MICROBIAL GENOMICS

RESEARCH ARTICLE
Bernaquez *et al.*, *Microbial Genomics* 2021;7:000672
DOI 10.1099/mgen.0.000672

Evaluation of whole-genome sequencing-based subtyping methods for the surveillance of *Shigella* spp. and the confounding effect of mobile genetic elements in long-term outbreaks

Isabelle Bernaquez[1], Christiane Gaudreau[2,3], Pierre A. Pilon[4,5] and Sadjia Bekal[1,3,*]

Abstract

Many public health laboratories across the world have implemented whole-genome sequencing (WGS) for the surveillance and outbreak detection of foodborne pathogens. PulseNet-affiliated laboratories have determined that most single-strain food-borne outbreaks are contained within 0–10 multi-locus sequence typing (MLST)-based allele differences and/or core genome single-nucleotide variants (SNVs). In addition to being a food- and travel-associated outbreak pathogen, most *Shigella* spp. cases occur through continuous person-to-person transmission, predominantly involving men who have sex with men (MSM), leading to long-term and recurrent outbreaks. Continuous transmission patterns coupled to genetic evolution under antibiotic treatment pressure require an assessment of existing WGS-based subtyping methods and interpretation criteria for cluster inclusion/exclusion. An evaluation of 4 WGS-based subtyping methods [SNVPhyl, coreMLST, core genome MLST (cgMLST) and whole-genome MLST (wgMLST)] was performed on 9 foodborne-, travel- and MSM-related retrospective outbreaks from a collection of 91 *Shigella flexneri* and 232 *Shigella sonnei* isolates to determine the methods' epidemiological concordance, discriminatory power, robustness and ability to generate stable interpretation criteria. The discriminatory powers were ranked as follows: coreMLST <SNVPhyl<cgMLST <wgMLST (range: 0.970–1.000). The genetic differences observed for non-MSM-related *Shigella* spp. outbreaks respect the standard 0–10 allele/SNV guideline; however, mobile genetic element (MGE)-encoded loci caused inflated genetic variation and discrepant phylogenies for prolonged MSM-related *S. sonnei* outbreaks via wgMLST. The *S. sonnei* correlation coefficients of wgMLST were also the lowest at 0.680, 0.703 and 0.712 for SNVPhyl, coreMLST and cgMLST, respectively. Plasmid maintenance, mobilization and conjugation-associated genes were found to be the main source of genetic distance inflation in addition to prophage-related genes. Duplicated alleles arising from the repeated nature of IS elements were also responsible for many false cg/wgMLST differences. The coreMLST approach was shown to be the most robust, followed by SNVPhyl and wgMLST for inter-laboratory comparability. Our results highlight the need for validating species-specific subtyping methods based on microbial genome plasticity and outbreak dynamics in addition to the importance of filtering confounding MGEs for cluster detection.

DATA SUMMARY

The fastq reads from all sequences analysed in this article are available at the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA), available at https://www.ncbi.nlm.nih.gov/sra under BioProject PRJNA639996. BioSample accession numbers for *S. flexneri* and *S. sonnei* can be found in Table S1 (available in the online version of this article).

Received 18 March 2021; Accepted 14 August 2021; Published 03 November 2021

Author affiliations: [1]Laboratoire de santé publique du Québec, Sainte-Anne-de-Bellevue, QC, H9X 3R5, Canada; [2]Microbiologie médicale et infectiologie, Centre Hospitalier de l'Université de Montréal (CHUM), Montreal, QC, H2X 3E4, Canada; [3]Département de microbiologie, infectiologie et immunologie, Université de Montréal, Montreal, QC, H3C 3J7, Canada; [4]Direction régionale de santé publique, Centre intégré universitaire de santé et de services sociaux du Centre-Sud-de-l'île-de-Montréal, Montreal, QC, H2L 4M1, Canada; [5]Département de médecine sociale et préventive, Université de Montréal, Montreal, QC, H3N 1X9, Canada.
***Correspondence:** Sadjia Bekal, sadjia.bekal@inspq.qc.ca

Keywords: *Shigella*; whole-genome sequencing; outbreak; whole/core genome MLST; SNVPhyl; mobile genetic elements.

Abbreviations: CG, Core Genome; DI, diversity index; hqSNV, high-quality single-nucleotide variant; LSPQ, Laboratoire de santé publique du Québec; MGE, mobile genetic element; ML, maximum-likelihood; MLST, multi-locus sequence typing; MS, minimum spanning; MSM, men who have sex with men; NDC, non-documented case; NML, National Microbiology Laboratory; PNC, PulseNet Canada; PNI, PulseNet International; SNVPhyl, single-nucleotide variant phylogenomics pipeline; ST, sequence type; WG, Whole Genome; WGS, whole-genome sequencing.

New short-read sequence data can be accessed in the NCBI's BioProject PRJNA639996.

Data statement: All supporting data, code and protocols have been provided within the article or through supplementary data files.Two supplementary tables and five supplementary figures are available with the online version of this article.

DATA STATEMENT

The authors confirm that all supporting data, code and protocols have been provided within the article or through supplementary data files. Two supplementary tables and five supplementary figures are available with the online version of this article and can be found at https://doi.org/10.6084/m9.figshare.14757978.v1 [1].

Impact Statement

The implementation of routine whole-genome sequencing (WGS) for bacterial pathogens has significantly improved international surveillance and outbreak detection in public health laboratories. Continuous efforts have been directed towards subtyping method improvement and development, as well as genomic interpretation guideline definitions for cluster detection. However, most studies have focused on foodborne pathogens and outbreaks, limiting the use of the established WGS interpretation guidelines for long-term outbreaks (i.e. continuous person-to-person transmissions). As these types of outbreaks were the dominant sources of *Shigella flexneri* and *Shigella sonnei* cases in the province of Quebec, this study provides an insight into the possible variability of interpretation criteria recommended for whole-genome multi-locus sequence typing (wgMLST). Our evaluation of wgMLST, which considers genetic variations in the accessory genome, against a robust core SNV typing pipeline and alternative core genome multi-locus sequence typing-based methods, highlights the need of validating species-specific subtyping methods based on microbial genome plasticity and outbreak dynamics in addition to the importance of filtering confounding mobile genetic elements (MGEs) for cluster detection. We also demonstrate the use of minimum spanning genetic distances as an alternative measure for WGS interpretation guidelines that can be applied to outbreaks of different transmission natures.

INTRODUCTION

Shigella spp. are facultative intracellular pathogens transmitted solely through the faecal–oral route and are the causative agents of shigellosis. While the rate of reported shigellosis cases in Canada has been steadily decreasing over time [2], *Shigella flexneri* and *Shigella sonnei* cases in large cities have become prevalent due to extensive sexual transmission networks involving men who have sex with men (MSM) [3–7], in addition to periodic food-, water- and travel-associated outbreaks in the general population [8–12]. Similar MSM-related outbreaks have also been documented globally and have been described as potential public health concerns due to their widespread resistances to antimicrobial treatments [13–18], occurrence of life-threatening invasive infections in patients co-infected with HIV [5, 6] and potential to spill into other at-risk communities.

At any given time, several strains of the same pathogen can circulate through a population, making outbreak investigations more challenging. Standardized, epidemiologically concordant, and highly discriminatory subtyping methods are thus needed to differentiate between sporadic and outbreak-related cases and to detect source attributions, amongst other interests [19]. In fact, PulseNet International (PNI), an international public health laboratory network [20], has started implementing whole-genome sequencing (WGS) in order to enhance their surveillance power, as it can provide the absolute maximum resolution profile possible for any organism [21]. Currently, a gene-by-gene approach, specifically, a whole-genome multi-locus sequence typing (wgMLST) method has been chosen as the official replacement for standardized foodborne infectious disease surveillance and outbreak response by PNI [21].

The wgMLST method detects allelic variations (e.g. point mutations, indels and small translocations) and infers phylogeny based on a large pan-genomic scheme of 17 380 loci, where each allele combination profile corresponds to a unique sequence type (ST). However, other high-quality WGS-based subtyping methods exist, by which public health laboratories may find useful to support challenging investigations. Such methods include core genome MLST-based approaches [e.g. EnteroBase core genome MLST (coreMLST) and core genome MLST (cgMLST)] and single-nucleotide variant (SNV)-based approaches (e.g. high-quality core genome SNV typing by SNVPhyl) [22–26]. cgMLST uses the same pan-genomic loci scheme as wgMLST but only infers phylogeny based on loci present in a user-defined percentage (e.g.≥95%) of the isolates under comparison. A second core genome MLST method was included in this study and differs from cgMLST by its use of the curated EnteroBase scheme (https://enterobase.warwick.ac.uk) consisting of 2513 highly conserved *Escherichia coli*/*Shigella* loci [27]. Similar core genome schemes have been shown to be standardized and scalable for inter-laboratory comparisons for other enteric pathogens [25, 28, 29]. Alternatively, SNVPhyl detects common high-quality SNVs from a reference-based mapping approach, which stands in contrast to the MLST-based subtyping methods through its capacity to include intergenic sequence variants [30]. Similar core genome SNV-based approaches are currently used for routine surveillance by Public Health England and the Centers for Disease Control and Prevention in the USA.

According to PNI, the wgMLST approach was selected based on its ability to integrate into routine public health surveillance, and to maintain real-time outbreak detection and inter-laboratory comparability [21]. However, an internal validation is required to verify the performance of any new method using a well-defined set of isolates and outbreaks. PulseNet Canada (PNC) has made progress towards the implementation of WGS for foodborne pathogen surveillance [22–26]. Studies on the top priority pathogens – *Listeria monocytogenes*, *E. coli* O157:H7 and common *Salmonella enterica* serovars (e.g.

Heidelberg, Enteritidis and Typhimurium) – have determined that most single-strain outbreaks are contained within a 0–10 allele and/or high-quality core SNV (hqSNV) diversity [23–25, 31, 32]. Similar performance assessment studies are now needed to determine the genetic interpretation guidelines for *Shigella* spp. outbreak detections. However, while *Shigella* spp. are phylogenetically considered *E. coli* [33], we hypothesize that existing WGS-based interpretation criteria might not be suitable for all *Shigella*-related outbreaks due to differences in epidemiology, environmental pressures and genomic plasticity. This diversity is thought to be due to the higher level of genetic variation caused by prolonged continuous transmission patterns, bottleneck-derived stochastic effects due to a low infectious dose [34, 35] and frequent use of antibiotics for other sexually transmitted diseases by MSM [13, 36], in addition to an inherent higher incidence of mobile genetic elements [37, 38], facility of horizontal gene transfer [39, 40] and dynamic prophage acquisition and loss [41] through time and across hosts.

In this study, four WGS-based subtyping methods (SNVPhyl, coreMLST, cgMLST and wgMLST) were evaluated based on discriminatory power, epidemiological concordance, compatibility, robustness and ability to generate stable WGS interpretation criteria for outbreak investigations using 9 foodborne-, travel- and MSM-related retrospective outbreaks of a collection of 91 *S. flexneri* and 232 *S. sonnei* isolates collected from the province of Quebec. We also aimed to define the number of allele or hqSNV differences regularly observed between outbreak-related isolates from different transmission natures (e.g. point source vs person-to-person transmission). These genetic differences will offer a preliminary guideline for public health officials to delimit outbreaks based on WGS data and to conduct effective routine surveillance of *S. flexneri* and *S. sonnei* in different contexts. The location of the genetic differences identified was also explored to determine the extent of genetic polymorphisms detected on mobile genetic elements (MGEs). This study therefore contributes to the ongoing validation of WGS-based subtyping methods for bacterial pathogen surveillance in Canada.

METHODS

Isolate characterization and dataset selection

Bacterial isolates from *Shigella* spp. cases in Quebec were sent to the Laboratoire de santé publique du Québec (LSPQ) in Sainte-Anne-de-Bellevue, QC for serotyping and molecular typing. Species identification and serotyping was performed at the LSPQ using slide agglutination (Denka Seiken Co., Ltd, Coventry, UK). *Shigella* species were also confirmed by bioinformatic analysis through the ecTyper pipeline (https://github.com/phac-nml/ecoli_serotyping).

Clinical cases of *S. flexneri* and *S. sonnei* were selected for analysis from the province of Quebec. Nine outbreaks involving three or more epidemiologically supported cases with a suspected source of infection were chosen for further analysis. Some subclusters were defined inside the large outbreaks to feature hypothetical transmission routes. Outbreaks and subclusters were delimited based on the SNVPhyl phylogenetic tree topology and distance metrics due to the recognition of SNV-based subtyping methods as strong subtyping tools in public health [22, 42], along with epidemiological evidence and temporality. Epidemiological curves for each outbreak clade can be observed in Fig. S1. Isolates with no epidemiological evidence available were designated as non-documented cases (NDCs). NDCs were included in the outbreaks when they clustered concordantly with the other epidemiologically supported isolates. Isolates clearly clustering outside of the nine well-characterized outbreaks were considered sporadic. Sporadic cases distanced by more than 100 alleles of the selected outbreaks via wgMLST were excluded for tree simplicity. The *S. flexneri* datasets included four outbreaks (SF-01 to SF-04), where two large MSM-related outbreaks (SF-01 and SF-02) were also further dissected into subclusters designated with letters (a through d). The *S. sonnei* dataset included five outbreaks (SS-01 to SS-05), where two large MSM-related outbreaks (SS-03 and SS-05) were also further divided into subclusters (a through d). Outbreaks and subclusters were characterized as 'MSM-related' when most cases were adult male with the presence of male cases with epidemiological data describing recent same-sex sexual relationships. All other outbreaks and subclusters are sometimes referred to as 'non-MSM' in this study. The final isolate collections of *S. flexneri* (*n*=91) and *S. sonnei* (*n*=232) used in this study are listed in Table S1 along with their BioProject and BioSample accession numbers.

WGS and quality control

Shigella spp. isolate sequencing was performed at the National Microbiology Laboratory (NML) in Winnipeg, MB. Sequencing protocols were followed as described previously [24]. The quality of the raw sequence reads and draft assemblies was evaluated using FastQC and BioNumerics v7.6. FastQC was incorporated within the Integrated Rapid Infectious Disease Analysis (IRIDA) platform to determine the average genome coverage and to check the basic sequencing quality metrics [43]. Only genomes with ≥40× coverage based on a 5 000 000 bp genome size were used in this study. The reads were then *de novo*-assembled using the SPAdes assembler (min contig size=1000) in BioNumerics v7.6 software [44]. Six quality metrics were also verified following *Shigella*-specific PNC guidelines: AvgQuality ≥30, N50 ≥18 000, NrContigs <500, length 4.4 to 5.0 Mb, NrConsensus (congruent allele calls by both assembly-based and assembly-free algorithms) >3300 and CorePercent ≥90%.

Phylogenomic analysis

SNV-based subtyping

Reference-based core genome SNV detection was performed using the Single Nucleotide Variant Phylogenomics (SNVPhyl) pipeline v.1.1 [30] integrated within the NML Galaxy platform [45]. Reference chromosomes of *S. flexneri* serotype 2 a strain 301 (NC_004337.2) and *S. sonnei* Ss046 (NC_007384.1) were used for mapping the reads of all *S.*

flexneri and *S. sonnei* isolates, respectively. PHASTER [46] and Island Viewer 4 [47] were used to identify prophage and genomic island sequences in the reference chromosomes for mapping exclusion. Pipeline parameters were followed as per the PNC guidelines and as described elsewhere [24]. Both vcf2core outputs were verified for ≥85% of valid and included positions in the core genome prior to PHASTER and IslandViewer 4 masking. A script [48] was used to supplement the maximum-likelihood (ML) phylogenetic tree outputs from SNVPhyl with the number of genetic differences associated for each branch for its visualization in FigTree v1.4.3 [49]. A second script [50] was used to generate the proper input files to visualize the clusters generated by SNVPhyl by constructing minimum-spanning (MS) trees in PHYLOViZ v2.0 with the goeBURST algorithm [51]. This script generates a profile table of each unique SNV site combination and assigns a number for each profile (i.e. sequence type) and also generates a table summarizing the profiles identified for each input isolate. The pairwise hqSNV distances were available in the snvMatrix output.

Allele-based subtyping

Three multi-locus sequence typing (MLST) methods were performed by the BioNumerics v7.6 software as described elsewhere [24]. MS and ML trees for categorical data were also constructed as follows. (a) EnteroBase core genome MLST (coreMLST) was performed by using the *E. coli/Shigella*-specific EnteroBase scheme consisting of 2513 core loci (https://enterobase.warwick.ac.uk). (b) Core genome MLST (cgMLST) was performed using the complete *E. coli/Shigella*-specific loci scheme consisting of 17 380 loci. However, loci used for analysis were required to be present in ≥95% of the isolates under comparison. (c) Whole-genome MLST (wgMLST) was also performed using the *E. coli/Shigella*-specific loci scheme consisting of 17 380 loci. Any loci present in at least one isolate were included in the wgMLST analysis.

When multiple consensus allele calls were identified for a single locus, only the allele call with the lowest allele identification number was tabulated via BioNumerics' default settings [27]. Hierarchical clustering was performed using the *categorical (values)* similarity coefficient measured as distances, which summarizes the pairwise allele differences between isolates by the following equation:

$$\text{categorical (values)}=(n\ \text{total}/n\ \text{common})\times n\ \text{differences},$$

where *n* total is the total number of loci present in the comparison of the total 91 *S. flexneri* or 232 *S. sonnei* isolates under study, *n* common is the number of loci that are present in the pair of isolates under examination and *n* differences is the number of common loci with different allele calls in the same pair of isolates.

All MLST distances reported in this study were the result of this equation and were rounded to their closest integer unless stated otherwise.

Discriminatory power and statistical analysis

Each subtyping method's ability to differentiate isolates and consequently assign different subtypes was evaluated using Simpson's diversity index (DI) [52]. The STs generated by SNVPhyl were resolved through the same script described in the previous section [50]. All STs from the MLST-based subtyping methods were identified through visualization of the MS trees, where all isolates included in a single node were given the same ST. The 95% confidence intervals and *P*-values were calculated using the jackknife pseudo-values method to indicate significant differences in discriminatory power [53]. All indices, intervals and *P*-values were calculated using the Comparing Partitions website (www.comparingpartitions.info). A *P*-value ≤0.05 was considered as statistically significant.

Genetic distance comparison and linear regressions

Pairwise MLST-based allelic differences (*y*-axis) for each outbreak were plotted against their respective hqSNV difference (*x*-axis). Pairwise differences involving sporadic isolates and NDCs were excluded in addition to inter-outbreak differences. A simple linear regression analysis was added for each MLST-based subtyping method to produce a best-fit line, its formula ($y=mx+b$) and its variability (R^2) via GraphPad Prism 8. The slope of the trendline (*m*) is indicative of the number of MLST-based allele differences found per hqSNV identified by SNVPhyl. Pearson correlation coefficients were also calculated via GraphPad Prism 8 in order to compare genetic distance correlations between all approaches for both species.

High-quality SNV and MLST loci locations and classification

The location of the hqSNVs identified by the SNVPhyl pipeline was determined by a third script [54], which mapped the position of the hqSNVs to the GenBank file of the reference genome used by SNVPhyl and then generated a table summarizing the features (e.g. CDS, miscellaneous feature, intergenic, mobile element, RNA, etc.) and gene products associated with each hqSNV position. Miscellaneous features were only present in the *S. flexneri* reference genome (NC_004337.2). IS elements were identified via the presence of transposase genes.

The loci corresponding to allele differences in BioNumerics were resolved by exporting the character data matrix for each dataset. The BioNumerics loci were divided into six categories according to their genome location, function, number of detected alleles per locus and number of mutations between different allele calls: chromosome, chromosome repeats, plasmid mutations, plasmid copies, IS element-related and prophage-related. The wgMLST quality assessment tool in BioNumerics enabled the identification of repeated loci and the number of mutations between allele calls. Loci with multiple allele calls for a single locus found in contigs of chromosomal origin were considered to be 'chromosome repeats'. Plasmid-related loci were classified on a case-by-case basis. MOB-recon was used to extract and type the plasmids from the BioNumerics SPAdes draft assemblies of key outbreak-related isolates. Each contig of plasmid origin is given a cluster code based on its similarity to a reference plasmid 'type' assigned by MOB-cluster. MOB-typer was used to predict conjugation potential [55]. The default minimum e-value threshold (1e-05) and minimum sequence identity (80%) were used for BLASTN. Generally, loci considered as 'plasmid mutations' represent allele differences with ≤2 mutations encoded on the same plasmid cluster, which would be indicative of the vertical evolution of plasmid sequences. Loci with >2 mutations were considered 'plasmid copies' whether or not these loci were present on the same plasmid cluster and were indicative of HGT events. Loci with multiple allele calls were also found in contigs of plasmid origin, however, these were not easily distinguishable from 'plasmid copies' and were thus grouped according to the number of mutations between allele calls as mentioned above. IS elements were identified via the presence of transposase genes. PHASTER [46] was used to identify prophage-related sequences in the draft assemblies, but only the loci mapped to intact prophages were categorized as 'prophage-related'.

MLST locus locations were not mutually exclusive. Therefore, a binning approach was followed based on the average sequence length and mobility range (chromosome <plasmids<prophages<IS elements). For example, if an IS element-related transposase gene was identified on a contig of plasmid origin, this gene was only tabulated as IS element-related.

Robustness comparisons

The variability of all four WGS-based subtyping methods was evaluated by comparing the intra-outbreak genetic distances found through the use of the complete collection of species-specific isolates and then isolating each outbreak and subcluster as distinct inputs. These two analyses should represent both extremes of distance range variability and therefore indicate their potential effect on future interpretation guidelines. Box and Tukey whisker plots of the pairwise genetic distances for all outbreaks and subclusters from both datasets were generated via GraphPad Prism 8. Phylogenetic trees were visually compared to detect major differences in clustering.

RESULTS

Phylogenomic comparisons between SNVPhyl and MLST-based subtyping methods

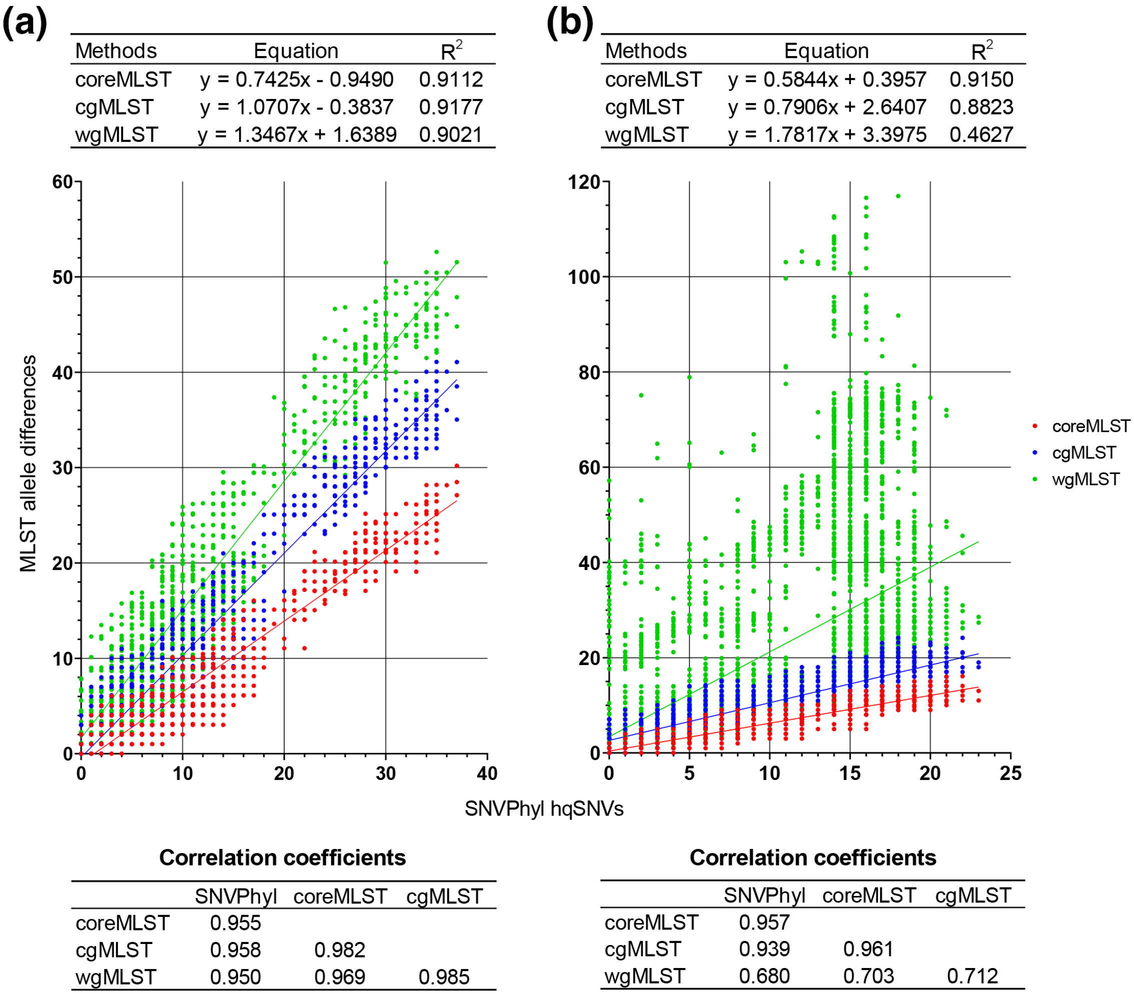
The ML trees generated via SNVPhyl for *S. flexneri* and *S. sonnei* and used as the basis of outbreak and subcluster delimitations are shown in Figs S2 and S3, respectively. Visual comparisons of the MS trees shown in Figs S4 and S5 generated by the four WGS-based subtyping methods for *S. flexneri* and *S. sonnei*, respectively, concluded that all outbreaks (indicated by 1 through 5) and most subclusters (indicated by a through d) delimited via SNVPhyl were similarly grouped by the MLST-based approaches. However, several single *S. sonnei* isolates seem to cluster separately from their other subcluster-related isolates (SS-03b and SS-05b) via wgMLST with relatively large distances and are indicated by W, X, Y and Z in Fig. S5.

Following these phylogenetic discrepancies observed via wgMLST, the extent of wgMLST's correlation to the core genome-based subtyping methods was evaluated. For each *Shigella* spp. dataset, all outbreak-related pairwise genetic distances of the three MLST-based subtyping methods were combined and plotted against their respective hqSNV distances in the scatter plot illustrated in Fig. 1. The amount of allele differences identified by all MLST-based subtyping methods for *S. flexneri* was consistent in comparison to the number of hqSNVs located via SNVPhyl as the R^2 were all above 0.90. These results correlate with the phylogenetic tree similarities for *S. flexneri* in Fig. S4. The slopes of the linear regressions were given as follows: coreMLST (0.74)<cgMLST (1.07)<wgMLST (1.35). Here, wgMLST detects 35% more differences than SNVPhyl. For *S. sonnei*, similar trends were observed, with the exception of wgMLST, as the number of pairwise allele differences is weakly correlated with the number of hqSNVs (R^2=0.46). However, wgMLST is the only method that could detect more variation (78%). Notably, the wgMLST-based allele differences can span upwards of 100 to 117 differences, while these same pairwise differences via SNVPhyl only span 11 to 18 hqSNVs. In addition, isolates indistinguishable via SNVPhyl (i.e. identical STs) can be differentiated by upwards of 57 whole-genome (WG) alleles. *S. sonnei* correlation coefficients for wgMLST were also the lowest at 0.680, 0.703 and 0.712 for SNVPhyl, coreMLST and cgMLST, respectively, while all other correlation coefficients ranged from 0.939 to 0.985 for both species. However, it is important to note that most discordant results seem to be caused by only a handful of isolates, particularly the isolates labelled W through Z in Fig. S5. These results therefore mirror the discrepant clustering of these *S. sonnei* isolates observed in the MS trees.

Comparison of the subtyping methods' discriminatory power

Visual comparison of the MS trees also clearly indicates an increase in discriminatory power due to the reduced size of the nodes notably observed for cgMLST and wgMLST, increasing the number of distinct STs. To assess the capacity of each subtyping method to differentiate between the *Shigella* isolates, the discriminatory powers were measured using Simpson's DI on the STs generated. This index calculates the average probability that a typing system will assign a different type to two randomly selected isolates in the collection under study. Table 1 indicates this measure along with ST data and *P*-values between confidence intervals (CIs).

As expected, the discriminatory power followed suit with the number of loci detected from the MLST-based



(a)

Methods	Equation	R^2
coreMLST	y = 0.7425x - 0.9490	0.9112
cgMLST	y = 1.0707x - 0.3837	0.9177
wgMLST	y = 1.3467x + 1.6389	0.9021

(b)

Methods	Equation	R^2
coreMLST	y = 0.5844x + 0.3957	0.9150
cgMLST	y = 0.7906x + 2.6407	0.8823
wgMLST	y = 1.7817x + 3.3975	0.4627

Correlation coefficients

	SNVPhyl	coreMLST	cgMLST
coreMLST	0.955		
cgMLST	0.958	0.982	
wgMLST	0.950	0.969	0.985

Correlation coefficients

	SNVPhyl	coreMLST	cgMLST
coreMLST	0.957		
cgMLST	0.939	0.961	
wgMLST	0.680	0.703	0.712

Fig. 1. Scatterplot of all outbreak-related pairwise differences of coreMLST, cgMLST and wgMLST against SNVPhyl for *S. flexneri* (a) and *S. sonnei* (b).

subtyping methods as they were arranged as follows: coreMLST <cgMLST<wgMLST for both *Shigella* spp. datasets. However, the number of hqSNVs used to generate phylogeny by SNVPhyl differed greatly between *Shigella* species (2717 and 1072 hqSNVs), but stayed as the third most discriminatory based on the Jackknife pseudo-values 95% CIs. Significant differences in discriminatory power are indicated in green, where the *P*-values are <0.05 and the CIs are not overlapping. However, this significance is not meaningful since all subtyping methods remained within a practical DI range of 97–100%, which makes them all highly discriminatory.

Inter-outbreak genetic distances and relatedness to sporadic cases

All outbreaks were easily distinguished by all four WGS-based subtyping methods. The minimum inter-outbreak distances separating the distinct *S. flexneri* and *S. sonnei* outbreaks were 215 hqSNVs/143 core alleles/198 core genome (CG) alleles/211 WG alleles and 42 hqSNVs/30 core alleles/40 CG alleles/51 WG alleles, respectively. Not many sporadic isolates clustered closely (<100 WG alleles) to the outbreaks under study. The minimum distances between the *S. flexneri* and *S. sonnei* outbreak-related isolates to any isolate considered sporadic were 44 hqSNVs/32 core alleles/41 CG alleles/70 WG alleles and 31 hqSNVs/20 core alleles/27 CG alleles/36 WG alleles, respectively. NDCs not considered as outbreak-related were separated by at least 11 hqSNVs/5 core alleles/9 CG alleles/11 WG alleles and 3 hqSNVs/1 core allele/5 CG alleles/6 WG alleles from the *S. flexneri* and *S. sonnei* outbreaks under study, respectively. Additional epidemiological evidence would have been needed to support their inclusion as outbreak-related cases or exclusion as sporadic cases.

Intra-outbreak/subcluster diversity and WGS-based interpretation guidelines

Table 2 summarizes the characteristics of the outbreaks and subclusters under study along with their intra-outbreak distance ranges and largest minimum spanning distances. The

Table 1. Summary of the SNVPhyl, coreMLST, cgMLST and wgMLST subtyping results, diversity indexes and statistical significances for the complete 91 *S. flexneri* and 232 *S. sonnei* isolate collections

Species	Method	No. of hqSNV locations/loci*	No. of ST	Size of largest ST	Simpson's DI	Jackknife pseudo-values 95% CIs	*P*-values between jackknife pseudo-values 95% CIs of Simpson's DIs†		
							SNVPhyl	coreMLST	cgMLST
S. flexneri (*n*=91)	SNVPhyl	2717	76	5	0.995	(0.990–0.999)			
	coreMLST	2511	63	5	0.987	(0.980–0.994)	0.039		
	cgMLST	3398	86	2	0.999	(0.997–1.000)	0.060	<0.001	
	wgMLST	4051	89	2	1.000	(0.999–1.000)	0.036	<0.001	0.213
S. sonnei (*n*=232)	SNVPhyl	1072	128	35	0.970	(0.957–0.984)			
	coreMLST	2511	110	28	0.970	(0.960–0.980)	0.976		
	cgMLST	3613	196	4	0.998	(0.997–0.999)	<0.001	<0.001	
	wgMLST	4426	209	4	0.999	(0.998–1.000)	<0.001	<0.001	0.013

*Number of sites used to generate phylogeny by SNVPhyl. Number of alleles included in the exported categorical data matrixes from BioNumerics.

†Significant differences in discriminatory power are indicated in green (*P*-value <0.05 and non-overlapping CIs).

CG, core genome; CI, confidence interval; DI, diversity index; hqSNV, high-quality single-nucleotide variant; MLST, multi-locus sequence typing; SNVPhyl, single-nucleotide phylogenomics pipeline; ST, sequence type; WG, whole genome.

intra-outbreak/subcluster distances indicate the range of pairwise genetic differences (hqSNVs or allele variants) observed between outbreak- and subcluster-related isolates, while the superior limits give a sense of the overall diversity observed within the outbreaks/subclusters. The largest minimum spanning distance illustrates the largest distance by which a single isolate differs from its neighbouring outbreak- or subcluster-related isolates according to the minimum spanning trees. For example, the three SS-05c-related isolates form a distinct subcluster in which they differ by *at least* 9 hqSNVs/6 core alleles/10 CG alleles/19 WG alleles from the other 117 isolates included in the SS-05 outbreak. This distance metric was included in this study in order to test an alternative interpretation criterion independent of the transmission nature, effective population size and duration of an outbreak. This was important since, naturally, prolonged outbreaks have more diversity caused by the constant accumulation of mutations in the outbreak strain, creating larger maximum intra-outbreak genetic distances, which has frequently been observed in the MSM community. Using a minimum spanning distance as a measure of relatedness would avoid such large distances between the first and last outbreak-related isolates and thus create a more stable guideline for public health decisions, assuming that the genetic variation correlates with the time of isolation from the index case and that a constant rate of cases is reported.

In the four non-MSM outbreaks studied, the maximum genetic diversity ranges from 0 to 7 hqSNVs/0–3 core alleles/3–7 CG alleles/4–8 WG alleles, which was consistent with the established 0–10 hqSNV/allele interpretation criteria measured for other single-strain outbreaks of foodborne enteric pathogens. In the MSM-related SF-01 outbreak, the genetic diversity ranges from 0 to 37 hqSNVs/0–30 core alleles/1–41 CG alleles/1–53 WG alleles, while in the other three large MSM-related outbreaks studied, the maximum genetic diversity ranges from 16 to 23 hqSNVs/9–16 core alleles/14–24 CG alleles/26–117 WG alleles. Notably, the diversity detected in large MSM-related outbreaks (SF-02, SS-03 and SS-05) by wgMLST is considerably larger than the other WGS-based subtyping methods. In particular, wgMLST added 93 alleles to the 24 allele diversity detected by cgMLST for the SS-05 outbreak. However, the diversity observed for the prolonged MSM-related SF-03 outbreak was inferior (6 hqSNVs/3 core alleles/6 CG alleles/8 WG alleles) as it only included seven cases. The maximum genetic diversity for the delimited subclusters ranges from 1 to 10 hqSNVs/1–9 core alleles/3–16 CG alleles/5–75 WG alleles.

The large genetic differences observed between MSM-related outbreak strains through wgMLST will make creating real-time interpretation guidelines problematic for routine surveillance of *Shigella* spp. If foodborne outbreak-validated criteria were applied (i.e. 10 WG allele threshold) to determine relatedness and epidemiological concordance, all MSM-related outbreaks involving more than seven cases would have had epidemiologically related isolates excluded solely on this basis.

Table 2. Features, intra-outbreak/subcluster genetic distances and largest minimum spanning distances determined by SNVPhyl, coreMLST, cgMLST and wgMLST of the four *S. flexneri* and five *S. sonnei* outbreaks under study in addition to the six *S. flexneri* and six *S. sonnei* subclusters

Species	Outbreaks/ Subclusters		Year	Serotypes	No. of sequenced isolates	Duration (days)*	Sex (M/F)	Age range	Epidemiology (frequency)	Intra-outbreak distance ranges†				Largest minimum spanning distances‡			
										SNVPhyl	coreMLST	cgMLST	wgMLST	SNVPhyl	coreMLST	cgMLST	wgMLST
S. flexneri	🟥	SF-01	2017–2019	2a and Y	31	1012	31/0	23–86	MSM (14)	0–37 (17)	0–30 (13)	1–41 (19)	1–53 (28)	23	16	23	26
	🟫	SF-01a	2018–2019	2a	4	289	4/0	23–67	MSM (4)+Travel (1)	1–6 (3.5)	0–2 (1)	1–5 (3)	1–8 (4)	4	1	3	5
	🟥	SF-01b	2019	2a	6	257	6/0	23–86	MSM (2)	0–4 (2)	0–4 (1)	1–5 (3)	2–7 (4)	3	3	3	3
	🟪	SF-01c	2019	2a	5	69	5/0	36–59	MSM (3)	0–3 (1.5)	0–2 (1)	2–6 (3)	2–8 (4.5)	2	1	3	4
	🟧	SF-01d	2018–2019	Y	7	322	7/0	29–78	MSM (1)+Travel (1)	0–10 (5)	0–9 (4)	2–16 (8)	5–18 (13)	5	4	8	8
	🟩	SF-02	2018–2019	1b and 3b	42	529	40/2	24–72	MSM (23)+Travel (1)	0–17 (6)	0–10 (3)	0–14 (6)	0–26 (9)	8	4	6	11
	🟩	SF-02a	2018–2019	1b and 3b	15	510	14/1	36–68	MSM (9)	0–10 (4)	0–4 (2)	0–8 (4)	1–12 (6)	3	2	3§	5§
	🟩	SF-02b	2019	1b	3	76	2/1	38–46	MSM (2)	0–1 (1)	0–1 (1)	0–3 (2)	2–5 (2)	1	1	2	2
	🟦	SF-03	2018–2019	2a	7	226	7/0	39–71	MSM (6)+Travel (1)	0–6 (3)	0–3 (2)	1–6 (4)	2–8 (5)	3	2	4	4
	🟨	SF-04	2018	1b	5	25	2/3	4–37	Food (3)+Familial (1)	0–2 (1)	0 (0)	1–6 (3)	1–7 (3)	1	0	3	3
S. sonnei	🟧	SS-01	2013	–	4	5	0/4	24–54	Non-MSM (4)	0 (0)	0–1 (0.5)	1–3 (2)	1–4 (2)	0	1	2	2
	🟦	SS-02	2014	–	3	22	0/3	46–53	Travel (3)	1–7 (6)	1–3 (1)	2–7 (4)	2–8 (5)	6	1	4	4
	🟪	SS-03	2017–2019	–	57	797	54/3	21–75	MSM (29)+Travel (1)	0–16 (5)	0–9 (3)	0–16 (6)	0–34 (8)	6	4	6	16
	🟪	SS-03a	2017–2018	–	16	155	16/0	23–75	MSM (12)+Travel (1)	0–3 (1)	0–2 (1)	0–6 (2)	0–8 (4)	1	1	4§	4§
	🟪	SS-03b	2018–2019	–	21	603	20/1	22–70	MSM (13)	0–4 (2)	0–5 (1)	0–10 (4)	0–29 (6)	2	4	4	16
	🟨	SS-04	2018	–	7	33	3/4	1–76	Food (7)	0–1 (0)	0–1 (0)	0–3 (1)	0–5 (2)	1	1	1	2
	🟩	SS-05	2018–2019	–	120	407	116/4	1–71	MSM (28)+Travel (2)	0–23 (8)	0–16 (6)	0–24 (10)	0–117 (15)	9	8	10	19
	🟩	SS-05a	2018–2019	–	69	396	67/2	24–70	MSM (19)+Travel (1)	0–7 (1)	0–5 (2)	0–10 (4)	0–57 (5)	3	2	5	7
	🟩	SS-05b	2018–2019	–	36	351	36/0	21–71	MSM (8)	0–8 (2)	0–6 (1)	0–10 (4)	0–75 (6)	3	3	4	31§
	🟩	SS-05c	2019	–	3	281	3/0	28–47	Unknown	0–4 (4)	0–3 (3)	2–6 (5)	2–7 (6)	1	3	5	5
	🟩	SS-05d	2019	–	4	29	2/2	1–41	Unknown	0–1 (0.5)	0–1 (1)	2–4 (3)	2–5 (4)	4	1	3	4

*Total number of days between the isolate dates of the first and last outbreak/subcluster-related isolates.

†Minimum and maximum pairwise differences between outbreak/subcluster-related isolates. Median pairwise genetic differences are in parentheses.

‡Largest genetic distance between neighbouring outbreak/subcluster-related isolates according to the minimum spanning trees (*n* differences).

§All isolates did not group together into a single minimum spanning cluster. The distance reported is the largest sum of minimum spanning distances between subcluster-related isolates.

CG, core genome; MLST, multi-locus sequence typing; MSM, men who have sex with men; SNVPhyl, single-nucleotide variant phylogenomics pipeline; WG, whole genome.

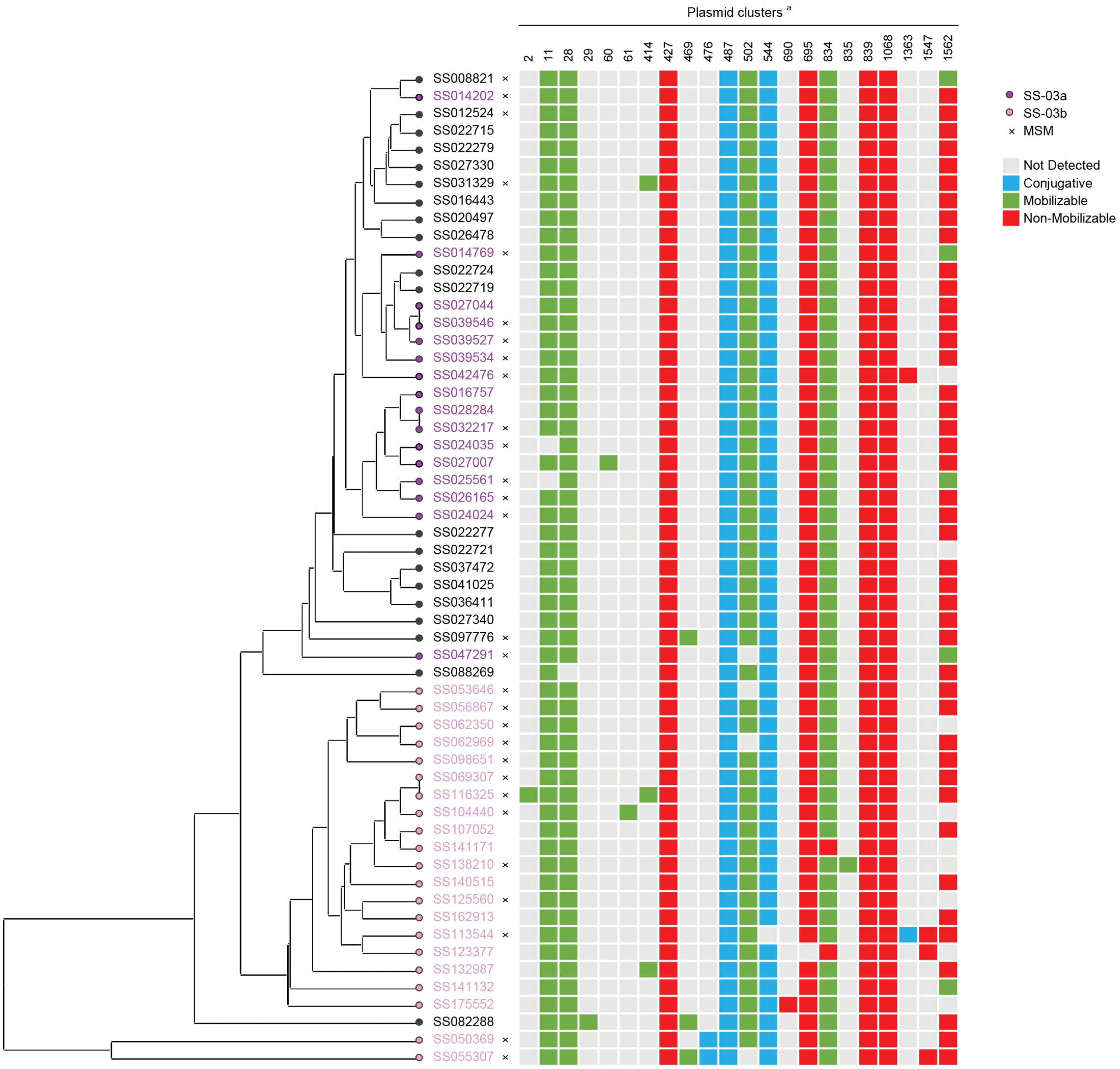


Fig. 2. WgMLST ML phylogenetic tree and plasmid profiles of the 57 isolates involved in the SS-03 outbreak. [a]Plasmid clusters are defined as a group of closed reference plasmids with high sequence similarity by MOB-cluster. Contigs of plasmid origin in the draft assemblies of the isolates under study were then assigned a plasmid cluster.

Alternatively, the largest minimum spanning distances observed for all 9 *S. flexneri* and 11 *S. sonnei* outbreaks/subclusters (excluding SF-01) under study were ≤8 hqSNVs/4 core alleles/8 CG alleles/11 WG alleles, and ≤9 hqSNVs/8 core alleles/10 CG alleles/35 WG alleles, respectively. Here, it is demonstrated that if largest minimum spanning distances were used instead of maximum intra-outbreak distances for outbreak delimitations, all WGS-based subtyping methods would have been able to characterize most outbreaks correctly based on the established 0–10 hqSNV/allele interpretation criteria, with the exception of wgMLST, where discrepant clustering was observed for several subclusters (e.g. SS-03b and SS-05b). However, by eliminating these isolates from the analysis, the largest minimum spanning distances for SS-03, SS-03b and SS-05b fall to 8, 4 and 12 WG alleles. In addition, if we separate the SS-05d subcluster from the SS-05 outbreak altogether due to differences in epidemiology, its largest minimum spanning distance falls to 12 from 19 WG alleles.

MGE-mediated phylogenetic discrepancies and inflated genetic distances

The phylogenetic discrepancies discussed in previous sections and illustrated in Fig. S5 were further investigated. A clear example is also illustrated in Fig. 2, where the epidemiologically supported isolates SS050369 (Y) and SS055307 (Z) clustered on separate branches from the rest of the SS-03 outbreak via wgMLST on the ML tree. A distance of 19 WG alleles was found between both isolates and a combined average distance of 26 (range, 18–35) WG alleles was found to separate them from the other 55 outbreak-related isolates. However, these isolates cluster concordantly with the other SS-03b subcluster through all core genome-based subtyping methods with a distance of only 1 hqSNV/2 core alleles/3 CG alleles between them and an average distance of 5 (range, 0–14) hqSNVs/3 (range, 0–6) core alleles/7 (range, 2–13) CG alleles to the other 55 SS-03 outbreak-related isolates.

The cause of the inflated genetic distances was thus investigated by analysing the allele variants used to generate the wgMLST phylogeny for these two isolates along with a temporally linked SS-03b-related isolate (SS053646). Isolates SS055307 and SS050369 differed by 0 to 1 hqSNVs/0–2 core alleles/3–5 CG alleles/22–26 WG alleles from SS053646. The combined WG allele differences of the three pairwise comparisons revealed that wgMLST detected differences in an additional 25 loci compared to cgMLST. MOB-recon determined that 24 out of the 25 allele differences were encoded on contigs associated to the plasmid clusters c476 or c487. The plasmid cluster profiles of the 57 outbreak-related isolates can be found in Fig. 2 alongside the wgMLST ML tree. Eleven out of the 22 different plasmid clusters were present in over 80% of the SS-03 outbreak-related isolates. The other 11 plasmid clusters were present sporadically in either 1 to 3 isolates. The plasmid profile of this outbreak strain was thus relatively stable through time. Two conjugative plasmid clusters, c487 and c544, were present in ≥98% of the isolates. Interestingly, the plasmid cluster c476 was also conjugative, but was only present in the discrepantly clustered isolates SS050369 (Y) and SS055307 (Z). The additional allelic differences observed for SS050369 and SS055307 from the other SS-03 isolates were thus caused by the acquisition of the conjugative plasmid c476, which generated multiple differences from homologous genes already present in the ubiquitous plasmid c487. However, the large diversity between both isolates was caused by the acquisition of a similar 'type' of plasmid according to MOB-typer (c476), while they seemed to have gone through different evolutionary paths, causing multiple allelic differences between them as well.

Locations of the hqSNVs and allele variants used to differentiate outbreak- and subcluster-related isolates

Following the results of the previous section, all outbreaks and subclusters were evaluated based on the location of the hqSNVs and allele variants used to generate phylogeny by the four WGS-based subtyping approaches. Fig. 3 illustrates the number of hqSNVs and allele differences according to these locations for all outbreaks and subclusters under study.

Most genetic differences were due to mutations in the chromosome for all four WGS-based subtyping methods, with the exception of the MSM-related outbreak SS-05 and subclusters SF-01c, SF-02b, SS-03b, SS-05a and SS-05b. The enhanced discrimination of wgMLST over cgMLST is largely due to allele differences mapped to plasmids and prophages. Most plasmid differences are caused by 'plasmid copies', where the same loci could be found on different plasmid sequences throughout an outbreak/subcluster and their corresponding alleles were sometimes discerned by upwards of several hundred mutations. Most of these loci were associated with plasmid maintenance, mobilization and conjugation. All outbreak- and subcluster-related isolates were also further distinguished by IS elements. However, due to the abundance and repeated nature of IS elements in the *Shigella* spp. genomes, most differences here would be considered false positives. In the 4 non-MSM outbreaks under study, most (15/19) genetic differences added by cgMLST and wgMLST over coreMLST are also considered false positives (e.g. chromosomal repeats and IS elements). The genetic differences due to 'plasmid copies' and 'prophages' were exceptionally high for the *S. sonnei* outbreak SS-05 and its subclusters SS-05a and SS-05b. Similarly to the SS-03 outbreak's phylogenetic discrepancies illustrated in Fig. 2, two SS-05a isolates, SS131423 (W) and SS121129 (X) illustrated in Fig. S5, seem to have acquired distinct conjugative c476 plasmids encoding multiple loci already encoded on the omnipresent conjugative plasmid c487, creating many allelic variations between each other and the rest of the SS-05a subcluster-related isolates. In the subcluster SS-05b, similar observations were found as 9 isolates who seem to have acquired several new conjugative plasmids are responsible for 93/95 'plasmid copies' differences. In addition, 4 SS-05b isolates seem to have acquired the *Salmonella* phage SSU5 (NC_018843) according to PHASTER, but 1 isolate seems to have acquired a variant, creating all 46 allele differences accorded to 'prophages'.

Table S2 lists the problematic loci (e.g. chromosome repeats, plasmid-associated loci, prophage-associated loci and IS element transposase genes) included in the MLST-based loci schemes. The inclusion of highly repeated genes, such as IS element transposases, might have also given a deceptive appearance of higher discriminatory power measured previously for cgMLST and wgMLST. In support of this, the loci ECOLI03837 encoding the fimbria biosynthesis transcriptional regulator *fimZ* had the allele calls 422 and 423 randomly assigned to the complete *S. sonnei* isolate collection due to different contig lengths caused by an IS element (IS2) truncation. However, while the *fimZ* gene is known as an *S. flexneri* serotype 2a strain 2457T-specific pseudogene [56], it was not detected by BioNumerics in the isolate collection for *S. flexneri*.

Robustness and the genetic distance variance of different input datasets

During surveillance, public health personnel evaluate the relatedness of isolates occurring within a determined time frame in order to detect potential outbreak clusters. In this time frame, the rate of unrelated isolates to the number of outbreak-associated cases can vary greatly depending on various factors (e.g. time of the year, region, etc.). The robustness of each WGS-based subtyping method was thus evaluated because of the need of an approach not easily influenced by the diversity of the input dataset under comparison, thus allowing interlaboratory comparisons.

Based on our previous results, the core genome-based subtyping methods (e.g. cgSNV, coreMLST and cgMLST)

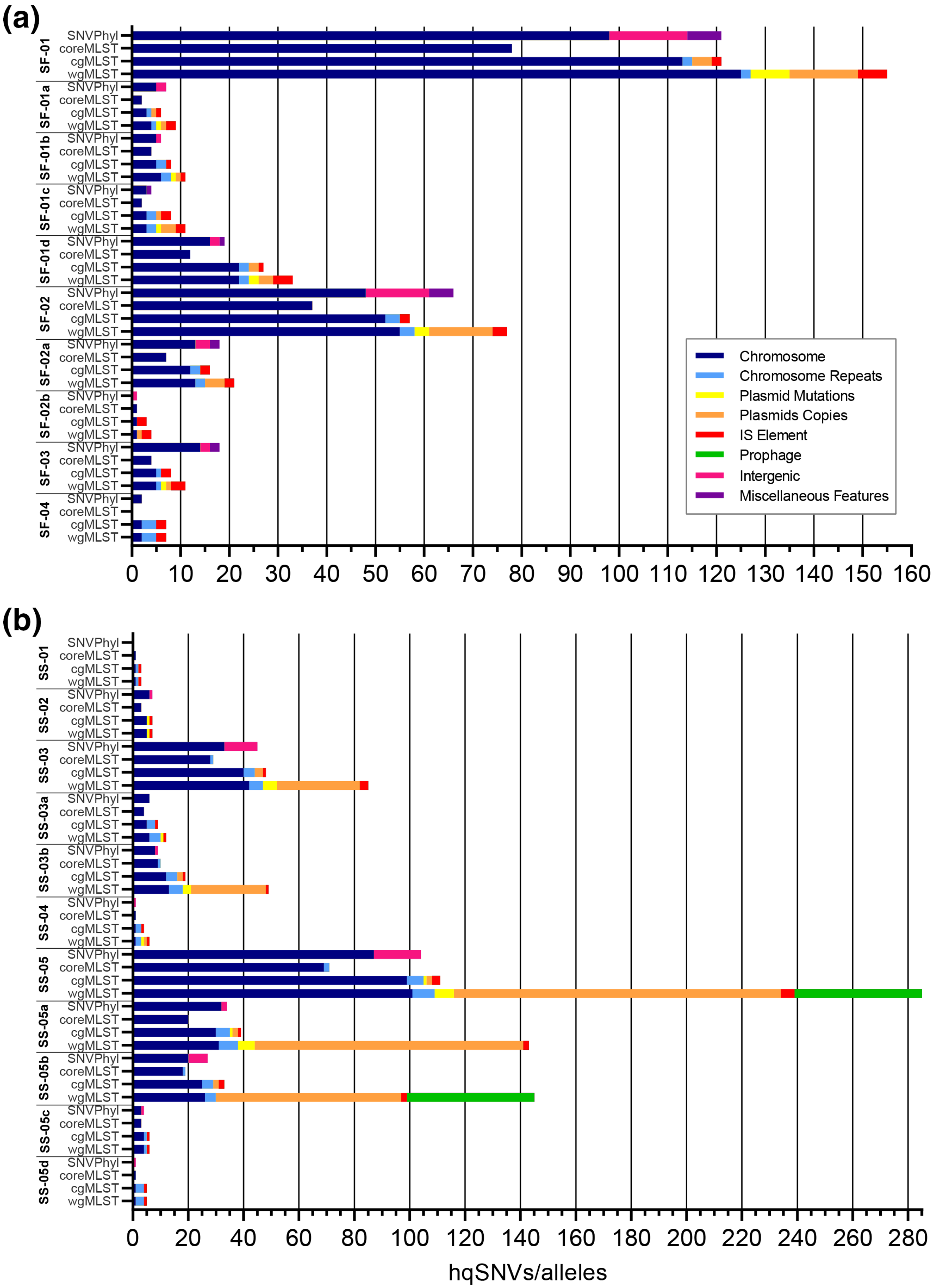


Fig. 3. Summary of the reference-based hqSNV locations and BioNumerics loci used to generate phylogeny via SNVPhyl, coreMLST, cgMLST and wgMLST for all 10 *S. flexneri* (a) and 11 *S. sonnei* (b) outbreaks and subclusters.

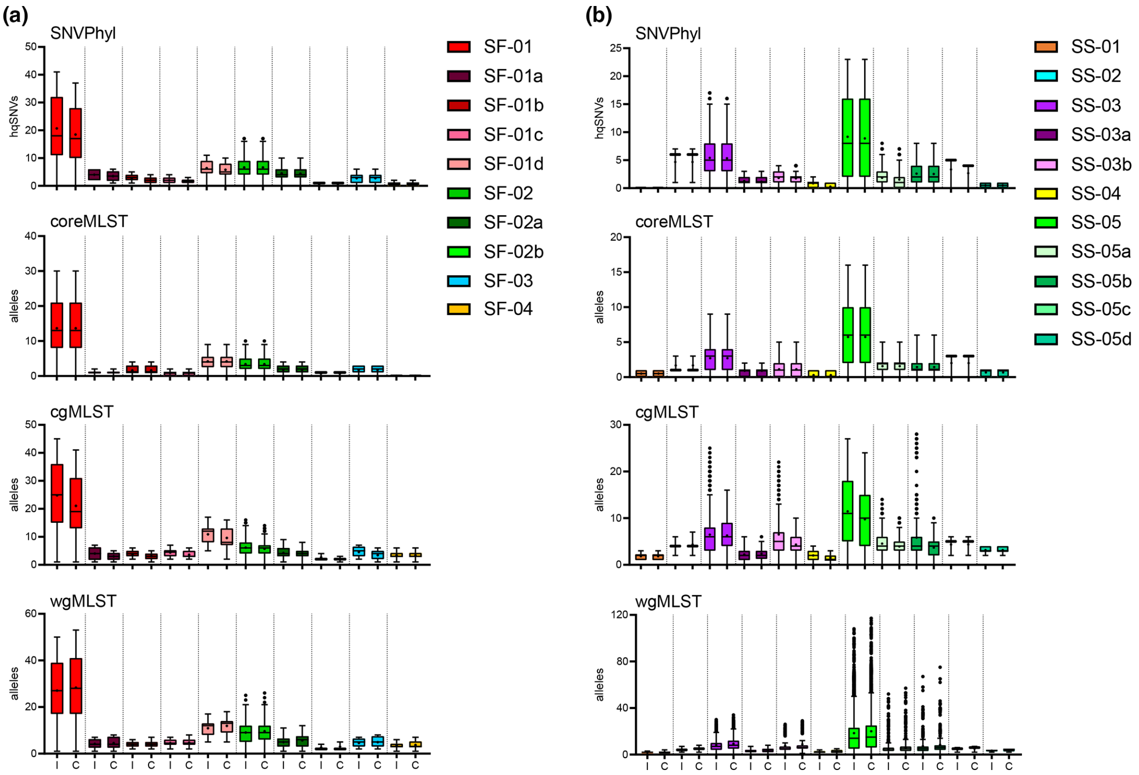


Fig. 4. Box and Tukey whisker plots of the pairwise genetic distances for all 10 *S. flexneri* (a) and 11 *S. sonnei* (b) outbreaks and subclusters via the use of isolated (I) outbreak/subcluster-specific collections and the complete (C) species-specific isolate collections as inputs to SNVPhyl, coreMLST, cgMLST and wgMLST. The black dots inside the boxes represent the mean genetic distances.

seem more reliable when investigating *Shigella* spp. outbreaks, as they eliminate most confounding MGEs (e.g. plasmids, insertion sequence elements, prophages). However, the results generated by core genome-based approaches, such as SNVPhyl and cgMLST, are inherently dependent on the collection of isolates under comparison [24]. The addition of distantly related isolates (as is the case when comparing various outbreaks and sporadic isolates together) will inevitably reduce the discriminatory power of these methods, in addition to reducing the genetic variation of the outbreaks in question by decreasing the size of the shared genome. Fig. 4 summarizes these variabilities by comparing the outbreak/subcluster diversities found earlier to the diversities generated when using each outbreak/subcluster as distinct inputs through box plots for *S. flexneri* and *S. sonnei*.

Minor increases in distance metrics were observed for isolated outbreaks via SNVPhyl for *S. flexneri* and *S. sonnei*. The coreMLST method proved to be the most robust, as no changes were detected for either species. The cgMLST method, as expected, generated the most notable changes in distance metrics, but MSM-related outbreaks of *S. sonnei* were the most affected. The distance metrics of isolated outbreaks via cgMLST increased the most, since the increased clonality observed in the inputs increased the core genome substantially so as to resemble the results generated by wgMLST. However, wgMLST had an inverse effect, as isolated outbreaks produced minor *decreases* in distances.

The phylogenies created by SNVPhyl, coreMLST and wgMLST were largely congruent, with the exception of changes in terminal branch lengths as a result of the differences in distances discussed above. However, the cgMLST tree topologies for certain isolated MSM-related outbreaks of *S. sonnei* resembled the discrepant topologies generated by wgMLST previously discussed in this paper. Therefore, cgMLST should not be used for sole outbreak investigations, since accessory genomes are likely stable throughout an outbreak and can cause epidemiologically discordant phylogenies similar to the use of wgMLST.

S. flexneri serotype conversions and impact on outbreak delimitations

In the collection of 91 *S. flexneri* isolates, 4 serotypes were included: 1b (*n*=47), 2a (*n*=34), Y (*n*=8) and 3b (*n*=2). Not all serotypes were monophyletic. Serotype conversion events were observed for outbreaks SF-01 and SF-02. The *gtrI* sequence from the GenBank accession number AF139596.1 [57] and the SfII bacteriophage sequence from the GenBank accession number AF021347.1 [58] were used as queries against the *S. flexneri* 1b/3b and *S. flexneri* 2a/Y isolates, respectively, in order to predict gene presence and mutations via local alignments.

Forty SF-02 isolates presented serotype 1b, while two isolates (SF118225 and SF164148) presented the 3b serotype, but clearly clustered within the same outbreak clade with a minimum distance of 1 and 2 hqSNVs/2 and 0 core alleles/3 and 2 CG alleles/6 and 3 WG alleles to a serotype 1b isolate. The only notifiable difference by agglutination between a 3b and a 1b serotype for *S. flexneri* is the addition of a glucosyl group on the N-acetylglucosamine unit of the O-antigen, which is mediated by the *gtrI* gene product [59]. A comparative DNA analysis of the 42 SF-02 outbreak-related isolates' *gtrI* gene with a reference sequence (AF139596.1) [57] resulted in a shared single-nucleotide deletion (−1A) at position 1175 that rendered a frameshift for both serotype 3b-presenting isolates. The other 40 isolates did not present any mutations, with the exception of 2 isolates (SF195966 and SF197491) with the last 26 and 2001 nucleotides of the *gtrI* gene deleted, respectively.

Additionally, the SF-01 outbreak included a Y serotype subcluster (SF-01d, *n*=7) among the other SF-01-related subclusters and isolates (*n*=24) presenting serotype 2a. The SF-01d subcluster was distanced by a minimum of 19 hqSNVs/14 core alleles/21 CG alleles/29 WG alleles to the other SF-01 isolates. The only detectable difference by agglutination between a Y serotype and a 2a serotype is the presence of a glucosyl group on the rhamnose III position, which is possible by the *gtrII* gene product, a serotype-specific glucosyltransferase. However, the *gtrII* gene is known to be encoded by the serotype-converting bacteriophage SfII, where the loss of this sequence is commonly known to cause these serotype conversions in addition to amino acid substitution events [58, 60]. Here, the 7 isolates presenting the Y serotype did not possess both sugar transferase genes (*bgt* and *gtrII*) regularly encoded by the SFII bacteriophage (AF021347), while the other 24 serotype 2a isolates possessed both genes.

DISCUSSION

The general goal of this study was to evaluate the WGS-based subtyping methods readily accessible by Canadian public health laboratories for the surveillance of *Shigella* spp. While other foodborne pathogens under the surveillance of PNC had been fully validated for wgMLST, there was a void for the interpretation of prolonged transmission patterns regularly observed in *S. flexneri* and *S. sonnei* outbreaks involving MSM.

In this study, several components from each subtyping method were therefore compared, including the methods' compatibility and discriminatory power. The four WGS-based subtyping methods tested were generally considered concordant due to high correlation coefficients (0.939–0.985), but wgMLST was not deemed compatible in the *S. sonnei* dataset (R^2=0.4627) due to its heightened ability to detect variation in the whole genome, which resulted in some abnormally large genetic distances between outbreak-related isolates (e.g. isolates W through Z). Similar to our results, previous studies on other bacterial species have shown that SNV-based and coreMLST/cgMLST methods were congruent [25, 61–63] and that correlation coefficients were greater for cgMLST than wgMLST [63]. However, an SNV-based subtyping approach was shown to be highly compatible with wgMLST for foodborne *L. monocytogenes* outbreaks (R^2=0.96) [42, 64]. Based on genetic distances alone, wgMLST also seemed compatible with core genome-based subtyping approaches when tested on other enteric pathogens [23, 24]. This difference in compatibility emphasizes the differences in environmental pressures and genomic plasticity between long-term *S. sonnei* clusters and the clonal pathogen outbreaks regularly studied (e.g. foodborne).

All WGS-based subtyping methods studied were highly discriminatory for both *Shigella* spp. datasets, but cgMLST and wgMLST were found to be the most discriminatory, as they were able to assign different sequence types to most isolates (Simpson's DI, 95% CIs: 0.997–1.000). In comparison to other enteric bacteria, Pearce *et al.* [25] found similar results, as their core genome SNV-based approach provided greater resolution than coreMLST for *S. enterica* serovar Enteritidis. However, Vincent *et al.* [23] found that SNVPhyl was the most discriminatory for distinguishing outbreak isolates of *S. enterica* serovar Heidelberg from Quebec in comparison to cgMLST and wgMLST. Jackson *et al.* [23] also found that a hqSNV analysis produced more differences among outbreak-related isolates compared to wgMLST for foodborne *L. monocytogenes* outbreaks, indicating a higher discriminatory power for SNV-based analyses, although Henri *et al.* [64] did not find any difference. These results showcase the limitation of MLST-based subtyping as they consider mutations, recombinations and indels as single evolutionary events. Therefore, a single allelic shift can comprise multiple genetic changes, which would otherwise be detected by an SNV-based subtyping approach. On the other hand, our results were similar to the findings of Rumore *et al.* [24], as wgMLST was the most discriminatory compared to SNVPhyl for verotoxigenic *E. coli* O157:H7. This concordance is indicative of a more dynamic and diverse *E. coli/Shigella* accessory genome, where their shared *E. coli/Shigella* wgMLST scheme may have measurable advantages when it comes to discriminatory power compared to the wgMLST schemes constructed for other enteric pathogens. This advantage is most likely due to the larger size of the *E. coli/Shigella* pan-genome, enabling the detection of a wide range of loci, including many involved with mobile genetic elements (e.g. plasmids, prophages and transposons).

These concordances and discriminatory powers were also translated in the intra-outbreak genetic distances observed. In the four non-MSM outbreaks studied, the maximum genetic diversity ranged from 0 to 7 hqSNVs/0–3 core alleles/3–7 CG alleles/4–8 WG alleles, which was consistent with documented non-MSM *S. flexneri* (≤5 hqSNVs) [65] and *S. sonnei* (≤7 hqSNVs) outbreaks [66]. This concordance demonstrates the shared limited transmission of the pathogen, thus resembling point source contaminations. However, our non-MSM outbreaks only included three to seven cases, which limits this observed consistency as non-MSM outbreaks involving a larger population may not always follow the same 0–10 hqSNV/allele guideline.

In the four MSM-related outbreaks studied (excluding the complete SF-01 outbreak), the maximum genetic diversity ranged from 1 to 23 hqSNVs/0–16 core alleles/3–24 CG alleles/5–117 WG alleles. Here, the SF-01 outbreak was excluded from the distance ranges described, since it seems to represent a genetically diverse reservoir population (or an endemic strain) where multiple outbreak events (i.e. subclusters a-d) and isolated cases seem to have emerged. Therefore, it was probably wrong of us to encompass every subcluster together as a single-strain outbreak due to its high diversity, as observed by the core genome-based approaches (0–37 hqSNVs/0–30 core alleles/1–41 CG alleles/1–53 WG alleles) and superior largest minimum spanning distances (23 hqSNVs/16 core alleles/23 CG alleles/26 WG alleles). This highlights the importance of gathering detailed epidemiological data to supplement WGS-based subtyping methods. While all SF-01 subclusters occurred in the same time frame and involved MSM cases, their differences in phylogeny may indicate deeper causes in epidemiology other than simply 'MSM'. Other subclusters were also defined in order to detect potential transmission routes and spillover events into other communities [67]. For example, the SS-05d subcluster involved women and children with unknown epidemiology, while the complete SS-05 outbreak would be considered MSM-related.

All MSM-related outbreaks surpassed the 10 genetic difference guidelines for SNVPhyl, cgMLST and wgMLST, with the exception of the SF-03 outbreak (*n*=7 isolates). The other 3 MSM-related outbreaks (SF-02, SS-03 and SS-05), which contained from 42 to 120 cases, have gained a substantial amount of diversity going from cgMLST to wgMLST, usually doubling the distances observed. Notably, for outbreak SS-05, it went from a maximum of 24 CG alleles to 117 WG alleles. The hqSNV diversities observed here seem concordant to the 64 SNP difference observed for an MSM-related *S. flexneri* outbreak involving 121 cases [65] when assuming that genetic diversity correlates with the size of the effective population of prolonged outbreak events. However, while defining the maximum number of WG allele differences required to classify isolates as related/unrelated to an outbreak strain may be a desirable course of action for outbreak delimitations, it is highly unlikely that a universal distance threshold will be able to consistently predict epidemiological relationships [24, 26, 68–70]. Some authors even predict misleading interpretations of transmission networks due to the inadequacy of only sequencing a single isolate per case, thus ignoring the within-host diversity of the pathogen population [34]. This heterogeneity between epidemiologically linked cases highlights the importance of including tree topologies as indicators of evolution in addition to epidemiological evidence and traceback data in combination with genetic distances to define *Shigella* spp. clusters.

Nevertheless, we have shown that the use of minimum spanning distances could be useful as an alternative measure of shared epidemiology for prolonged outbreak events, since all outbreak- and subcluster-related isolates did not surpass 10 hqSNV/allele differences with their neighbouring isolates through SNVPhyl, coreMLST and cgMLST. Most wgMLST minimum spanning distances also respected this threshold, with the exception of several instances, where the clustering was considered discordant or where the distances were measured outside the delimited subclusters. Further studies are thus warranted due to its flexibility to create stable WGS interpretation guidelines. However, a major weakness of this study was the limited epidemiological evidence available and the assumptions made on outbreak/subcluster delimitations based on the SNVPhyl phylogeny. Therefore, some outbreaks/subclusters analysed here might have included some isolates with no true shared epidemiology and vice versa.

The inflated distances and discordant clustering observed with the use of the wgMLST approach were investigated through a comparison of the predicted plasmid profiles detected in the SS-03 outbreak. We have shown that homologous genes encoded on co-inhabiting plasmids can cause multiple allelic differences only detected by the wgMLST method. However, these discrepancies are partly due to the default BioNumerics settings of only tabulating the allele call with the lowest allele identification number when multiple allele calls are present for a single locus. An easy setting change in BioNumerics could be of interest, as it could eliminate the loci with multiple allele calls from the analysis and remove the effect of repeated loci on the genetic distances measured. It is also important to mention that while MOB-recon has shown high sensitivity and specificity for plasmid detection, contigs of plasmid origin detected through short-read sequencing can be split across multiple clusters and/or merged into clusters with other non-related contigs [55]. In addition, repetitive elements with multiple copies will only be assigned to one plasmid cluster code via MOB-recon's winner-takes-all strategy [55]. The cluster codes mentioned in this paper are therefore used for descriptive purposes only and further studies are recommended in order to better understand the plasmid diversity observed in the MSM-related outbreaks, as plasmids can be rapid disseminators of antimicrobial resistance traits [71].

The drawbacks of the use of wgMLST were also illustrated in our analysis of the locations of the hqSNV and allele variants used by the different subtyping approaches to determine sequence types. While subtyping methods should approximate epidemiology, the inclusion of MGE-encoded loci in the cg/wgMLST schema was shown to reduce this reliability via the presence of plasmid-, prophage- and IS element-related loci. Although the scientific community had expressed their concern regarding including pseudogenes and paralogous genes within a dataset used for surveillance purposes and estimated that such datasets could generate spurious genetic differences leading to inaccurate clustering [21, 25, 29, 63], most outbreaks studied did not present such erroneous results. However, some homologous recombinations, plasmid acquisitions and prophage sequence variations have been documented to affect phylogeny in prior outbreak investigations [41, 63, 72]. The emergence of such false diversity in our *S. sonnei* MSM-related outbreaks is thought to coincide with *S. sonnei*'s competitive edge over *S. flexneri* to accept and maintain with more stability horizontally transferred DNA [39, 40]. In addition, prolonged transmission patterns providing greater opportunities for horizontal gene transfers [65] and the environmental pressures of antibiotic use by MSM for the treatment of other sexually transmitted diseases [13, 73, 74] are probable causes for such wgMLST results. This pressure would be weaker in cases of non-MSM outbreaks relating to short-term food or community transmissions, which would consequently present more clonal clusters of *Shigella* in the absence of a complex accessory genome, as was shown in our results. In addition, serotype conversions have been observed in prolonged outbreaks due to serotype-converting bacteriophages and frameshift mutations similarly described elsewhere [75, 76]. These results highlight the caution needed during outbreak investigations to not solely exclude cases based on serotypes. Phylogeny is therefore a more trustworthy tool, since we have data demonstrating that serotype conversions can happen spontaneously in an outbreak strain.

So far, our data have shown that core genome-based subtyping methods have been given the upper hand over wgMLST, although some have been deemed to impact global inter-laboratory comparisons (e.g. SNVPhyl and cgMLST) [21] due to differences in reference genome choice, pipeline parameters and input datasets used by different jurisdictions [68]. Nevertheless, by utilizing a consistent set of conserved loci and allele designations, such as the EnteroBase scheme used in coreMLST, these variations vanish and enable global data comparisons in addition to a potential classification nomenclature due to the absence of differences observed for different input datasets. While minor increases in distance metrics were observed for isolated outbreaks via SNVPhyl for *S. flexneri* and *S. sonnei*, these results were concordant with Reimer *et al.*'s [31] comparison of different input datasets for *L. monocytogenes* outbreaks, as the size of the shared genome had little influence on the outbreak diversities, but increased nonetheless. However, the cgMLST approach was shown to be unreliable for inter-laboratory comparisons, while wgMLST produced minor decreases in distances. This decrease is due to the *categorical(values)* coefficient used in BioNumerics to calculate the pairwise distances. During a wgMLST analysis with isolated outbreaks, only the *n* total component will vary compared to the analysis of complete collections. Therefore, only including closely related isolates in the wgMLST analysis will reduce the total number of loci (*n* total) detected in the input dataset due to minimal genetic diversities and similar accessory genomes, and will consequently minimize the effect of the *n* total/*n* common factor on the *n* differences component. Applying the *categorical(differences)* coefficient instead could eliminate the fluctuating nature of the wgMLST analysis by only comparing true genetic variants (*n* differences) and remove the effect of loci presence/absence.

To conclude, it is of utmost importance that subtyping methods approximate epidemiology in an outbreak or surveillance scenario. In other words, isolates that cluster closely together and share a recent common ancestor via phylogenomic approaches should be indicative of a shared source of infection or chain of transmission [63]. It is also assumed that subtyping methods with a higher discriminatory power will indicate stronger epidemiological relationships. With that being said, even though wgMLST proved to be the most discriminatory for our *S. flexneri* and *S. sonnei* datasets, core genome-based subtyping methods were more phylogenetically consistent and epidemiologically concordant due to their exclusion of genetic variation in the accessory genome and filtering processes of repetitive loci and confounding MGEs. While MGE-mediated discrepancies were only observed in prolonged MSM-related outbreaks of *S. sonnei*, similar irregularities could theoretically occur at any time in any pathogen, but a higher degree of caution is needed for highly recombinant species and long spanning transmission patterns with strong selective pressures. These (1) highlight the importance of validating WGS-based subtyping methods for each pathogen separately on a basis of genome plasticity and outbreak dynamics, (2) support the use of a combination of different WGS analyses in order to obtain a better perspective of a given outbreak and (3) emphasize the application of a certain flexibility to proposed interpretation thresholds of relatedness. However, horizontally acquired genetic elements should not be deemed unusable, as plasmid-, IS element-, phage-, antimicrobial resistance-, virulence- and CRISPR-based subtyping have been shown to be valuable tools for public health investigations and risk assessments for other pathogens [70, 77, 78]. Further studies are therefore recommended in order to validate the parallel use of similar approaches to improve the typeability of *Shigella* spp. in the WGS era. We have also shown the utility of using a minimum spanning distance metric for future outbreak investigations, since it can be applied to outbreaks of different transmission natures and is not influenced by the time span of a given outbreak.

Funding information
The authors received no specific grant from any funding agency.

Acknowledgements
The authors would like to thank the Bacteriology Unit staff at the LSPQ for bacterial serotyping and the members of the Genomics Core at the NML for whole-genome sequencing, in addition to the PNC staff and the Division of Enteric Diseases Reference Services at the NML. The authors would also like to thank Khadidja Yousfi, Éric Fournier and Florence Doualla-Bell for their help in generating the proper outputs needed for data analysis and their knowledgeable insights.

Authors and contributors
Conceptualization, S.B. and I.B. Methodology, S.B. and I.B. Formal analysis, I.B. Investigation, I.B. Data curation, C.G., P.A.P. and S.B. Writing – original draft, I. B. Writing – review and editing, S. B. Visualization, I.B. Supervision, S.B.

Conflicts of interest
The author(s) declare that there are no conflicts of interest.

Ethical statement

This study uses strains isolated from humans and obtained in the context of a provincial surveillance programme for enteric pathogens. The collected isolates' secondary use for research purposes therefore did not require a review or approval by an ethics committee at the LSPQ.

References

1. Bernaquez I, Gaudreau C, Pilon PA, Bekal S. Evaluation of whole genome sequencing-based subtyping methods for the surveillance of *Shigella* spp. and the confounding effect of mobile genetic elements in long-term outbreaks. *Figshare*. 2021https://doi.org/10.6084/m9.figshare.14757978.v1
2. Government of Canada, PHAC. Reported cases from 1991 to 2018 in Canada – notifiable diseases on-line. https://diseases.canada.ca/notifiable/charts?c=yl
3. Gaudreau C. Ciprofloxacin-Resistant *Shigella sonnei* among Men Who Have Sex with Men, Canada, 2010. *Emerg Infect Dis* 2011;17:1747–1750.
4. Gaudreau C, Barkati S, Leduc J-. M, Pilon PA, Favreau J. Shigella spp. with Reduced Azithromycin Susceptibility, Quebec, Canada, 2012–2013. *Emerg Infect Dis* 2014;20:854–856.
5. Gaudreau C, Pilon PA, Cornut G, Marchand-Senecal X. *Shigella flexneri* with Ciprofloxacin Resistance and Reduced Azithromycin Susceptibility, Canada, 2015. *Emerg Infect Dis* 2016;22:3.
6. Wilmer A, Romney MG, Gustafson R, Sandhu J, Chu T. *Shigella flexneri* serotype 1 infections in men who have sex with men in Vancouver, Canada. *HIV Med* 2015;16:168–175.
7. Gilbart VL, Simms I, Jenkins C, Furegato M, Gobin M. Sex, drugs and smart phone applications: findings from semistructured interviews with men who have sex with men diagnosed with *Shigella flexneri* 3a in England and Wales. *Sex Transm Infect* 2015;91:598–602.
8. Pilon P, Camara B, Bekal S. Outbreak of *Shigella sonnei* in Montréal's ultra-Orthodox Jewish community, 2015. *Can Commun Dis Rep* 2016;42:88–90.
9. Naimi TS, Wicklund JH, Olsen SJ, Krause G, Wells JG. Concurrent outbreaks of Shigella sonnei and enterotoxigenic Escherichia coli infections associated with parsley: implications for surveillance and control of foodborne illness. *J Food Prot* 2003;66:535–541.
10. Drews SJ, Lau C, Andersen M, Ferrato C, Simmonds K. Laboratory based surveillance of travel-related Shigella sonnei and Shigella flexneri in Alberta from 2002 to 2007. *Global Health* 2010;6:20.
11. Kozak GK, MacDonald D, Landry L, Farber JM. Foodborne outbreaks in Canada linked to produce: 2001 through 2009. *J Food Prot* 2013;76:173–183.
12. Pons W, Young I, Truong J, Jones-Bitton A, McEwen S, *et al*. A systematic review of waterborne disease outbreaks associated with small non-community drinking water systems in Canada and the United States. *PLOS ONE* 2015;10:e0141646.
13. Baker KS, Dallman TJ, Ashton PM, Day M, Hughes G. Intercontinental dissemination of azithromycin-resistant shigellosis through sexual transmission: a cross-sectional study. *The Lancet Infect Dis* 2015;15:913–921.
14. Simms I, Field N, Jenkins C, Childs T, Gilbart VL *et al*. Intensified shigellosis epidemic associated with sexual transmission in men who have sex with men --Shigella flexneri and S. sonnei in England, 2004 to end of February 2015. *Euro Surveill* 2015;20:21097.
15. Bowen A, Grass J, Bicknese A, Campbell D, Hurd J. Elevated Risk for Antimicrobial Drug–Resistant Shigella Infection among Men Who Have Sex with Men, United States, 2011–2015. *Emerg Infect Dis* 2016;22:1613–1616.
16. Ingle DJ, Easton M, Valcanis M, Seemann T, Kwong JC. Co-circulation of Multidrug-resistant Shigella Among Men Who Have Sex With Men in Australia. *Clin Infect Dis* 2019;69:1535–1544.
17. Chiou C-S, Izumiya H, Kawamura M, Liao Y-S, Su Y-S, *et al*. The worldwide spread of ciprofloxacin-resistant *Shigella Sonnei* among hiv-infected men who have sex with men, Taiwan. *Clinical Microbiology and Infection* 2016;22:383.
18. Sati HF, Bruinsma N, Galas M, Hsieh J, Sanhueza A. Characterizing Shigella species distribution and antimicrobial susceptibility to ciprofloxacin and nalidixic acid in Latin America between 2000–2015. *PLOS ONE* 2019;14:e0220445.
19. Ronholm J, Nasheri N, Petronella N, Pagotto F. Navigating Microbiological Food Safety in the Era of Whole-Genome Sequencing. *Clin Microbiol Rev* 2016;29:837–857.
20. Swaminathan B, Gerner-Smidt P, L-K N, Lukinmaa S, Kam K-. M. Building PulseNet International: an interconnected system of laboratory networks to facilitate timely public health recognition and response to foodborne disease outbreaks and emerging foodborne diseases. *Foodborne Pathog Dis* 2006;3:36–50.
21. Nadon C, Walle V, Gerner-Smidt P, Campos J, Chinen I. PulseNet International: Vision for the implementation of whole genome sequencing (WGS) for global food-borne disease surveillance. *Euro Surveill* 2017;22:30544.
22. Bekal S, Berry C, Reimer AR, Van Domselaar G, Beaudry G. Usefulness of high-quality core genome single-nucleotide variant analysis for subtyping the highly clonal and the most prevalent salmonella enterica serovar heidelberg clone in the context of outbreak investigations. *J Clin Microbiol* 2016;54:289–295.
23. Vincent C, Usongo V, Berry C, Tremblay DM, Moineau S. Comparison of advanced whole genome sequence-based methods to distinguish strains of Salmonella enterica serovar Heidelberg involved in foodborne outbreaks in Québec. *Food Microbiol* 2018;73:99–110.
24. Rumore J, Tschetter L, Kearney A, Kandar R, McCormick R. Evaluation of whole-genome sequencing for outbreak detection of Verotoxigenic *Escherichia coli* O157:H7 from the Canadian perspective. *BMC Genomics* 2018;19:870.
25. Pearce ME, Alikhan N-. F, Dallman TJ, Zhou Z, Grant K. Comparative analysis of core genome MLST and SNP typing within a European *Salmonella* serovar Enteritidis outbreak. *Int J Food Microbiol* 2018;274:1–11.
26. Jackson BR, Tarr C, Strain E, Jackson KA, Conrad A. Implementation of Nationwide Real-time Whole-genome Sequencing to Enhance Listeriosis Outbreak Detection and Investigation. *Clin Infect Dis* 2016;63:380–386.
27. Applied Maths. Escherichia coli - Shigella Schema for whole genome sequence typing Release Note. https://www.applied-maths.com/sites/default/files/extra/Release-Note-Escherichacoli-Shigella-schema.pdf
28. Ruppitsch W, Pietzka A, Prior K, Bletz S, Fernandez HL. Defining and Evaluating a Core Genome Multilocus Sequence Typing Scheme for Whole-Genome Sequence-Based Typing of Listeria monocytogenes. *J Clin Microbiol* 2015;53:2869–2876.
29. Moura A, Criscuolo A, Pouseele H, Maury MM, Leclercq A *et al*. Whole genome-based population biology and epidemiological surveillance of Listeria monocytogenes. *Nat Microbiol* 2017;2:16185.
30. Petkau A, Mabon P, Sieffert C, Knox NC, Cabral J. SNVPhyl: a single nucleotide variant phylogenomics pipeline for microbial genomic epidemiology. *Microb Genom* 2017;3:e000116.
31. Reimer A, Weedmark K, Petkau A, Peterson C-. L, Walker M. Shared genome analyses of notable listeriosis outbreaks, highlighting the critical importance of epidemiological evidence, input datasets and interpretation criteria. *Microb Genom* 2019;5:e000237.
32. Dallman TJ, Byrne L, Ashton PM, Cowley LA, Perry NT. Whole-Genome Sequencing for National Surveillance of Shiga Toxin–Producing Escherichia coli O157. *Clin Infect Dis* 2015;61:305–312.
33. Pupo GM, Lan R, Reeves PR. Multiple independent origins of *Shigella* clones of *Escherichia coli* and convergent evolution of many of their characteristics. *Proc Natl Acad Sci USA* 2000;97:10567–10572.
34. Worby CJ, Lipsitch M, Hanage WP. Within-Host Bacterial Diversity Hinders Accurate Reconstruction of Transmission Networks from Genomic Distance Data. *PLOS Comput Biol* 2014;10:e1003549.
35. Kothary MH, Babu US. Infective dose of foodborne pathogens in volunteers: A review. *J Food Safety* 2001;21:49–68.

36. Baker S, Thomson N, Weill F-. X, Holt KE. Genomic insights into the emergence and spread of antimicrobial-resistant bacterial pathogens. *Science* 2018;360:733–738.

37. Newton ILG, Bordenstein SR. Correlations Between Bacterial Ecology and Mobile DNA. *Curr Microbiol* 2011;62:198–208.

38. Jin Q, Yuan Z, Xu J, Wang Y, Shen Y. Genome sequence of Shigella flexneri 2a: insights into pathogenicity through comparison with genomes of *Escherichia coli* K12 and O157. *Nucleic Acids Res* 2002;30:4432–4441.

39. Anderson M, Sansonetti PJ, Marteyn BS. Shigella Diversity and Changing Landscape: Insights for the Twenty-First Century. *Front Cell Infect Microbiol* 2016;6:45.

40. Thompson CN, Duy PT, Baker S. The Rising Dominance of Shigella sonnei: An Intercontinental Shift in the Etiology of Bacillary Dysentery. *PLOS Negl Trop Dis* 2015;9:e0003708.

41. Greig DR, Jenkins C, Dallman TJ. A Shiga Toxin-Encoding Prophage Recombination Event Confounds the Phylogenetic Relationship Between Two Isolates of Escherichia coli O157:H7 From the Same Patient. *Front Microbiol* 2020;11:588769.

42. Katz LS, Griswold T, Williams-Newkirk AJ, Wagner D, Petkau A. A Comparative Analysis of the Lyve-SET Phylogenomics Pipeline for Genomic Epidemiology of Foodborne Pathogens. *Front Microbiol* 2017;8:375.

43. Andrews S. Fastqc: A quality control tool for high throughput sequence data. . https://www.bioinformatics.babraham.ac.uk/projects/fastqc/ [accessed 06 Jan 2021].

44. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *J Comput Biol* 2012;19:455–477.

45. Afgan E, Baker D, Batut B, van den Beek M, Bouvier D. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Res* 2018;46:W537–W544.

46. Arndt D, Grant JR, Marcu A, Sajed T, Pon A. PHASTER: a better, faster version of the PHAST phage search tool. *Nucleic Acids Res* 2016;44:W16–W21.

47. Bertelli C, Laird MR, Williams KP, Simon Fraser University Research Computing Group, Lau BY. IslandViewer 4: expanded prediction of genomic islands for larger-scale datasets. *Nucleic Acids Res* 2017;45:W30–W35.

48. Government of Canada, PHAC-NML. rearrange_snv_matrix.pl. https://github.com/phac-nml/snvphyl-tools/blob/master/bin/rearrange_snv_matrix.pl

49. The Institute of Evolutionary Biology. FigTree. Molecular evolution, phylogenetics and epidemiology. http://tree.bio.ed.ac.uk/software/figtree/

50. Goverment of Canada, PHAC-NML. 2017. positions2phyloviz.pl. https://github.com/phac-nml/snvphyl-tools/blob/master/bin/positions2phyloviz.pl

51. Francisco AP, Vaz C, Monteiro PT, Melo-Cristino J, Ramirez M. PHYLOViZ: phylogenetic inference and data visualization for sequence based typing methods. *BMC Bioinformatics* 2012;13:87.

52. Hunter PR, Gaston MA. Numerical index of the discriminatory ability of typing systems: an application of Simpson's index of diversity. *J Clin Microbiol* 1988;26:2465–2466.

53. Severiano A, Carriço JA, Robinson DA, Ramirez M, Pinto FR. Evaluation of jackknife and bootstrap for defining confidence intervals for pairwise agreement measures. *PLOS ONE* 2011;6:e19539.

54. Fournier É. SNPlocalisationV6.py. 2019. https://github.com/Eric-Fournier3/EnteroWGS/blob/master/SNPlocalisationV6.py

55. Robertson J, Nash JHE. MOB-suite: software tools for clustering, reconstruction and typing of plasmids from draft assemblies. *Microb Genom* 2018;4:e000206.

56. Wei J, Goldberg MB, Burland V, Venkatesan MM, Deng W. Complete genome sequence and comparative genomics of Shigella flexneri serotype 2a strain 2457T. *Infect Immun* 2003;71:2775–2786.

57. Adhikari P, Allison G, Whittle B, Verma NK. Serotype 1a O-antigen modification: molecular characterization of the genes involved and their novel organization in the Shigella flexneri chromosome. *J Bacteriol* 1999;181:4711–4718.

58. Mavris M, Manning PA, Morona R. Mechanism of bacteriophage SfII-mediated serotype conversion in Shigella flexneri. *Mol Microbiol* 1997;26:939–950.

59. Allison GE, Verma NK. Serotype-converting bacteriophages and O-antigen modification in Shigella flexneri. *Trends Microbiol* 2000;8:17–23.

60. Chen J-. H, Hsu W-. B, Chiou C-. S, Chen C-. M. Conversion of Shigella flexneri serotype 2a to serotype Y in a shigellosis patient due to a single amino acid substitution in the protein product of the bacterial glucosyltransferase gtrII gene. *FEMS Microbiol Lett* 2003;224:277–283.

61. Kohl TA, Diel R, Harmsen D, Rothgänger J, Walter KM. Whole-genome-based Mycobacterium tuberculosis surveillance: a standardized, portable, and expandable approach. *J Clin Microbiol* 2014;52:2479–2486.

62. Lee Y, Kim B-. S, Chun J, Yong JH, Lee YS. Clonality and Resistome analysis of KPC-producing Klebsiella pneumoniae strain isolated in Korea using whole genome sequencing. *Biomed Res Int* 2014;2014:352862.

63. Blanc DS, Magalhães B, Koenig I, Senn L, Grandbastien B. Comparison of Whole Genome (wg-) and Core Genome (cg-) MLST (BioNumerics™) Versus SNP Variant Calling for Epidemiological Investigation of Pseudomonas aeruginosa. *Front Microbiol* 2020;11:1729.

64. Henri C, Leekitcharoenphon P, Carleton HA, Radomski N, Kaas RS. An Assessment of Different Genomic Approaches for Inferring Phylogeny of Listeria monocytogenes. *Front Microbiol* 2017;8:2351.

65. Chattaway MA, Greig DR, Gentle A, Hartman HB, Dallman TJ. Whole-Genome Sequencing for National Surveillance of Shigella flexneri. *Front Microbiol* 2017;8:1700.

66. McDonnell J, Dallman T, Atkin S, Turbitt DA, Connor TR. Retrospective analysis of whole genome sequencing compared to prospective typing data in further informing the epidemiological investigation of an outbreak of Shigella sonnei in the UK. *Epidemiol Infect* 2013;141:2568–2575.

67. Mitchell HD, Mikhail AFW, Painset A, Dallman TJ, Jenkins C. Use of whole-genome sequencing to identify clusters of Shigella flexneri associated with sexual transmission in men who have sex with men in England: a validation study using linked behavioural data. *Microb Genom* 2019;5:e000311:11.:.

68. Schürch AC, Arredondo-Alonso S, Willems RJL, Goering RV. Whole genome sequencing options for bacterial strain typing and epidemiologic analysis based on single nucleotide polymorphism versus gene-by-gene–based approaches. *Clin Microbiol Infect* 2018;24:350–354.

69. Pightling AW, Pettengill JB, Luo Y, Baugher JD, Rand H. Interpreting Whole-Genome Sequence Analyses of Foodborne Bacteria for Regulatory Applications and Outbreak Investigations. *Front Microbiol* 2018;9:1482.

70. EFSA Panel on Biological Hazards (EFSA BIOHAZ Panel), Koutsoumanis K, Allende A, Alvarez-Ordóñez A, *et al*. Whole genome sequencing and metagenomics for outbreak investigation, source attribution and risk assessment of food-borne microorganisms. *EFSA J* 2019;17:12.

71. Thanh Duy P, Thi Nguyen TN, Vu Thuy D, Chung The H, Alcock F. Commensal *Escherichia coli* are a reservoir for the transfer of XDR plasmids into epidemic fluoroquinolone-resistant *Shigella* sonnei. *Nat Microbiol* 2020;5:256–264.

72. Cowley LA, Dallman TJ, Fitzgerald S, Irvine N, Rooney PJ. Short-term evolution of Shiga toxin-producing *Escherichia coli* O157:H7 between two food-borne outbreaks. *Microb Genom* 2016;2:e000084.

73. Baker KS, Dallman TJ, Field N, Childs T, Mitchell H. Horizontal antimicrobial resistance transfer drives epidemics of multiple Shigella species. *Nat Commun* 2018;9:1462.

74. Baker KS, Dallman TJ, Field N, Childs T, Mitchell H. Genomic epidemiology of Shigella in the United Kingdom shows transmission of pathogen sublineages and determinants of antimicrobial resistance. *Sci Rep* 2018;8:7389.

75. Gentle A, Ashton PM, Dallman TJ, Jenkins C. Evaluation of Molecular Methods for Serotyping Shigella flexneri. *J Clin Microbiol* 2016;54:1456–1461.

76. Brengi SP, Sun Q, Bolaños H, Duarte F, Jenkins C. PCR-Based Method for Shigella flexneri Serotyping: International Multicenter Validation. *J Clin Microbiol* 2019;57:e01592-18.

77. Werner G, Fleige C, Geringer U, van Schaik W, Klare I. IS element IS16 as a molecular screening tool to identify hospital-associated strains of Enterococcus faecium. *BMC Infect Dis* 2011;11:80.

78. Yousfi K, Usongo V, Berry C, Khan RH, Tremblay DM. Source Tracking Based on Core Genome SNV and CRISPR Typing of Salmonella enterica Serovar Heidelberg Isolates Involved in Foodborne Outbreaks in Québec, 2012. *Front Microbiol* 2020;11:1317.