobial resistance genes. J Antimicrob Chemother 67:2640–2644. https://doi.org/10.1093/jac/dks261.

82. Hunt M, Mather AE, Sanchez-Buso L, Page AJ, Parkhill J, Keane JA, Harris SR. 2017. ARIBA: rapid antimicrobial resistance genotyping directly from sequencing reads. Microb Genom 3:e000131. https://doi.org/10.1099/mgen.0.000131.

83. Inouye M, Dashnow H, Raven LA, Schultz MB, Pope BJ, Tomita T, Zobel J, Holt KE. 2014. SRST2: rapid genomic surveillance for public health and hospital microbiology labs. Genome Med 6:90. https://doi.org/10.1186/s13073-014-0090-6.

84. Li H. 2011. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics 27:2987–2993. https://doi.org/10.1093/bioinformatics/btr509.

85. Gupta A, Jordan IK, Rishishwar L. 2017. stringMLST: a fast k-mer based tool for multilocus sequence typing. Bioinformatics 33:119–121. https://doi.org/10.1093/bioinformatics/btw586.

86. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol 19:455–477. https://doi.org/10.1089/cmb.2012.0021.

87. Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, Formsma K, Gerdes S, Glass EM, Kubal M, Meyer F, Olsen GJ, Olson R, Osterman AL, Overbeek RA, McNeil LK, Paarmann D, Paczian T, Parrello B, Pusch GD, Reich C, Stevens R, Vassieva O, Vonstein V, Wilke A, Zagnitko O. 2008. The RAST Server: rapid annotations using subsystems technology. BMC Genomics 9:75. https://doi.org/10.1186/1471-2164-9-75.

88. Overbeek R, Olson R, Pusch GD, Olsen GJ, Davis JJ, Disz T, Edwards RA, Gerdes S, Parrello B, Shukla M, Vonstein V, Wattam AR, Xia F, Stevens R. 2014. The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). Nucleic Acids Res 42:D206–D214. https://doi.org/10.1093/nar/gkt1226.

89. Brettin T, Davis JJ, Disz T, Edwards RA, Gerdes S, Olsen GJ, Olson R, Overbeek R, Parrello B, Pusch GD, Shukla M, Thomason JA, III, Stevens R, Vonstein V, Wattam AR, Xia F. 2015. RASTtk: a modular and extensible implementation of the RAST algorithm for building custom annotation pipelines and annotating batches of genomes. Sci Rep 5:8365. https://doi.org/10.1038/srep08365.

90. Darling AE, Mau B, Perna NT. 2010. progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. PLoS One 5:e11147. https://doi.org/10.1371/journal.pone.0011147.

91. Jin Q, Yuan Z, Xu J, Wang Y, Shen Y, Lu W, Wang J, Liu H, Yang J, Yang F, Zhang X, Zhang J, Yang G, Wu H, Qu D, Dong J, Sun L, Xue Y, Zhao A, Gao Y, Zhu J, Kan B, Ding K, Chen S, Cheng H, Yao Z, He B, Chen R, Ma D, Qiang B, Wen Y, Hou Y, Yu J. 2002. Genome sequence of Shigella flexneri 2a: insights into pathogenicity through comparison with genomes of Escherichia coli K12 and O157. Nucleic Acids Res 30:4432–4441. https://doi.org/10.1093/nar/gkf566.

92. Jiang Y, Yang F, Zhang X, Yang J, Chen L, Yan Y, Nie H, Xiong Z, Wang J, Dong J, Xue Y, Xu X, Zhu Y, Chen S, Jin Q. 2005. The complete sequence and analysis of the large virulence plasmid pSS of Shigella sonnei. Plasmid 54:149–159. https://doi.org/10.1016/j.plasmid.2005.03.002.

93. Jakhetia R, Talukder KA, Verma NK. 2013. Isolation, characterization and comparative genomics of bacteriophage SfIV: a novel serotype converting phage from Shigella flexneri. BMC Genomics 14:677. https://doi.org/10.1186/1471-2164-14-677.

94. Huan PT, Whittle BL, Bastin DA, Lindberg AA, Verma NK. 1997. Shigella flexneri type-specific antigen V: cloning, sequencing and characterization of the glucosyl transferase gene of temperate bacteriophage SfV. Gene 195:207–216. https://doi.org/10.1016/S0378-1119(97)00144-3.

95. Huan PT, Taylor R, Lindberg AA, Verma NK. 1995. Immunogenicity of the Shigella flexneri serotype Y (SFL 124) vaccine strain expressing cloned glucosyl transferase gene of converting bacteriophage SfX. Microbiol Immunol 39:467–472. https://doi.org/10.1111/j.1348-0421.1995.tb02230.x.

96. Tang SS, Carlin NI, Talukder KA, Cam PD, Verma NK. 2016. Shigella flexneri serotype 1c derived from serotype 1a by acquisition of gtrIC gene cluster via a bacteriophage. BMC Microbiol 16:127. https://doi.org/10.1186/s12866-016-0746-z.

97. Casjens S, Winn-Stapley DA, Gilcrease EB, Morona R, Kuhlewein C, Chua JE, Manning PA, Inwood W, Clark AJ. 2004. The chromosome of Shigella flexneri bacteriophage Sf6: complete nucleotide sequence, genetic

mosaicism, and DNA packaging. J Mol Biol 339:379–394. https://doi.org/10.1016/j.jmb.2004.03.068.

98. Ye C, Lan R, Xia S, Zhang J, Sun Q, Zhang S, Jing H, Wang L, Li Z, Zhou Z, Zhao A, Cui Z, Cao J, Jin D, Huang L, Wang Y, Luo X, Bai X, Wang Y, Wang P, Xu Q, Xu J. 2010. Emergence of a new multidrug-resistant serotype X variant in an epidemic clone of Shigella flexneri. J Clin Microbiol 48:419–426. https://doi.org/10.1128/JCM.00614-09.
99. Yang F, Yang J, Zhang X, Chen L, Jiang Y, Yan Y, Tang X, Wang J, Xiong Z, Dong J, Xue Y, Zhu Y, Xu X, Sun L, Chen S, Nie H, Peng J, Xu J, Wang Y, Yuan Z, Wen Y, Yao Z, Shen Y, Qiang B, Hou Y, Yu J, Jin Q. 2005. Genome dynamics and diversity of Shigella species, the etiologic agents of bacillary dysentery. Nucleic Acids Res 33:6445–6458. https://doi.org/10.1093/nar/gki954.
100. Suzuki S, Ono N, Furusawa C, Ying BW, Yomo T. 2011. Comparison of sequence reads obtained from three next-generation sequencing platforms. PLoS One 6:e19534. https://doi.org/10.1371/journal.pone.0019534.
101. Fasano A, Noriega FR, Maneval DR, Jr, Chanasongcram S, Russell R, Guandalini S, Levine MM. 1995. Shigella enterotoxin 1: an enterotoxin of Shigella flexneri 2a active in rabbit small intestine in vivo and in vitro. J Clin Invest 95:2853–2861. https://doi.org/10.1172/JCI117991.
102. Nataro JP, Seriwatana J, Fasano A, Maneval DR, Guers LD, Noriega F, Dubovsky F, Levine MM, Morris JG, Jr. 1995. Identification and cloning of a novel plasmid-encoded enterotoxin of enteroinvasive Escherichia coli and Shigella strains. Infect Immun 63:4721–4728.
103. Guyer DM, Radulovic S, Jones FE, Mobley HL. 2002. Sat, the secreted autotransporter toxin of uropathogenic Escherichia coli, is a vacuolating cytotoxin for bladder and kidney epithelial cells. Infect Immun 70:4539–4546. https://doi.org/10.1128/IAI.70.8.4539-4546.2002.