

REPORT DOCUMENTATION PAGE			Form Approved OMB NO. 0704-0188	
The public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA, 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number. PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.				
1. REPORT DATE (DD-MM-YYYY)		2. REPORT TYPE New Reprint		3. DATES COVERED (From - To) -
4. TITLE AND SUBTITLE Comparative Analysis of Subtyping Methods against a Whole-Genome-Sequencing Standard for Salmonella enterica Serotype Enteritidis		5a. CONTRACT NUMBER W911NF-11-1-0442		
		5b. GRANT NUMBER		
		5c. PROGRAM ELEMENT NUMBER 611102		
6. AUTHORS Xiangyu Deng, Nikki Shariat, Elizabeth M. Driebe, Chandler C. Roe, Beth Tolar, Eija Trees, Paul Keim, Wei Zhang, Edward G. Dudley, Patricia I. Fields, David M. Engelthaler, N. A. Ledebore		5d. PROJECT NUMBER		
		5e. TASK NUMBER		
		5f. WORK UNIT NUMBER		
7. PERFORMING ORGANIZATION NAMES AND ADDRESSES Pennsylvania State University Office of Sponsored Programs 110 Technology Center Building University Park, PA 16802 -7000			8. PERFORMING ORGANIZATION REPORT NUMBER	
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS (ES) U.S. Army Research Office P.O. Box 12211 Research Triangle Park, NC 27709-2211			10. SPONSOR/MONITOR'S ACRONYM(S) ARO	
			11. SPONSOR/MONITOR'S REPORT NUMBER(S) 60012-LS.6	
12. DISTRIBUTION AVAILABILITY STATEMENT Approved for public release; distribution is unlimited.				
13. SUPPLEMENTARY NOTES The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision, unless so designated by other documentation.				
14. ABSTRACT A retrospective investigation was performed to evaluate whole-genome sequencing as a benchmark for comparing molecular subtyping methods for Salmonella enterica serotype Enteritidis and survey the population structure of commonly encountered S. enterica serotype Enteritidis outbreak isolates in the United States. A total of 52 S. enterica serotype Enteritidis isolates representing 16 major outbreaks and three sporadic cases collected between 2001 and 2012 were sequenced and subjected to subtyping by four different methods: (i) whole-genome single-nucleotide polymorphism typing (WGSST), (ii) multiple locus variable number tandem repeat (VNTR) analysis,				
15. SUBJECT TERMS Salmonella, subtyping, CRISPR				
16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	15. NUMBER OF PAGES
a. REPORT UU	b. ABSTRACT UU	c. THIS PAGE UU	UU	19a. NAME OF RESPONSIBLE PERSON Edward Dudley
				19b. TELEPHONE NUMBER 814-867-0439

Report Title

Comparative Analysis of Subtyping Methods against a Whole-Genome-Sequencing Standard for *Salmonella enterica* Serotype Enteritidis

ABSTRACT

A retrospective investigation was performed to evaluate whole-genome sequencing as a benchmark for comparing molecular subtyping methods for *Salmonella enterica* serotype Enteritidis and survey the population structure of commonly encountered *S. enterica* serotype Enteritidis outbreak isolates in the United States. A total of 52 *S. enterica* serotype Enteritidis isolates representing 16 major outbreaks and three sporadic cases collected between 2001 and 2012 were sequenced and subjected to subtyping by four different methods: (i) whole-genome single-nucleotide-polymorphism typing (WGST), (ii) multiple-locus variable-number tandem-repeat (VNTR) analysis (MLVA), (iii) clustered regularly interspaced short palindromic repeats combined with multi-virulence-locus sequence typing (CRISPR-MVLST), and (iv) pulsed-field gel electrophoresis (PFGE). WGST resolved all outbreak clusters and provided useful robust phylogenetic inference results with high epidemiological correlation. While both MLVA and CRISPR-MVLST yielded higher discriminatory power than PFGE, MLVA outperformed the other methods in delineating outbreak clusters whereas CRISPR-MVLST showed the potential to trace major lineages and ecological origins of *S. enterica* serotype Enteritidis. Our results suggested that whole-genome sequencing makes a viable platform for the evaluation and benchmarking of molecular subtyping methods.

REPORT DOCUMENTATION PAGE (SF298) (Continuation Sheet)

Continuation for Block 13

ARO Report Number 60012.6-LS
Comparative Analysis of Subtyping Methods agi...

Block 13: Supplementary Note

© 2015 . Published in Journal of Clinical Microbiology, Vol. 53 (1) (2015), (3 (1)). DoD Components reserve a royalty-free, nonexclusive and irrevocable right to reproduce, publish, or otherwise use the work for Federal purposes, and to authorize others to do so (DODGARS §32.36). The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision, unless so designated by other documentation.

Approved for public release; distribution is unlimited.

Comparative Analysis of Subtyping Methods against a Whole-Genome-Sequencing Standard for *Salmonella enterica* Serotype Enteritidis

Xiangyu Deng,^a Nikki Shariat,^b Elizabeth M. Driebe,^c Chandler C. Roe,^c Beth Tolar,^d Eija Trees,^d Paul Keim,^{c,e} Wei Zhang,^f Edward G. Dudley,^{b,g} Patricia I. Fields,^d David M. Engelthaler^c

Center for Food Safety, Department of Food Science and Technology, University of Georgia, Griffin, Georgia, USA^a; Department of Food Science, The Pennsylvania State University, University Park, Pennsylvania, USA^b; Translational Genomics Research Institute, Flagstaff, Arizona, USA^c; Division of Foodborne, Waterborne and Environmental Diseases, Centers for Disease Control and Prevention, Atlanta, Georgia, USA^d; Northern Arizona University, Flagstaff, Arizona, USA^e; Institute for Food Safety and Health, Illinois Institute of Technology, Bedford Park, Illinois, USA^f; Center for Molecular Immunology and Infectious Diseases, The Pennsylvania State University, University Park, Pennsylvania, USA^g

A retrospective investigation was performed to evaluate whole-genome sequencing as a benchmark for comparing molecular subtyping methods for *Salmonella enterica* serotype Enteritidis and survey the population structure of commonly encountered *S. enterica* serotype Enteritidis outbreak isolates in the United States. A total of 52 *S. enterica* serotype Enteritidis isolates representing 16 major outbreaks and three sporadic cases collected between 2001 and 2012 were sequenced and subjected to subtyping by four different methods: (i) whole-genome single-nucleotide-polymorphism typing (WGST), (ii) multiple-locus variable-number tandem-repeat (VNTR) analysis (MLVA), (iii) clustered regularly interspaced short palindromic repeats combined with multi-virulence-locus sequence typing (CRISPR-MVLST), and (iv) pulsed-field gel electrophoresis (PFGE). WGST resolved all outbreak clusters and provided useful robust phylogenetic inference results with high epidemiological correlation. While both MLVA and CRISPR-MVLST yielded higher discriminatory power than PFGE, MLVA outperformed the other methods in delineating outbreak clusters whereas CRISPR-MVLST showed the potential to trace major lineages and ecological origins of *S. enterica* serotype Enteritidis. Our results suggested that whole-genome sequencing makes a viable platform for the evaluation and benchmarking of molecular subtyping methods.

Salmonella enterica is currently the most common bacterial foodborne pathogen in the United States, causing over 1 million cases of illnesses annually, including approximately 20,000 hospitalizations and 400 deaths (1). Serotyping is commonly used to subtype strains below the species level for epidemiologic purposes. *Salmonella enterica* serotype Enteritidis was the serotype most commonly linked to foodborne outbreaks between 1998 and 2008 in the United States, with shell eggs being the major vehicle for foodborne transmission (2). In recent years, *S. enterica* serotype Enteritidis was also found to cause multistate outbreaks associated with other foods such as ground beef (2012), Turkish pine nuts (2011), and alfalfa and spicy sprouts (2011), in addition to shelled eggs (2010) (3).

During outbreak investigations, it is critical to employ subtyping methods capable of distinguishing outbreak isolates from epidemiologically distinct but genetically related bacterial strains. Most *S. enterica* serotype Enteritidis isolates have been shown to be genetically homogeneous, making it difficult for conventional subtyping methods such as pulsed-field gel electrophoresis (PFGE), the current gold standard for strain-level *Salmonella* subtyping, to discriminate between strains (4, 5). Among the *S. enterica* serotype Enteritidis isolates reported to PulseNet (6), approximately 45% display a single PFGE pattern using XbaI (JEGX01.0004), rendering PFGE ineffective in some foodborne outbreak investigations. One strategy to improve subtype resolution is to target hypervariable regions (i.e., regions of the bacterial chromosome with less genetic stability) in the bacterial genome to produce sufficient polymorphism for strain differentiation. Two such methods have been developed and evaluated with *S. enterica*

serotype Enteritidis isolates. Multilocus variable-number tandem-repeat analysis (MLVA) utilizes the polymorphism in the copy numbers of tandemly repeated sequences at multiple loci in the *S. enterica* serotype Enteritidis genome. It provides higher resolution than PFGE (7, 8) and has become a supplementary subtyping technique for surveillance and investigation of *S. enterica* serotype Enteritidis outbreaks by PulseNet. Analysis using clustered regularly interspaced short palindromic repeats (CRISPRs) combined with multi-virulence-locus sequence typing (designated CRISPR-MVLST) takes advantage of combined sequence variations in the spacer regions of the two CRISPR loci in *Salmonella* and two virulence genes (*fimH* and *sseL*) (9). This recently proposed subtyping scheme allowed better discrimination of *S. enterica* serotype Enteritidis isolates than PFGE (10).

Received 13 August 2014 Returned for modification 5 September 2014

Accepted 29 October 2014

Accepted manuscript posted online 5 November 2014

Citation Deng X, Shariat N, Driebe EM, Roe CC, Tolar B, Trees E, Keim P, Zhang W, Dudley EG, Fields PI, Engelthaler DM. 2015. Comparative analysis of subtyping methods against a whole-genome-sequencing standard for *Salmonella enterica* serotype Enteritidis. J Clin Microbiol 53:212–218. doi:10.1128/JCM.02332-14.

Editor: N. A. Ledebore

Address correspondence to Xiangyu Deng, xdeng@uga.edu.

Supplemental material for this article may be found at <http://dx.doi.org/10.1128/JCM.02332-14>.

Copyright © 2015, American Society for Microbiology. All Rights Reserved.

doi:10.1128/JCM.02332-14

Common criteria to evaluate the efficacy of subtyping methods include discriminatory power and clustering concordance with epidemiological data. Both MLVA and CRISPR-MVLST have been assessed in *Salmonella* based on these criteria (7, 8, 10–13). Evaluation of subtyping methods is often conducted through comparisons with PFGE; however, PFGE is not sufficiently discriminatory against clonal organisms such as *S. enterica* serotype Enteritidis and its utility as a benchmark for other subtyping techniques can be compromised. In recognition of this, multiple enzymes have been used as part of a PFGE scheme to improve discrimination (5). Nevertheless, the lack of diversity in PFGE patterns, as in the case of *S. enterica* serotype Enteritidis subtyping, may prevent the differentiation of epidemiologically unrelated isolates.

Powered by whole-genome-sequencing (WGS) technologies, recent implementations of whole-genome single-nucleotide-polymorphism (SNP) typing (WGST) have led to substantial improvements of both molecular subtyping and phylogenetic analyses, particularly for genetically homogenous bacterial pathogens such as *S. enterica* serotype Enteritidis (14, 15). A recent WGS-based survey of *S. enterica* serotype Enteritidis isolates resolved the commonly circulating *S. enterica* serotype Enteritidis populations in the United States into five major genetic lineages, revealing potential patterns in their geographical and epidemiological distribution (15).

WGS allows discovery of SNPs across entire bacterial genomes, thereby providing superior subtyping resolution and phylogenetic accuracy, which can be utilized for benchmarking other subtyping methods. In this study, we assembled a cohort of 52 *S. enterica* serotype Enteritidis isolates from 15 major foodborne disease outbreaks and three sporadic cases in the United States and 1 outbreak in Mauritius between 2001 and 2012. A retrospective investigation of these isolates was performed with a combination of WGST, MLVA, CRISPR-MVLST, and PFGE analyses to compare their respective performances in delineating each individual outbreak under the guidance of the recently proposed phylogenetic framework and population structure of *S. enterica* serotype Enteritidis (15).

MATERIALS AND METHODS

Bacterial isolates. A total of 52 *S. enterica* serotype Enteritidis isolates were obtained from the National *Salmonella* Reference Laboratory at the Centers for Disease Control and Prevention (Table 1). Forty-nine isolates were epidemiologically linked to 16 outbreaks, and three were isolated from sporadic cases. The sporadic isolates were isolated during a 2012 outbreak of ground beef infection (outbreak D; <http://www.cdc.gov/salmonella/enteritidis-07-12/>). They were included to test the ability of a particular subtyping method to distinguish between sporadic and outbreak isolates.

WGST. Bacterial strains were grown in Luria broth at 37°C to the stationary phase. Genomic DNA was prepared using a GenElute genomic DNA isolation kit (Sigma-Aldrich, St. Louis, MO). WGS was performed at TGen North using Illumina technology (100-bp paired-end reads) as described in previous studies (16, 17). All WGS data files were deposited in the NCBI Sequence Read Archive (<http://www.ncbi.nlm.nih.gov/bioproject>) under project number PRJNA251730. Average coverage of sequencing is summarized in Table S3 in the supplemental material. SNP detection was performed similarly to what was described in our previous study (15). Briefly, trimmed and filtered sequencing reads were mapped to a reference genome (P125109; GenBank accession no. [AM933172.1](http://www.ncbi.nlm.nih.gov/nuclseq/AM933172.1)) to call variants (SNPs, insertions, and deletions). For each genome analyzed, a list of high-quality SNPs was derived by subjecting initial SNP calls to a

TABLE 1 Isolates used in this study^a

Isolate	Outbreak	Epidemiologic information
J0900	A	Almonds, CA, 2001
J0905	A	Almonds, CA, 2001
2011K-1845	B	Fast food restaurant, TX, 2011
2011K-1846	B	Fast food restaurant, TX, 2011
H9556	C	Juice, CA, 2003
H9558	C	Juice, CA, 2003
2012K-0627	D	Ground beef, VT, 2012
2012K-0628	D	Ground beef, VT, 2012
2012K-0644	D	Ground beef, VT, 2012
2012K-0738	NA	Sporadic case during outbreak D, MD, 2012
2012K-0619	NA	Sporadic case during outbreak D, TX, 2012
2012K-0597	NA	Sporadic case during outbreak D, GA, 2012
2009K-1740	E	Chicken, MD, 2009
2009K-1742	E	Chicken, MD, 2009
2010K-0338	F	Chili sauce, Mauritius, 2009
2010K-0348	F	Uncooked chicken tikka, Mauritius, 2009
2010K-0351	F	Mauritius, 2009
2010K-0358	F	Raw chicken, Mauritius, 2009
2010K-0362	F	Mauritius, 2009
2011K-1667	G	Turkish pine nuts, NY, 2011
2011K-1668	G	Turkish pine nuts, NY, 2011
K3308	H	Stuffed chicken products, MN, 2006
K3310	H	Stuffed chicken products, MN, 2006
K2330	I	OH, 2005
K2331	I	OH, 2005
2012K-0284	J	Elderly care facility, MA, 2012
2012K-0285	J	Elderly care facility, MA, 2012
2012K-0283	J	Elderly care facility, MA, 2012
2010K-2617	K	Guinea pig, WI, 2011
2011K-0019	K	Guinea pig, CA, 2011
2011K-0079	K	Guinea pig, OR, 2011
2011K-0104	K	Guinea pig, IL, 2011
2012K-0499	L	Restaurant, NC, 2012
2012K-0500	L	Restaurant, NC, 2012
2012K-0501	L	Restaurant, NC, 2012
2009K-1553	M	Eggs, PA, 2009
2009K-1559	M	Eggs, PA, 2009
2009K-1562	M	Eggs, PA, 2009
2010K-1946	N	Tall ships, PA, 2010
2010K-1947	N	Tall ships, PA, 2010
2009K-1545	M	Eggs, PA, 2009
K2082	O	Hospital eggs, GA, 2005
K2083	O	Hospital eggs, GA, 2005
2010K-0666	P	Restaurant, CT, 2010
2010K-0667	P	Restaurant, CT, 2010
2010K-0668	P	Restaurant, CT, 2010
2010K-0669	P	Restaurant, CT, 2010
2010K-0672	P	Restaurant, CT, 2010
2010K-0673	P	Restaurant, CT, 2010
2010K-0677	P	Food worker, CT, 2010
2010K-0678	P	Food worker, CT, 2010
2010K-0675	P	Restaurant, CT, 2010

^a Isolates J0900 and J0905 were collected from the environment; isolates 2012K-0644, 2010K-0338, 2010K-0348, 2010K-0358, and 2011K-1668 were collected from foods; all the other isolates were collected from humans. NA, not applicable.

set of quality filters, including a minimum Phred base score of 60, a minimum read-mapping score of 20, a mapping depth ranging from 5 to 100 reads per locus, and a maximum alternative allele percentage of 25%. SNPs were accepted only when confirmed by reads mapped to both the forward and reverse strands. High-quality SNPs detected from the conserved genome regions (i.e., core genome SNPs) among the 52 *S. enterica*

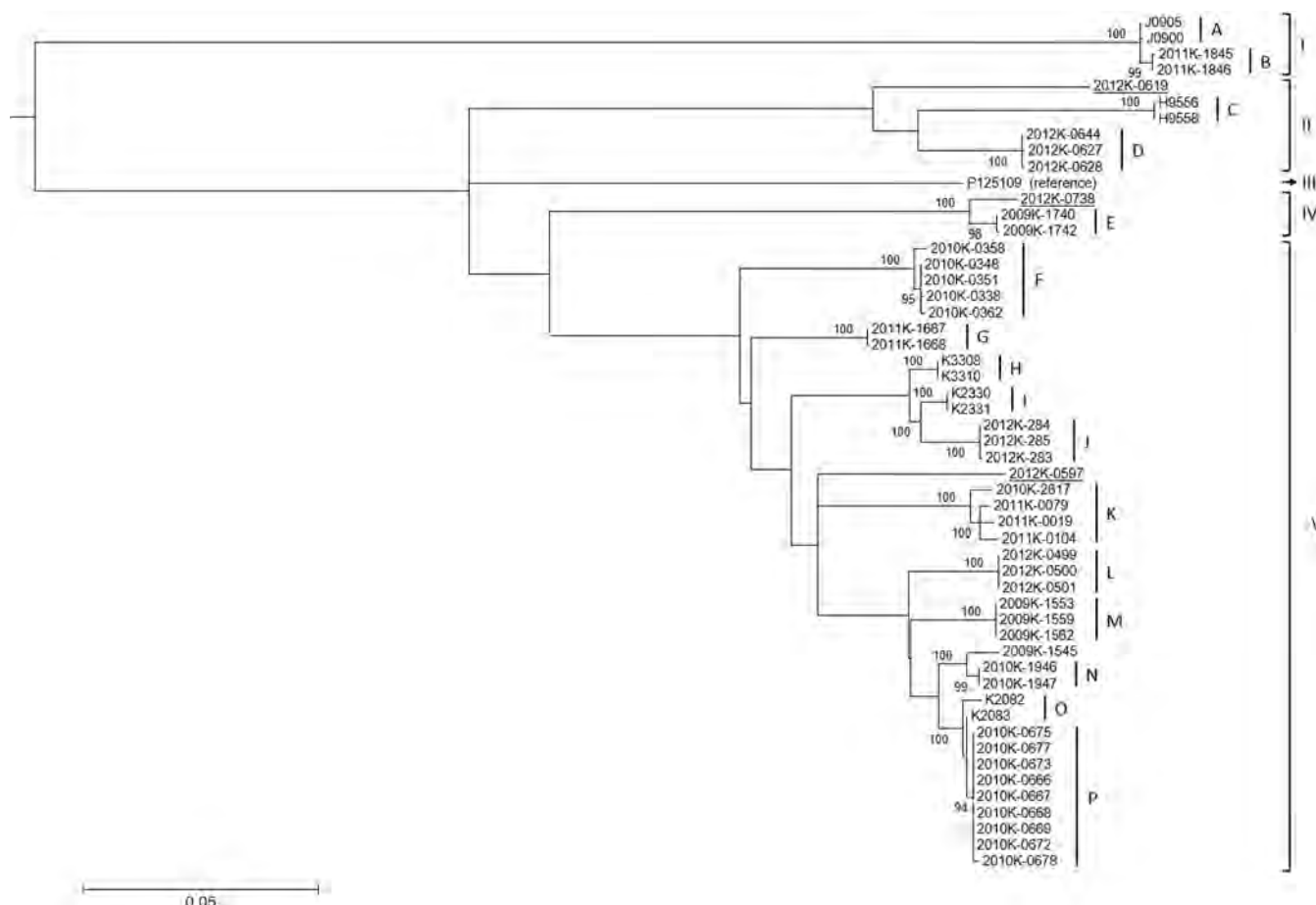


FIG 1 Phylogeny and outbreak clusters inferred by WGST. Different lineages (I, II, III, IV, and V) are labeled. A total of 16 outbreak clusters (A through P) are identified and labeled. Bootstrapping values of branches leading to individual outbreak clusters are labeled. The designations of three isolates from sporadic cases (2012K-0619, 2012K-0738, and 2012K-0597) are underlined.

serotype Enteritidis genomes and the reference genome were used to construct a maximum-likelihood (ML) tree using MEGA 5 (18). A similar ML tree was built by further incorporating 125 *S. enterica* serotype Enteritidis and 3 *Salmonella enterica* serotype Nitra recently sequenced genomes that represent the population structure of commonly circulating *S. enterica* serotype Enteritidis isolates in the United States (15).

PFGE and MLVA. PFGE (using XbaI) and MLVA were performed according to standard PulseNet protocols (19) (<http://www.cdc.gov/pulsenet/pathogens/>). Dendrograms of PFGE and MLVA patterns were generated by BioNumerics software (Applied-Maths, St.-Martens-Latem, Belgium).

CRISPR-MVLST. For each sequenced genome, contigs were *de novo* assembled by Velvet (20). The sequence of each marker (CRISPR1, CRISPR2, *fimH*, and *sseL*) was extracted from the respective contigs. Individual alleles were given a numeric identifier, as shown previously (9), and a CRISPR-MVLST sequence type was determined based on unique allelic combinations of each marker. The presence of homologous direct repeats and duplicated spacers can complicate contig assembly for the CRISPR arrays. The majority of CRISPR alleles were determined using the WGS data. For the few CRISPR sequences where we were unable to extract the CRISPR sequences, we PCR amplified and sequenced the CRISPR array as previously described (12). To depict the clustering of subtypes determined by CRISPR-MVLST, the binary distribution (presence or absence) of every spacer in CRISPR1 and CRISPR2 and every SNP in *fimH* and *sseL* was profiled for each isolate. Specifically, if a spacer or a SNP was present in an isolate, it was designated “1”; otherwise, it was designated

“0.” The binary distribution patterns of all isolates were then combined and input into SplitTree (21) to build a dendrogram by employing the unweighted-pair group method using average linkages (UPGMA) algorithm.

Discriminatory power. The ability to differentiate sampled *S. enterica* serotype Enteritidis isolates by the use of each subtyping method evaluated in this study was calculated using Simpson’s index of diversity (22).

RESULTS

WGST-based investigation of outbreak and sporadic isolates. A total of 2,353 SNPs were identified from the core genome of the 52 *S. enterica* serotype Enteritidis isolates and the reference strain. These SNPs resolved the cohort of outbreak and sporadic isolates into 34 SNP haplotypes and allowed the delineation of all 16 outbreak clusters (Fig. 1, clusters A through P). The inferred phylogeny of these isolates was highly consistent with their outbreak association. All but one outbreak isolate (2009K-1545) fell into their respective outbreak clusters. 2009K-1545 was considered to be associated with a shelled-egg outbreak in Pennsylvania in 2009 (outbreak M). However, it appeared to be phylogenetically more closely related to another outbreak among crew members of a historic sailing ship in the same state in 2010 (outbreak N). The three isolates (2012K-0619, 2012K-0738, and 2012K-0597) from sporadic cases that occurred during the 2012 ground beef out-

TABLE 2 Comparison of outbreak delineations of different subtyping methods^a

Outbreak	Result by:			
	WGST	MLVA	CRISPR-MLVST	PFGE
A	++	+	+	—
B	++	+	+	++
C	++	++	+	++
D	++	++	+	+
E	++	+	+	+
F	++	++	++	+
G	++	++	++	++
H	++	+	+	+
I	++	++	+	+
J	++	—	+	+
K	++	—	—	+
L	++	++	++	+
M	++	+	+	+
N	++	+	—	+
O	++	+	—	+
P	++	+	—	+

^a Symbols are used to report evaluations of subtyping methods. ++, isolates from the outbreak formed a cluster, and the cluster did not include isolates from other outbreaks or sporadic cases; +, isolates from the outbreak clustered with each other but also with isolates from other outbreaks or sporadic cases; —, isolates from the outbreak did not form a cluster.

break (outbreak D) were dispersed throughout the tree with substantial phylogenetic distances from the outbreak, indicating their separate origins from sources other than the contaminated ground beef (Fig. 1).

PFGE, MLVA, and CRISPR-MVLST subtyping. PFGE and MLVA results are summarized in Table S1 in the supplemental material. CRISPR-MVLST results are summarized in Table S2. Briefly, 15 different *S. enterica* serotype Enteritidis sequence types (ESTs) were identified among 52 different isolates using CRISPR-MLVST. Eleven ESTs were previously observed in other *S. enterica* serotype Enteritidis clinical isolates, and four (EST43, EST44, EST45, and EST46) appeared to be new (10, 23). The most frequent EST was EST12 (17% of isolates; 9/52), followed by EST4 (12%; 6/51). The four new ESTs were designated due to new alleles identified for *sseL* and CRISPR1 (EST43), CRISPR1 and CRISPR2 (EST44), CRISPR2 (EST45), or *sseL* (EST46).

Comparison of subtyping methods. Analysis of all *S. enterica* serotype Enteritidis isolates with three distinct subtyping methods (PFGE, MLVA, and CRISPR-MVLST) allowed a comparison of their relative subtyping efficacies, which were benchmarked by WGST and evaluated by three criteria: (i) discriminatory power, (ii) delineation of outbreak clusters, and (iii) phylogenetic concordance with WGST.

A total of 8, 18, 16, and 34 subtypes were identified from the 52 isolates by PFGE, MLVA, CRISPR-MVLST, and WGST, respectively, resulting in their respective discriminatory powers of 0.81, 0.92, 0.93, and 0.97.

Each of the 16 outbreak clusters was unequivocally identified by WGST; isolates from each outbreak formed distinct clades (Table 2 and Fig. 2). MLVA resolved six outbreak clusters (outbreaks C, D, F, G, I, and L), CRISPR-MLVST identified three (F, G, and L), and PFGE differentiated three (B, C, and G). For another eight outbreak clusters, MLVA was able to cluster the corresponding isolates, but the clusters did not definitively exclude other isolates.

Similarly, 9 and 12 outbreaks were inconclusively clustered by CRISPR-MLVST and PFGE, respectively. Isolates from two outbreak clusters, four outbreak clusters, and one outbreak cluster failed to cluster by MLVA, CRISPR-MLVST, and PFGE, respectively (Table 2 and Fig. 2).

While CRISPR-MVLST, MLVA, and PFGE are not intended for phylogenetic inference, CRISPR-MVLST correctly identified all four major lineages defined by WGST (Fig. 2).

DISCUSSION

The exceptional performance of WGST in the fine-scale delineation of outbreaks of infectious disease has been demonstrated in recent investigations (24–31). In the current study, we expanded the evaluation of WGST by retrospectively investigating isolates from 15 recent *S. enterica* serotype Enteritidis outbreaks in the United States and 1 in Mauritius. This collection of isolates represents the known phylogenetic diversity and epidemiological prevalence of commonly circulating *S. enterica* serotype Enteritidis lineages in the United States in recent years as previously surveyed in reference 15, therefore providing a realistic assessment of WGST in discriminating this otherwise difficult-to-subtype pathogen. With the exception of 2009K-1545 (discussed below), WGST was able to unequivocally discriminate each particular outbreak cluster by exclusively assigning outbreak isolates to it.

Three sporadic strains (2012K-0619, 2012K-0738, and 2012K-0579) were isolated during a multistate outbreak linked to ground beef in 2012 and found to display a PFGE pattern indistinguishable from that of the outbreak strain. Both MLVA and CRISPR-MVLST separated them from temporally related outbreak D isolates (Fig. 2), and WGST was further able to identify these isolates as epidemiologically unrelated to this and any outbreak as well as to each other as shown in Fig. 1.

WGST also indicated that outbreak M might have been polyclonal (i.e., that multiple strains might have been involved in the same outbreak), as a previously identified outbreak isolate (2009K-1545) fell outside the major outbreak cluster, which was also shown by the CRISPR-MVLST result (Fig. 2; see also Table S2 in the supplemental material). Interestingly, WGST suggested that 2009K-1545 was phylogenetically close to outbreak N, which was temporally and geographically related to outbreak M (Pennsylvania, 2009 to 2010). Therefore, some isolates from the two outbreaks may have originated from a recent common ancestor, which is consistent with the fact that the patterns of the outbreak M and N isolates were indistinguishable by MLVA and PFGE (Fig. 2). Together, these results suggest that WGST makes a superior subtyping tool that can reliably define *S. enterica* serotype Enteritidis outbreak clusters in the epidemiological setting of recent *S. enterica* serotype Enteritidis outbreaks in the United States.

The ability of WGST to concurrently provide superior discriminatory power and accurate phylogenetic inferences has the potential to bridge outbreak investigations with long-term and large-scale epidemiological studies. WGST defines outbreaks by resolving phylogenetic relationships rather than by targeting hypervariable but phylogenetically uninformative markers. This provides information regarding the evolutionary dynamics and population structure of the pathogen, which, in turn, can help increase understanding of the patterns and trends of its distribution and infection. To further demonstrate the robustness of WGST in delineating outbreak clusters among closely related *S. enterica* serotype Enteritidis isolates, including those of the same

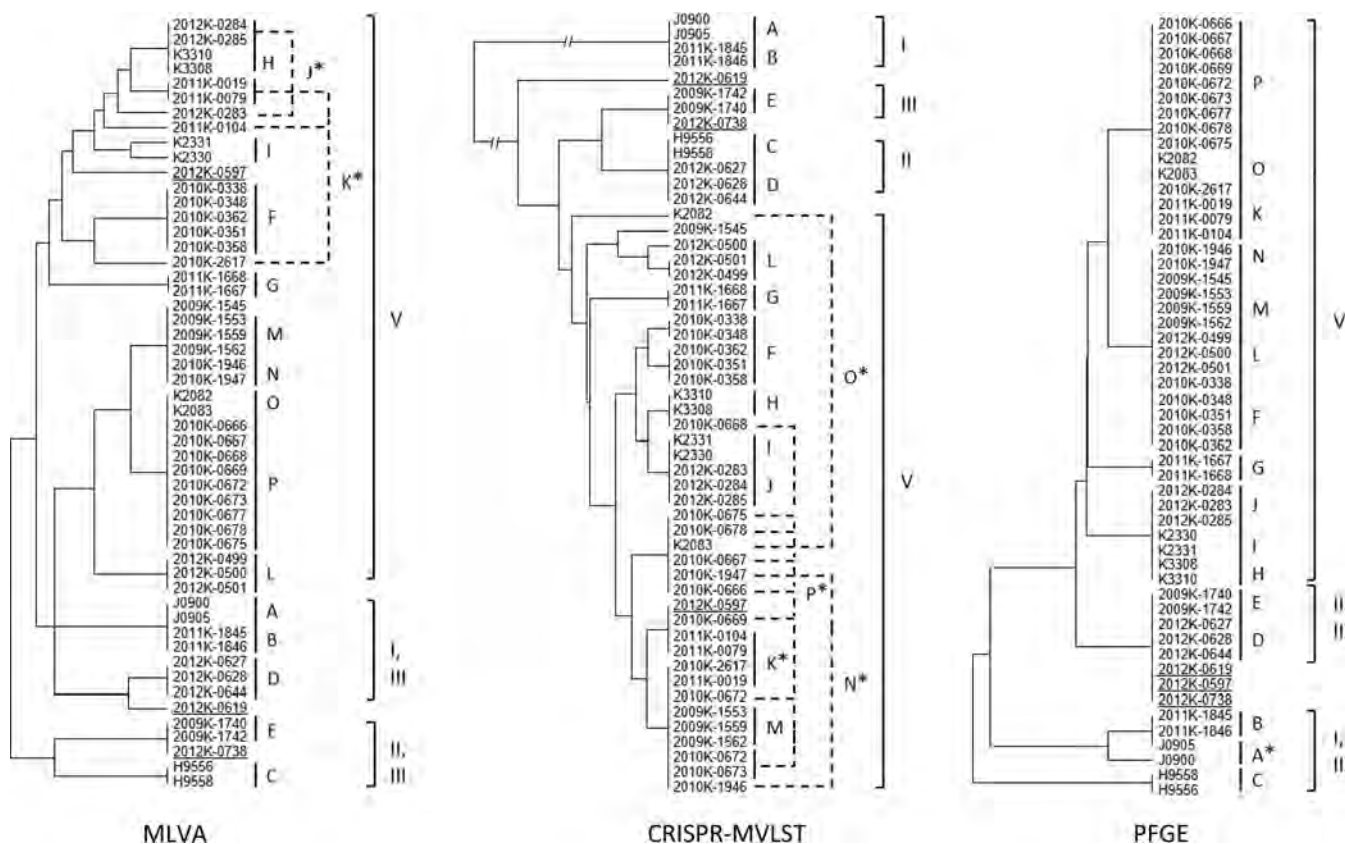


FIG 2 Clustering of outbreak isolates by MLVA, CRISPR-MVLST, and PFGE. Outbreaks are labeled A through P according to Table 1 data. Outbreaks that included isolates not clustered together are labeled with a single asterisk (*) and indicated by dashed lines. The designations of three isolates from sporadic cases (2012K-0619, 2012K-0738, and 2012K-0597) are underlined. The lineages to which each isolate belonged are also labeled. These dendrograms are intended to show the hierarchical clustering of isolates, and their branch lengths are not comparable between the different methods.

PFGE patterns, we included a total of 125 previously sequenced *S. enterica* serotype Enteritidis isolates from a phylogenetic and epidemiologic survey of this serotype (15). As shown in Fig. S1 in the supplemental material, all the outbreak isolates analyzed in the current study formed distinct clusters (highlighted in red in Fig. S1) consistent with their epidemiological information in the background of the additional isolates, including the ones with the same PFGE patterns as some of the outbreak isolates (highlighted in blue in Fig. S1). Only one previously sequenced isolate (02-2966) grouped within an outbreak cluster (outbreak K, a 2011 multistate outbreak associated with guinea pigs). Isolate 2-2966 was collected from a rodent in California in 2002, with its PFGE pattern unknown.

The isolates investigated in the current study fell within four of the five previously defined lineages (15) when they were incorporated into the previous phylogeny (see Fig. S1 in the supplemental material). Furthermore, the epidemiological information and phylogenetic distribution for the newly sequenced strains corresponded with the geographic characteristics and observed prevalence of the five lineages. Specifically, the isolates from outbreaks A and B were from California and Texas, respectively, which is consistent with their clustering in a major clade of lineage I mainly consisting of western American isolates (15). J0900 and J0905 were isolated from environmental sources, similarly to other isolates in this clade that were predominately associated with envi-

ronmental origins. Whereas none of the American isolates surveyed in the current study clustered in lineage III, a lineage characteristic of its international spread, the majority of them (39 of 52 isolates; 10 of the 16 outbreak clusters) were found in lineage V, a typical domestic lineage often associated with poultry products. Lineage IV was represented by only one isolate and was considered to be rare or undersampled in the previous study (15). It remained the least sampled lineage in the current study, with three isolates (2012K-0738, 2009K-1740, and 2012K-1742). All the isolates identified in lineage IV so far were isolated in Maryland. Interestingly, less-sampled lineage II, which was previously recognized as a population associated with marine mammals in California, was found to also include isolates from outbreak and sporadic cases widespread on the west coast (California; outbreak C), east coast (Vermont; outbreak D), and Gulf coast (Texas; 2012K-0619). It was hypothesized that free-ranging and migratory marine mammals and the birds that share their habitats could potentially play a role in long-distance dispersal of this pathogen (15). While CRISPR-MVLST was able to delineate major lineages, it was not possible to reveal such patterns by MLVA and PFGE.

Most comparative studies of different molecular subtyping schemes have focused on performance parameters such as discriminatory power and subtype correlation (32, 33). This approach is sometimes confounded by the limited resolution of common subtyping markers and/or lack of coherence between

them, as dictated by their inherent differences in mutation rates and evolutionary history. PFGE has been used as a standard to facilitate comparison, but it is not ideal, especially for genetically homogenous organisms such as *S. enterica* serotype Enteritidis. Using WGS to benchmark molecular subtyping enables more-rigorous evaluation by interrogating subtypes defined by particular methods with unparalleled resolution and superior phylogenetic accuracy. In the present study, the side-by-side comparison of commonly used and recently developed subtyping methods for *S. enterica* serotype Enteritidis was guided by the robust subtyping and phylogenetic inference of WGST. This allowed a thorough evaluation of the relative performances of MLVA and CRISPR-MVLST that was otherwise not possible.

For instance, CRISPR-MVLST was outperformed by MLVA in outbreak cluster delineation but was able to resolve each of the major lineages. We also observed that it was the CRISPR components, rather than the virulence genes, in the CRISPR-MVLST scheme that afforded the differentiation of the lineages (see Table S2 in the supplemental material). Originated from phages and plasmids that might be characteristic of particular environments (34), CRISPRs might capture signals of ecological relevance. Given the dynamic nature of CRISPR loci with respect to spacer acquisition, loss, and duplication, we hypothesized that the identification of major *S. enterica* serotype Enteritidis lineages by CRISPR-MVLST was due to the imprinting of exogenous genetic cues on the CRISPRs that reflect the different ecological origins of major lineages. However, a recent study that included various *Salmonella* serotypes suggested that such signals might not be phylogenetically informative at the species level due to factors such as horizontal gene transfer and acquisition of common CRISPRs by different lineages (35). Further studies are necessary to investigate the robustness and scope of CRISPR subtyping in detecting ecological and evolutionary patterns of *Salmonella* and other organisms.

It is anticipated that WGS will eventually become the new gold standard for microbial pathogen subtyping. Ongoing efforts such as the 100K Genome Project (<http://100kgenome.vetmed.ucdavis.edu/>), GenomeTrakr Network (<http://www.fda.gov/Food/FoodScienceResearch/WholeGenomeSequencingProgramWGS/>), Global Microbial Identifier (<http://www.globalmicrobialidentifier.org/>), and Advanced Molecular Detection (<http://www.cdc.gov/amd/>) are creating a vast resource of microbial genomes and piloting the real-time, WGS-based surveillance of microbial pathogens. Instead of viewing WGS as the ultimate tool that will soon spell the end of other subtyping methods, we recommend using WGS as a comprehensive platform that will provide access to all existing and future genetic markers for subtyping. For example, CRISPRs and virulence genes were retrieved from sequencing data in the present study, and tools using WGS for other subtyping schemes have been developed (36, 37). Additionally, WGS provides a wealth of genomic data for interrogation for additional features beyond phylogenetic analysis, including gene content (e.g., antibiotic resistance and virulence genes), accessory genome changes (e.g., plasmids and genomic islands), and the presence of phenotypically relevant SNPs (e.g., nonsynonymous and regulatory effectors). The incorporation of various subtyping methods into the WGS platform will provide both backward compatibility to existing markers and data and extensibility to newly developed schemes, thus facilitating the evaluation and benchmarking of molecular subtyping.

In this study, we evaluated only Illumina sequencing technology. Comparisons of different sequencing platforms have been reported elsewhere (38, 39). Also, we did not attempt to address the important issue of routine and broad implementation of WGS in clinical and public health applications, which has recently been investigated (40).

ACKNOWLEDGMENTS

We thank Philippe Horvath for developing the Excel macro for visualizing the CRISPR spacers.

This work was supported in part by University of Georgia startup funds to X.D. and a United States Army Research Office grant (W911NF-11-1-0442) to E.G.D.

REFERENCES

- 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.
- Jackson BR, Griffin PM, Cole D, Walsh KA, Chai SJ. 2013. Outbreak-associated *Salmonella enterica* serotypes and food commodities, United States, 1998–2008. *Emerg Infect Dis* 19:1239–1244. <http://dx.doi.org/10.3201/eid1908.121511>.
- Centers for Disease Control and Prevention. 2013. Reports of Selected *Salmonella* Outbreak Investigations. Centers for Disease Control and Prevention, Atlanta, GA. <http://www.cdc.gov/salmonella/outbreaks.html>.
- Olson AB, Andrysiak AK, Tracz DM, Guard-Bouldin J, Demczuk W, Ng LK, Maki A, Jamieson F, Gilmour MW. 2007. Limited genetic diversity in *Salmonella enterica* serovar Enteritidis PT13. *BMC Microbiol* 7:87. <http://dx.doi.org/10.1186/1471-2180-7-87>.
- Zheng J, Keys CE, Zhao S, Meng J, Brown EW. 2007. Enhanced subtyping scheme for *Salmonella enteritidis*. *Emerg Infect Dis* 13:1932–1935. <http://dx.doi.org/10.3201/eid1312.070185>.
- Swaminathan B, Barrett TJ, Hunter SB, Tauxe RV, CDC PulseNet Task Force. 2001. PulseNet: the molecular subtyping network for foodborne bacterial disease surveillance, United States. *Emerg Infect Dis* 7:382–389. <http://dx.doi.org/10.3201/eid0703.017303>.
- Boxrud D, Pederson-Gulrud K, Wotton J, Medus C, Lyszkowicz E, Besser J, Bartkus JM. 2007. Comparison of multiple-locus variable-number tandem repeat analysis, pulsed-field gel electrophoresis, and phage typing for subtype analysis of *Salmonella enterica* serotype Enteritidis. *J Clin Microbiol* 45:536–543. <http://dx.doi.org/10.1128/JCM.01595-06>.
- Cho S, Boxrud DJ, Bartkus JM, Whittam TS, Saeed M. 2007. Multiple-locus variable-number tandem repeat analysis of *Salmonella enteritidis* isolates from human and non-human sources using a single multiplex PCR. *FEMS Microbiol Lett* 275:16–23. <http://dx.doi.org/10.1111/j.1574-6968.2007.00875.x>.
- Liu F, Barrangou R, Gerner-Smidt P, Ribot EM, Knabel SJ, Dudley EG. 2011. Novel virulence gene and clustered regularly interspaced short palindromic repeat (CRISPR) multilocus sequence typing scheme for subtyping of the major serovars of *Salmonella enterica* subsp. *enterica*. *Appl Environ Microbiol* 77:1946–1956. <http://dx.doi.org/10.1128/AEM.02625-10>.
- Shariat N, DiMarzio MJ, Yin S, Dettinger L, Sandt CH, Lute JR, Barrangou R, Dudley EG. 2013. The combination of CRISPR-MVLST and PFGE provides increased discriminatory power for differentiating human clinical isolates of *Salmonella enterica* subsp. *enterica* serovar Enteritidis. *Food Microbiology* 34:164–173. <http://dx.doi.org/10.1016/j.fm.2012.11.012>.
- Dewaele I, Rasschaert G, Bertrand S, Wildemaue C, Wattiau P, Imberechts H, Herman L, Ducatelle R, De Reu K, Heyndrickx M. 2012. Molecular characterization of *Salmonella enteritidis*: comparison of an optimized multi-locus variable-number of tandem repeat analysis (MLVA) and pulsed-field gel electrophoresis. *Foodborne Pathog Dis* 9:885–895. <http://dx.doi.org/10.1089/fpd.2012.1199>.
- Shariat N, Kirchner MK, Sandt CH, Trees E, Barrangou R, Dudley EG. 2013. Subtyping of *Salmonella enterica* serovar Newport outbreak isolates by CRISPR-MVLST and determination of the relationship between CRISPR-MVLST and PFGE results. *J Clin Microbiol* 51:2328–2336. <http://dx.doi.org/10.1128/JCM.00608-13>.

13. Shariat N, Sandt CH, DiMarzio MJ, Barrangou R, Dudley EG. 2013. CRISPR-MVLST subtyping of *Salmonella enterica* subsp. *enterica* serovars Typhimurium and Heidelberg and application in identifying outbreak isolates. *BMC Microbiol* 13:254. <http://dx.doi.org/10.1186/1471-2180-13-254>.
14. Allard MW, Luo Y, Strain E, Pettengill J, Timme R, Wang C, Li C, Keys CE, Zheng J, Stones R, Wilson MR, Musser SM, Brown EW. 2013. On the evolutionary history, population genetics and diversity among isolates of *Salmonella enteritidis* PFGE pattern JEGX01.0004. *PLoS One* 8:e55254. <http://dx.doi.org/10.1371/journal.pone.0055254>.
15. Deng X, Desai PT, den Bakker HC, Mikoleit M, Tolar B, Trees E, Hendriksen RS, Frye JG, Porwollik S, Weimer BC, Wiedmann M, Weinstock GM, Fields PI, McClelland M. 2014. Genomics epidemiology of *Salmonella enterica* serotype Enteritidis based on population structure of prevalent lineages. *Emerg Infect Dis* 20:1481–1489. <http://dx.doi.org/10.3201/eid2009.131095>.
16. Etienne KA, Gillice J, Hilsabeck R, Schupp JM, Colman R, Lockhart SR, Gade L, Thompson EH, Sutton DA, Neblett-Fanfair R, Park BJ, Turabelidze G, Keim P, Brandt ME, Deak E, Engelthaler DM. 2012. Whole genome sequence typing to investigate the Apophysomyces outbreak following a tornado in Joplin, Missouri, 2011. *PLoS One* 7:e49989. <http://dx.doi.org/10.1371/journal.pone.0049989>.
17. Gillice JD, Schupp JM, Balajee SA, Harris J, Pearson T, Yan Y, Keim P, DeBess E, Marsden-Haug N, Wohlrle R, Engelthaler DM, Lockhart SR. 2011. Whole genome sequence analysis of *Cryptococcus gattii* from the Pacific Northwest reveals unexpected diversity. *PLoS One* 6:e28550. <http://dx.doi.org/10.1371/journal.pone.0028550>.
18. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol* 28:2731–2739. <http://dx.doi.org/10.1093/molbev/msr121>.
19. Ribot EM, Fair MA, Gautom R, Cameron DN, Hunter SB, Swaminathan B, Barrett TJ. 2006. Standardization of pulsed-field gel electrophoresis protocols for the subtyping of *Escherichia coli* O157:H7, *Salmonella*, and *Shigella* for PulseNet. *Foodborne Pathog Dis* 3:59–67. <http://dx.doi.org/10.1089/fpd.2006.3.59>.
20. Zerbino DR, Birney E. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* 18:821–829. <http://dx.doi.org/10.1101/gr.074492.107>.
21. Huson DH, Bryant D. 2006. Application of phylogenetic networks in evolutionary studies. *Mol Biol Evol* 23:254–267. <http://dx.doi.org/10.1093/molbev/msj030>.
22. Hunter PR, Gaston MA. 1988. Numerical index of the discriminatory ability of typing systems: an application of Simpson's index of diversity. *J Clin Microbiol* 26:2465–2466.
23. Liu F, Kariyawasam S, Jayarao BM, Barrangou R, Gerner-Smidt P, Ribot EM, Knebel SJ, Dudley EG. 2011. Subtyping *Salmonella enterica* serovar Enteritidis isolates from different sources by using sequence typing based on virulence genes and clustered regularly interspaced short palindromic repeats (CRISPRs). *Appl Environ Microbiol* 77:4520–4526. <http://dx.doi.org/10.1128/AEM.00468-11>.
24. Rasko DA, Webster DR, Sahl JW, Bashir A, Boisen N, Scheutz F, Paxinos EE, Sebra R, Chin CS, Iliopoulos D, Klammer A, Peluso P, Lee L, Kislyuk AO, Bullard J, Kasarskis A, Wang S, Eid J, Rank D, Redman JC, Steyert SR, Frimodt-Moller J, Struve C, Petersen AM, Krogfelt KA, Nataro JP, Schadt EE, Waldor MK. 2011. Origins of the *E. coli* strain causing an outbreak of hemolytic-uremic syndrome in Germany. *N Engl J Med* 365:709–717. <http://dx.doi.org/10.1056/NEJMoa1106920>.
25. Mellmann A, Harmsen D, Cummings CA, Zentz EB, Leopold SR, Rico A, Prior K, Szczepanowski R, Ji Y, Zhang W, McLaughlin SF, Henkhaus JK, Leopold B, Bielaszewska M, Prager R, Brzoska PM, Moore RL, Guenther S, Rothberg JM, Karch H. 2011. Prospective genomic characterization of the German enterohemorrhagic *Escherichia coli* O104:H4 outbreak by rapid next generation sequencing technology. *PLoS One* 6:e22751. <http://dx.doi.org/10.1371/journal.pone.0022751>.
26. Hendriksen RS, Price LB, Schupp JM, Gillice JD, Kaas RS, Engelthaler DM, Bortolaia V, Pearson T, Waters AE, Upadhyay BP, Shrestha SD, Adhikari S, Shakya G, Keim PS, Aarestrup FM. 2011. Population genetics of *Vibrio cholerae* from Nepal in 2010: evidence on the origin of the Haitian outbreak. *mBio* 2:e00157-00111. <http://dx.doi.org/10.1128/mBio.00157-11>.
27. Chin CS, Sorenson J, Harris JB, Robins WP, Charles RC, Jean-Charles RR, Bullard J, Webster DR, Kasarskis A, Peluso P, Paxinos EE, Yamai-chi Y, Calderwood SB, Mekalanos JJ, Schadt EE, Waldor MK. 2011. The origin of the Haitian cholera outbreak strain. *N Engl J Med* 364:33–42. <http://dx.doi.org/10.1056/NEJMoa1012928>.
28. Bakker HC, Switt AI, Cummings CA, Hoelzer K, Degoricija L, Rodriguez-Rivera LD, Wright EM, Fang R, Davis M, Root T, Schoonmaker-Bopp D, Musser KA, Villamil E, Waechter H, Kornstein L, Furtado MR, Wiedmann M. 2011. A whole-genome single nucleotide polymorphism-based approach to trace and identify outbreaks linked to a common *Salmonella enterica* subsp. *enterica* serovar Montevideo pulsed-field gel electrophoresis type. *Appl Environ Microbiol* 77:8648–8655. <http://dx.doi.org/10.1128/AEM.06538-11>.
29. Lienau EK, Strain E, Wang C, Zheng J, Ottesen AR, Keys CE, Hammack TS, Musser SM, Brown EW, Allard MW, Cao G, Meng J, Stones R. 2011. Identification of a salmonellosis outbreak by means of molecular sequencing. *N Engl J Med* 364:981–982. <http://dx.doi.org/10.1056/NEJMc1100443>.
30. Engelthaler DM, Chiller T, Schupp JA, Colvin J, Beckstrom-Sternberg SM, Driebe M, Moses T, Tembe W, Sinari S, Beckstrom-Sternberg JS, Christoforides A, Pearson JV, Carpten J, Keim P, Peterson A, Terashita D, Balajee SA. 2011. Next-generation sequencing of *Coccidioides immitis* isolated during cluster investigation. *Emerg Infect Dis* 17:227–232. <http://dx.doi.org/10.3201/eid1702.100620>.
31. Harris SR, Feil EJ, Holden MT, Quail MA, Nickerson EK, Chantratita N, Gardete S, Tavares A, Day N, Lindsay JA, Edgeworth JD, de Lencastre H, Parkhill J, Peacock SJ, Bentley SD. 2010. Evolution of MRSA during hospital transmission and intercontinental spread. *Science* 327:469–474. <http://dx.doi.org/10.1126/science.1182395>.
32. Foley SL, White DG, McDermott PF, Walker RD, Rhodes B, Fedorka-Cray PJ, Simjee S, Zhao S. 2006. Comparison of subtyping methods for differentiating *Salmonella enterica* serovar Typhimurium isolates obtained from food animal sources. *J Clin Microbiol* 44:3569–3577. <http://dx.doi.org/10.1128/JCM.00745-06>.
33. Hyeon JY, Chon JW, Park JH, Kim MS, Oh YH, Choi IS, Seo KH. 2013. A comparison of subtyping methods for differentiating *Salmonella enterica* serovar Enteritidis isolates obtained from food and human sources. *Osong Public Health Res Perspect* 4:27–33. <http://dx.doi.org/10.1016/j.phrp.2012.12.005>.
34. Kunin V, He S, Warnecke F, Peterson SB, Garcia Martin H, Haynes M, Ivanova N, Blackall LL, Breitbart M, Rohwer F, McMahon KD, Hugenholtz P. 2008. A bacterial metapopulation adapts locally to phage predation despite global dispersal. *Genome Res* 18:293–297. <http://dx.doi.org/10.1101/gr.6835308>.
35. Timme RE, Pettengill JB, Allard MW, Strain E, Barrangou R, Wehnes C, Van Kessel JS, Karns JS, Musser SM, Brown EW. 2013. Phylogenetic diversity of the enteric pathogen *Salmonella enterica* subsp. *enterica* inferred from genome-wide reference-free SNP characters. *Genome Biol Evol* 5:2109–2123. <http://dx.doi.org/10.1093/gbe/evt159>.
36. Larsen MV, Cosentino S, Rasmussen S, Friis C, Hasman H, Marvig RL, Jelsbak L, Sicheritz-Ponten T, Ussery DW, Aarestrup FM, Lund O. 2012. Multilocus sequence typing of total-genome-sequenced bacteria. *J Clin Microbiol* 50:1355–1361. <http://dx.doi.org/10.1128/JCM.06094-11>.
37. Inouye M, Conway TC, Zobel J, Holt KE. 2012. Short read sequence typing (SRST): multi-locus sequence types from short reads. *BMC Genomics* 13:338. <http://dx.doi.org/10.1186/1471-2164-13-338>.
38. Loman NJ, Misra RV, Dallman TJ, Constantinidou C, Gharbia SE, Wain J, Pallen MJ. 2012. Performance comparison of benchtop high-throughput sequencing platforms. *Nat Biotechnol* 30:434–439. <http://dx.doi.org/10.1038/nbt.2198>.
39. Jünemann S, Sedlazeck FJ, Prior K, Albersmeier A, John U, Kalinowski J, Mellmann A, Goesmann A, von Haeseler A, Stoye J, Harmsen D. 2013. Updating benchtop sequencing performance comparison. *Nat Biotechnol* 31:294–296. <http://dx.doi.org/10.1038/nbt.2522>.
40. Joensen KG, Scheutz F, Lund O, Hasman H, Kaas RS, Nielsen EM, Aarestrup FM. 2014. Real-time whole-genome sequencing for routine typing, surveillance, and outbreak detection of verotoxigenic *Escherichia coli*. *J Clin Microbiol* 52:1501–1510. <http://dx.doi.org/10.1128/JCM.03617-13>.