JCM Accepted Manuscript Posted Online 11 March 2015
J. Clin. Microbiol. doi:10.1128/JCM.00323-15

TITLE: *Salmonella* Serotype Determination Utilizing High-throughput Genome Sequencing Data

Shaokang Zhang[1], Yanlong Yin[2]*, Marcus B. Jones[3], Zhenzhen Zhang[4], Brooke L. Deatherage Kaiser[5], Blake A. Dinsmore[6], Collette Fitzgerald[6], Patricia I. Fields[6], Xiangyu Deng[1]#

[1]Center for Food Safety, Department of Food Science and Technology, University of Georgia, 1109 Experiment Street, Griffin, Georgia 30223

[2]Department of Computer Science, Illinois Institute of Technology, Chicago, IL

[3]Department of Infectious Diseases, J. Craig Venter Institute, Rockville, MD

[4]Department of Biostatistics, School of Public Health, University of Michigan, Ann Arbor, MI

[5]Biological Sciences Division, Pacific Northwest National Laboratory, Richard, WA

[6]Division of Foodborne, Waterborne and Environmental Diseases, Centers for Disease Control and Prevention, MS-C03, 1600 Clifton Road, Atlanta, GA 30333

RUNNING HEAD: *Salmonella* serotyping from WGS

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

*Current address: Bloomberg L.P., New York City, NY, 10022

33 **ABSTRACT**

34 Serotyping forms the basis of national and international surveillance networks for
35 *Salmonella,* one of the most prevalent foodborne pathogens worldwide (1-3). Public
36 health microbiology is currently being transformed by whole genome sequencing (WGS)
37 which opens the door to serotype determination using WGS data. SeqSero
38 (www.denglab.info/SeqSero) is a novel web-based tool for determining *Salmonella*
39 serotypes using high-throughput genome sequencing data. SeqSero is based on curated
40 databases of *Salmonella* serotype determinants (*rfb* gene cluster, *fliC* and *fljB* alleles) and
41 is predicted to determine serotype rapidly and accurately for nearly the full spectrum of
42 *Salmonella* serotypes (more than 2,300 serotypes), from both raw sequencing reads and
43 genome assemblies. The performance of SeqSero was evaluated by testing: 1) raw reads
44 from genomes of 308 *Salmonella* isolates of known serotype; 2) raw reads from genomes
45 of 3,306 *Salmonella* isolates sequenced and made publicly available by GenomeTrakr, a
46 U.S. national monitoring network operated by the Food and Drug Administration; and 3)
47 354 other publicly available draft or complete *Salmonella* genomes. We also
48 demonstrated *Salmonella* serotype determination from raw sequencing reads of fecal
49 metagenomes from mice orally infected with this pathogen. SeqSero can help to maintain
50 the well-established utility of *Salmonella* serotyping when integrated into a platform of
51 WGS-based pathogen subtyping and characterization.

52

53

54

55

Journal of Clinical Microbiology

INTRODUCTION

Salmonella is the most prevalent foodborne pathogen in the United States, causing 1.2 million cases of illness annually and the largest health burden among all bacterial pathogens (4). The U.S. National *Salmonella* Surveillance System has been built upon serotyping in public health laboratories, a subtyping method traditionally performed through the agglutination of *Salmonella* cells with specific antisera that detects lipopolysaccharide O antigen and flagellar H antigens. Specific combinations of O and H antigenic types represent serotypes (or serovars). More than 2,500 *Salmonella* serotypes have been described in the White-Kauffmann-Le Minor scheme (5, 6). The phenotypic determination of serotypes is labor intensive and time-consuming (takes at least 2 days), which has led to the development of genetic methods for serotype determination (7, 8), These methods generally use two categories of targets for serotype determination: 1) indirectly using random surrogate genomic markers associated with particular serotypes; and 2) directly, using genetic determinants of serotype, including the *rfb* gene cluster responsible for somatic (O) group synthesis (9, 10) and *fliC (11)* and *fljB* (12) genes encoding the two flagellar antigens present in *Salmonella*. The latter approach has the advantage of determining serotype using the same markers as the phenotypic method, providing continuity between serotypes determined by phenotypic and genetic markers (13, 14). While this approach may result in distinct genetic lineages being assigned the same serotype due to horizontal gene transfer of the serotype determinants, phylogenetic reconstruction is beyond the scope of serotyping and can be better performed by other subtyping methods. Also, through the identification of individual serotype determinants,

79 methods based on serotype determinants have the potential to predict a wide range of
80 *Salmonella* serotypes. By contrast, methods based on random surrogate genomic markers
81 rely on the presumed correspondence between the markers and particular serotypes and
82 therefore need to be validated for each new serotype tested.
83
84 Routine and real-time implementation of whole genome sequencing (WGS) (15, 16) is
85 poised to transform public health microbiology. Efforts have been made to enable a
86 variety of pathogen subtyping and characterization analyses through WGS data, such as
87 multi-locus sequence typing (17, 18), antimicrobial resistance identification (19) and
88 virulence characterization (16). Beyond WGS of pure cultures, recent application of
89 metagenome sequencing in diagnosis and outbreak investigation of infectious diseases
90 (20, 21) has demonstrated the potential for culture-independent detection of pathogens
91 from complex clinical samples.
92
93 Here we present a novel application of whole genome and metagenome sequence data for
94 *Salmonella* serotype determination. Curated databases for major serotype determinants
95 were constructed: the *rfb* gene clusters responsible for somatic O group antigen synthesis
96 (22); O-antigen flipase gene *wzx* and the O-antigen polymerase gene *wzy* which are
97 typically found in the *rfb* cluster and are highly specific for the majority of O groups (23);
98 additional genes from the *rfb* cluster useful for characterization of specific O groups; and,
99 *fliC* and *fljB* genes that encode *Salmonella* flagellar antigens. Based on mapping raw
100 sequencing reads to these databases for the identification of individual antigen types, our
101 bioinformatics approach allows robust and comprehensive prediction of *Salmonella*

serotype without genome assembly. A web application of our serotyping tool (named SeqSero) is publicly available at www.denglab.info/SeqSero.

MATERIALS AND METHODS

Whole genome sequences. A total of 229 *Salmonella enterica* isolates of various relatively uncommon serotypes (Table S1) were sequenced on an Illumina HiSeq 2000 platform (100 bp, paired-end reads) per manufacturer's instructions by the 100K Foodborne Pathogen Genome Project at University of California, Davis (http://100kgenome.vetmed.ucdavis.edu/). An additional 79 *Salmonella* genomes representing common serotypes from the WGS collection of CDC (NCBI BioProject PRJNA186441) were included, for a total of 308 genomes in the CDC strain set. The serotypes of these isolates were confirmed using traditional (24) and genetic (13, 14) serotyping assays. For the GenomeTrakr strain set, *Salmonella* genomes sequenced by the Illumina platform and uploaded to the GenomeTrakr depository (NCBI BioProject 183844) as of June 1st, 2014 were reviewed for suitability for inclusion in a validation dataset. Genomes were excluded for the following reasons: 1) no serotype or more than one serotypes indicated for a specific genome (n=766); 2) rough, nonmotile strains (n=39); 3) monophasic variants (n=76); and, 4) less than 10x sequencing coverage (n=11). A total of 354 assembled genomes with N50 contig size >150,000 bases were downloaded from GenBank for validation analysis.

Mouse infections, fecal sample preparation and metagenome sequencing. Mouse infections, fecal sample preparation and DNA extraction were performed as previously

described (25). *S. enterica* serotype Typhimurium strain 14028s was used to orally challenge female, age-matched (6-8 weeks) 129SvJ mice (25). Fecal samples from control mice had not been sequenced and were not available for the current study. For deep metagenomic sequencing, extracted DNA were bar-coded, multiplexed and sequenced using the Illumina V3 chemistry on the HiSeq 2000 platform. We implemented automation for the construction of up to 96 fragment or paired end libraries at one time. Paired-end libraries were constructed using the Illumina TruSeq protocol. Approximately 1 Gb of shotgun sequence data per sample were generated.

Databases for *Salmonella* serotype determinants. For O group determination, two databases were built: 1) sequences from the entire *rfb* cluster were used for O group determination from genome assemblies; and, 2) *wzx* (O-antigen flippase), *wzy* (O-antigen polymerase), and other genes or markers from the *rfb* cluster useful for O group determination (Table S4) were used when the input data was raw sequencing reads. Two O antigen groups, those that possess O9 (O9, O2, O9, 46 and O9, 46, 27) and those that possess O3 (O3, 10 and O1, 3, 19), require additional markers for differentiation including the *rfb* sequence specific to O3, 10 and a frame shift mutation in *tyv* (Table S4). The combined use of the six markers allowed the differentiation of 273 O3, 10 and 72 O1, 3, 19 strains (data not shown). In the two O group databases, each of the 46 O antigens was represented by a single *rfb* cluster (26) or a single allele of *wzx* or *wzy* gene (27).

For H antigen determination, a single database that contained both *fliC* and *fljB* alleles was built; the sequences were primarily from (28) and supplemented with *fliC* and *fljB*

genes extracted from *Salmonella* genomes (closed and draft assemblies) available at GenBank. Multiple, distinct alleles for the same flagellar antigenic type were allowed to accommodate the multiphyletic nature of some H antigens (28).

For the multiple rounds of reads mapping for H antigen determination, three additional datasets were developed. 1) *fliC* and *fljB* alleles were grouped into clusters based on sequence similarity (Table S5). This grouping was used to identify the mostly likely H antigen group after the first two rounds of reads mapping (see details below). 2) A representative allele for each H antigen type was selected and used to extract sequencing reads relevant to H antigens in the third round of reads mapping. This allele was near the midpoint between the root and the tip of longest branch of the phylogenetic tree that contained all the alleles for an antigen. 3) For H antigen clusters that had multiple antigen types (Table S5) and therefore required a BLAST analysis for final H antigen determination, a database of the middle, variable sequences of the alleles for every antigen in the cluster was used for the BLAST alignment (see details below). All the databases and additional datasets are available at www.denglab.info/SeqSero. They are regularly curated and updated when new sequences become available. Text S1 provides a discussion of considerations for serotype determination in *Salmonella* using the conventions of the White-Kauffmann-Le Minor Scheme.

Serotype prediction from raw sequencing reads. A reads mapping-based strategy was developed for prediction of O and H antigenic types. In general, raw sequencing reads without any quality filtering or trimming were mapped to individual antigen sequence databases using Burrows-Wheeler Aligner (BWA) with the default parameter setting of

171 the sampe/samse algorithm (29). The allele to which the highest number of reads mapped
172 was chosen as the allele potentially present in the genome tested.
173
174 Some *fliC* and *fljB* alleles share high levels of sequence similarity (28), creating
175 challenges for the determination of antigenic type based on DNA sequence. This issue
176 was aggravated in our pipeline because multiple, closely related alleles were present in
177 the database. When the test genome contains a gene for an antigen type that is
178 represented by a single allele in the database, most reads map to that one allele and only a
179 few to other alleles in the database, producing a pronounced difference. When the
180 database contains multiple, closely related alleles, reads can map to multiple alleles,
181 diminishing or even eliminating the otherwise pronounced excess in the number of reads
182 mapped to the allele expected for the genome being tested (Figure S1). To minimize
183 these problems, we implemented a stepwise identification approach using two rounds of
184 reads mapping for all analyses and adding an additional round of mapping plus a
185 subsequent BLAST analysis in cases where multiple antigenic types are present in a
186 predefined H antigen cluster (Table S5).
187
188 An example workflow of *fliC* identification is depicted in Figure 4; a similar workflow is
189 used for *fljB* determination. 1) Round one mapping: The raw sequencing reads of a
190 serotype Typhimurium genome (NCBI SRA accession SRX528051) were mapped to the
191 entire H antigen database. The *fliC* alleles were then ranked according to the number of
192 reads mapped to each allele, from the largest to the smallest. Up to three antigen clusters
193 (Table S5) that contain the highest ranking alleles were selected. In this example, clusters

fliC_eh (including antigenic type "e, h"), fliC_ir (including antigenic types "i", "r" and "r, i") and fliC_z35 (including antigenic type "z35") were selected. 2) Round two mapping: The allele in each cluster that had the most mapped reads was selected (up to three alleles) and reads were mapped to just those three alleles. The alleles were again ranked as described above. In this example, the order of the top ranking clusters changed to cluster fliC_ir, fliC_eh and fliC_z35, suggesting an error caused by the "dilution effect" (Figure S1) between clusters fliC_ir and fliC_eh had been corrected and the antigen of the test genome was determined to belong to cluster fliC_ir. 3) Round three mapping: The representative alleles for the antigenic types in cluster fliC_ir were used in another round of reads mapping to extract relevant reads with homology to the *fliC* locus. 4) BLAST analysis: The extracted reads were aligned using BLAST (30) to a collection of variable regions of the alleles in cluster fliC_ir and a BLAST score was assigned to each read/allele alignment. BLAST scores of all alignments associated with the same allele were summed and the highest score pointed to the most likely allele and its corresponding antigen for the test genome, in this example flagellar antigen "i".

Serotype prediction from genome assembly. For O antigen group determination, the *galF* and *gnd* genes that flank the *rfb* cluster were located by aligning the two genes against a *Salmonella* genome assembly (30). If both genes resided in the same contig, the *rfb* gene cluster between the two loci was extracted. If the two genes fell into two separate contigs, the corresponding contigs were split at *galF* or *gnd* in order to separate the sequence with homology to the *rfb* cluster from flanking sequences, producing four contig fragments. Then the *rfb* cluster or the set of 4 contig fragments that might or might

not contain partial *rfb* cluster was aligned against the *rfb* database using BLAST. The resulting hits were ranked by BLAST scores, with the highest ranking *rfb* hit determining the O antigen group of the genome. For H antigen determination, *fliC* and *fljB* alleles were obtained from a genome assembly by *in silico* PCR (http://hgwdev.cse.ucsc.edu/~kent/src/). Primers used for *in silico* PCR were summarized in Table S7. Since the sequences flanking *fljB* may vary, multiple sets of primers were used to maximize the possibility of obtaining *fljB* amplicons. *In silico* amplicons of *fliC* and *fljB* were aligned against the H antigen database using BLAST and the antigen types were identified similar to the determination of O antigen as described above.

Statistical analysis. We assessed how well we could identify *fliC* and *fljB* antigens using the GenomeTrakr dataset, by calculating the difference between the number of reads (x and y) aligned to the top two best mapped alleles for each of the two genes in the H antigen database. We used logistic regression to estimate the probability of making an incorrect identification given the size of the mapped reads difference ($x - y$). The outcome of the model was a binary indicator of whether the correct H antigen is identified. The covariate was the logarithmic scaled reads difference. We used scaling to account for the fact that the larger number of sequencing reads (z) tend to yield a bigger reads difference. The scaled reads difference (α) was calculated as $\alpha = [(x - y)/z] \times 10^6$.

RESULTS

SeqSero pipeline. The major components and workflows of the SeqSero system are outlined in Figure 1 and detailed in the Method section.

Antigen determinants databases. A total of 473 alleles representing 56 antigenic types for *fliC*, and 190 alleles representing 18 antigenic types for *fljB* were included in a combined H antigen database. A second database consisting of the 46 described *rfb* clusters was used for O group determination from genome assemblies. A third database containing *wzx, wzy* and other targets (Table S4) was used for O group determination from raw sequencing reads (see the Method section for details). The alleles represented in the databases will theoretically identify 2,389 out of the 2,577 serotypes described in the White-Kauffmann-Le minor Scheme.

Serotype prediction from whole genome sequencing. The results of predictions were summarized in Table 1. For raw sequencing reads, two sets of isolates were tested: 1) 308 isolates that were serotyped at CDC and represented 72 serotypes (Table S1); and 2) 3,306 isolates of 228 serotypes sequenced as of June, 2014 by GenomeTrakr of the Food and Drug Administration, a network of state and federal public health laboratories for the monitoring of foodborne pathogens isolated from food; the serotype of the strain was extracted from the meta data deposited with the sequence (Table S2). For genome assemblies, 354 draft or finished genomes of 44 serotypes were tested including all the assemblies deposited in GenBank as of April, 2014 with serotype information available in the associated meta data and an N50 contig size (31) more than 150,000 bases (Table S3). This empirical N50 cutoff was used to exclude poorly assembled genomes.

For the 308 isolates with confirmed serotype, 304 (98.7%) were correctly identified, two produced partial serotype information, and two produced an unexpected serotype result (Table 1). The accuracy of serotype predictions based on annotated serotypes was 92.6% and 91.5% for the GenomeTrakr and assembled genome datasets, respectively.

We analyzed the four WGS from the confirmed serotype dataset that produced a partial or unexpected serotype result in order to determine whether the result pointed to a problem in the SeqSero pipeline. Two of the six serotype Hvittingfoss (antigenic formula I 16:b:e,n,x) genomes tested failed to generate O antigen calls resulting in a partial serotype. Those genomes lacked sequencing reads that mapped to any *rfb* cluster including the *wzx/wzy* genes. One of the three serotype London (antigenic formula I 3,10:l,v:1,6) genomes produced a *fljB* determination of "e,n,x" instead of the expected "1,6"; reads that could be assembled into both "1, 6" and "e, n, x" alleles were found. One of the five serotype Weltevreden (antigenic formula I 3,10:r:z6) genomes produced a *fliC* determination "i" instead of the expected "r" allele. Again, reads to both "r" and "i" were found to be present in the WGS.

Together, 200 serotypes were successfully predicted from the three datasets (Table S6), including 85 out of the top 100 most commonly reported *Salmonella* serotypes from human infections o the U.S. national *Salmonella* surveillance (http://www.cdc.gov/nationalsurveillance/salmonella-surveillance.html).

Robustness of H antigen identification by reads mapping. The phenotypic nature of serotyping and the diversity of *Salmonella* flagellar antigens make it difficult to map specific antigen types to individual genotypes or sequence variations (e.g., point mutations, insertions and deletions). Closely related H antigens such as the G complex

and 1 complex (11) constitute a particular challenge for robust identification of antigenic type based on sequence comparison. We defined and calculated median scaled reads difference (see the Method section for details) to evaluate how well we can use reads mapping to identify H antigens of the genomes in the GenomeTrakr dataset. The median scaled reads difference was 3.59 for *fliC*, and 1.82 for *fljB*, corresponding to a predicted probability of incorrect antigen call at 2.7% and 5.6% respectively (Figure 2). These results suggested that our method of H antigen identification based on reads mapping was robust. It should be noted that the statistical modeling was based on the results after only the first round of reads mapping (Figure 4); therefore, it included errors that might later be corrected by the subsequent mapping and BLAST analyses.

Serotype prediction from metagenome sequencing. Serotype Typhimurium was detected in metagenomes of mouse stool samples 1 day before and 3, 6 and 14 days after the oral infection of *S. enterica* serotype Typhimurium strain 14028s (Table 2). A small number of reads were mapped to the serotype determinants on day -1, far fewer than later samples (Table 2). The strain detected on day -1 appeared to be phylogenetically distinct from the strain used for infection and serotyped on Day 3, 6 and 14 (Figure 3).

We also tested metagenome sequencing reads from a study to detect *Salmonella* from tomato phyllosphere microbiome (32); we did not find any *Salmonella* serotype markers, likely due to the low abundance of *Salmonella* in those samples and consistent with the fact that no *Salmonella* was detected in this study using RT-PCR or culture methods.

To test whether *Escherichia coli* DNA might produce a false positive signal in metagenomic samples, we tested metagenome sequences from 45 fecal specimens from

Journal of Clinical Microbiology

patients involved in the 2011 outbreak of Shiga-toxigenic *E. coli* (STEC) O104:H4 in Germany (21). No reads from any of the metagenomes mapped to any allele in the *Salmonella* antigen databases.

DISCUSSION

The bioinformatics pipeline reported here determined *Salmonella* serotype directly from raw sequencing reads or assembled genomes. The O group is determined primarily by *wzx* and *wzy* sequences for raw reads and by the *rfb* cluster for assembled genomes. Both H phases are determined through *fliC* and *fljB* sequences combined in the same H antigen database. Serotype determination from raw reads is recommended for high throughput sequencing technologies that generate short reads, such as Illumina. Using raw sequencing reads reduces analysis time and allows serotype determination from raw data without the need for high quality genome assembly and subsequent extraction of serotype determinants. With computing capacity of 4 CPU cores and 4 GB RAM, the serotype predictions of most isolates from raw WGS reads (an average of 2.17 million reads per genome) were finished within 10 minutes.

SeqSero proved accurate in determining serotype using genomes from strains in the CDC collection, which represented most of the 100 most common serotypes identified in the US (Table 1). An O group was not identified for two isolates because no reads with homology to the entire or vast majority (the first 11,325 bases of the 12,901 bases O16 *rfb*) of *rfb* cluster were present in the WGS. Since an O group was detected in these strains using conventional methods, the *rfb* cluster is presumably present in those strains;

we are currently investigating why no sequence reads were generated. Two additional isolates were not identified as the expected serotype due to the identification of a flagellar antigenic type that was not detected by conventional methods; for those genomes, reads corresponding to all three antigenic types (two expected for the confirmed serotype and a third detected by SeqSero) were identified, suggesting that these strains may have a third flagellin allele. This phenomenon has been described before (33). The accuracy of the GenomeTrakr and assembled genomes dataset was somewhat lower; we were unable to confirm the accuracy of the annotated serotype for those strains. Since the serotype of these strain likely had been determined in a variety of laboratories and reported to GenomeTrakr, it is possible that at least some of these misidentifications were serotyping errors and not errors of our application. Since the isolates of these sequences were not available to us we could not confirm if the results of the original serotype determination were correct. Also, they represented a somewhat more diverse set of serotypes; partial serotype determination may be due allelic diversity in previously uncharacterized serotypes.

The option to input genome assemblies for analysis was designed to support high quality assemblies especially those made possible by long read sequencing platforms, such as PacBio. However, since O group prediction from assembled genomes is based on the entire *rfb* cluster and *Salmonella* and *E. coli* share some *rfb* cluster (26), the presence of an *E. coli* genome may produce a false positive *Salmonella* O group call (data not shown). The raw sequencing reads approach used the more discriminatory targets *wzx* and *wzy* for O group identification and is less likely to produce false positive calls. Also,

the genome assemblies in our validation data set produced a higher proportion of partial serotypes than did raw reads (Table 1), likely due to the failure in extracting serotype determinants from draft assemblies.

To improve differentiation of closely related H antigens, the assembly-free approach used a combination of reads mapping for efficiency and BLAST alignment for resolution. The first two of three rounds of mapping were used to identify a group of related H antigens (Table S5). The third round extracted reads that could be aligned to *fliC* and *fljB* loci, followed by a BLAST alignment to determine specific *fliC* and *fljB* antigenic types. This strategy has the potential to detect *Salmonella* serotypes from voluminous and noise-rich metagenome sequences of complex microbial communities such as fecal samples used for culture-independent diagnosis.

Rough, nonmotile and monophasic variants were excluded from the initial validation of the tool because they may possess non-expressed serotype determinants and may serotype differently by phenotypic and genetic methods. *fljB* may be deleted in some monophasic strains, in which case they will type the same by phenotypic and genetic methods. In other instances, some or all of *fljB* remains or the monophasic nature arises from mutation in the phase inversion mechanism; for those strains, flagellar antigen determinants not detected by phenotypic method may be detected by genetic methods. Although they were excluded here, the ability to more fully characterize these strains is an added benefit of serotype determined by genetic markers.

We were able to detect serotype Typhimurium from mouse fecal samples at four sampling times, including one day before oral infection. The strain on day -1 appeared to be present in a small amount and phylogenetically distinct from the challenge strain; its origin is unknown. Metagenomic samples known to contain *E. coli* O104:H4 did not produce any signal, suggesting that no false positive serotyping had been generated by pathogenic or commensal *Enterobacteriaceae* other than *Salmonella* in the fecal samples. Due the limited data available for the evaluation of serotype determination from metagenomic datasets, further investigation is needed to test the sensitivity and specificity of our tool when applied to metagenome sequencing data, especially when multiple strains of *Salmonella* with different serotypes are present in the same sample. While the serotype determination from WGS workflow consists of multiple steps and relies on various databases for reads mapping and BLAST alignment, a self-explanatory and easy-to-use web user interface is provided for public access to the tool. The web application runs on a cloud server and is compatible with all major Internet browsers and mobile devices; it requires no empirical or arbitrary parameters to be set for analysis, thus being user friendly for novice users.

Since serotype antigens are subject to horizontal transfer, serotype does not always correlate with phylogenetic relationships among *Salmonella* strains, i.e., strains from distinct genetic lineages may have the same complement of serotype antigens. It has been suggested that *Salmonella* serotyping be replaced by a genetic subtyping scheme, such as multilocus sequence typing (MLST) (34). However, serotyping continues to serve a key role as a first line subtyping method for *Salmonella*, with decades' worth of epidemiological data based on serotype identification. Our tool provides a simple and fast

Journal of Clinical Microbiology

means for determining serotype from a WGS using the determinants responsible for serotype. MLST and other genetic subtyping methods play an important role in public health surveillance and can provide phylogenetic context within a serotype when needed. The ongoing transition into advanced technologies such as WGS (35) will enable the integration of multiple identification, subtyping and characterization workflows typically employed in public health laboratories into a single, comprehensive and highly efficient platform, featuring *in silico* identification and prediction of various genotypic and phenotypic features (e.g., https://cge.cbs.dtu.dk/services/). Multiple methods can then be selected depending on the nature and scale of a particular investigation. Towards this prospect, the serotyping tool we present here maintains the well-established utility of *Salmonella* serotyping by bridging the gap between this historically important subtyping method and the cutting-edge application of whole- and metagenome sequencing in clinical and public health practices.

ACKNOWLEDGEMENTS

We are grateful to the FDA GenomeTrakr network for making large volumes of *Salmonella* whole genome sequences publicly available. We thank the 100K Foodborne Pathogen Genome Project for sequencing CDC isolates used in this study. This work was supported in part by contributions from the Board of Advisors, Center for Food Safety, University of Georgia and University of Georgia startup funds to X.D.

TABLES

Table 1. Accuracy of serotype predictions

Result	Number of genomes (% of total)		
	Reads mapping, CDC strains	Reads mapping, GenomeTrakr strains	Assembled genomes
Expected serotype[a]	304 (98.7%)	3,061 (92.6%)	324 (91.5%)
Unexpected serotype	2 (0.65%)	205 (6.2%)[b]	11 (3.1%)[b]
Partial or no serotype[c]	2 (0.65%)	40 (1.2%)	19 (5.4%)
Total tested	308	3306	354

[a]Predicted serotype was considered correct if serotype antigens detected corresponded to antigens detected by conventional methods. See Text S1 for discussion of interpretation of serotype results. For GenomeTrakr and genome assembly datasets, serotype prediction in consensus with annotated serotype was considered correct.

[b]Numbers represent serotype prediction inconsistent with annotated serotype; the accuracy of the annotated serotype is unknown.

[c]Some or all of the expected serotype determinants were not detected.

Table 2. Serotype determination from stool metagenomes of mice orally infected with *Salmonella.*

Sample		Number of reads mapped to individual antigen alleles[c]		
Sampling time	Accession[a]	*wzx/wzy* (O4)[b]	*fliC* (i)	*fljB* (1,2)
Day -1	SRR916930	273	2	2
Day 3	SRR916932	521	10	11
Day 6	SRR916933	519	12	10
Day 14	SRR916931	1572	21	21

[a]NCBI SRA accession number of the metagenome sequence.

[b]Predicted antigen type.

[c]The number of reads aligned to the best mapped antigen allele after the first round of reads mapping.

FIGURE LEGENDS

Figure 1. Major components and workflows of SeqSero. Two workflows are represented including serotype determination from 1. Genome assembly and 2. Raw sequencing reads.

Figure 2. Predicted incorrect H antigen identification using reads mapping with 95% confidence limits. A. Prediction for *fliC* identificaiton. B. Prediction for *fljB* identification. Logistic regression was used to estimate the probability of making an incorrect identification given the size of the mapped reads difference scaled by total number of reads sequenced from a genome. The GenomeTrakr dataset selected for SeqSero validation was used for this analysis. Observed correct and incorrect antigens calls were based on the first round of reads mapping.

Figure 3. Phylogenetic relationship among detected *Salmonella enterica* serotype Typhimurium strains from mice fecal metagenomes. A maximum likelihood tree showing the phylogenetic distance among the *Salmonella* strains serotyped from mice stool metagenomes before and after oral infection. Raw reads from each metagenome were mapped to the genome (GenBank accession number CP001363) of the infection strain (str. 14028s), high quality SNPs were identified and a core genome SNP maximum likelihood tree was built using similar methods as previously described in (36).

Figure 4. An example workflow of *fliC* H antigen prediction. Detailed description can be found in the Materials and Methods section. [a]Pre-defined antigen clusters are summarized in Table S5.

REFERECES

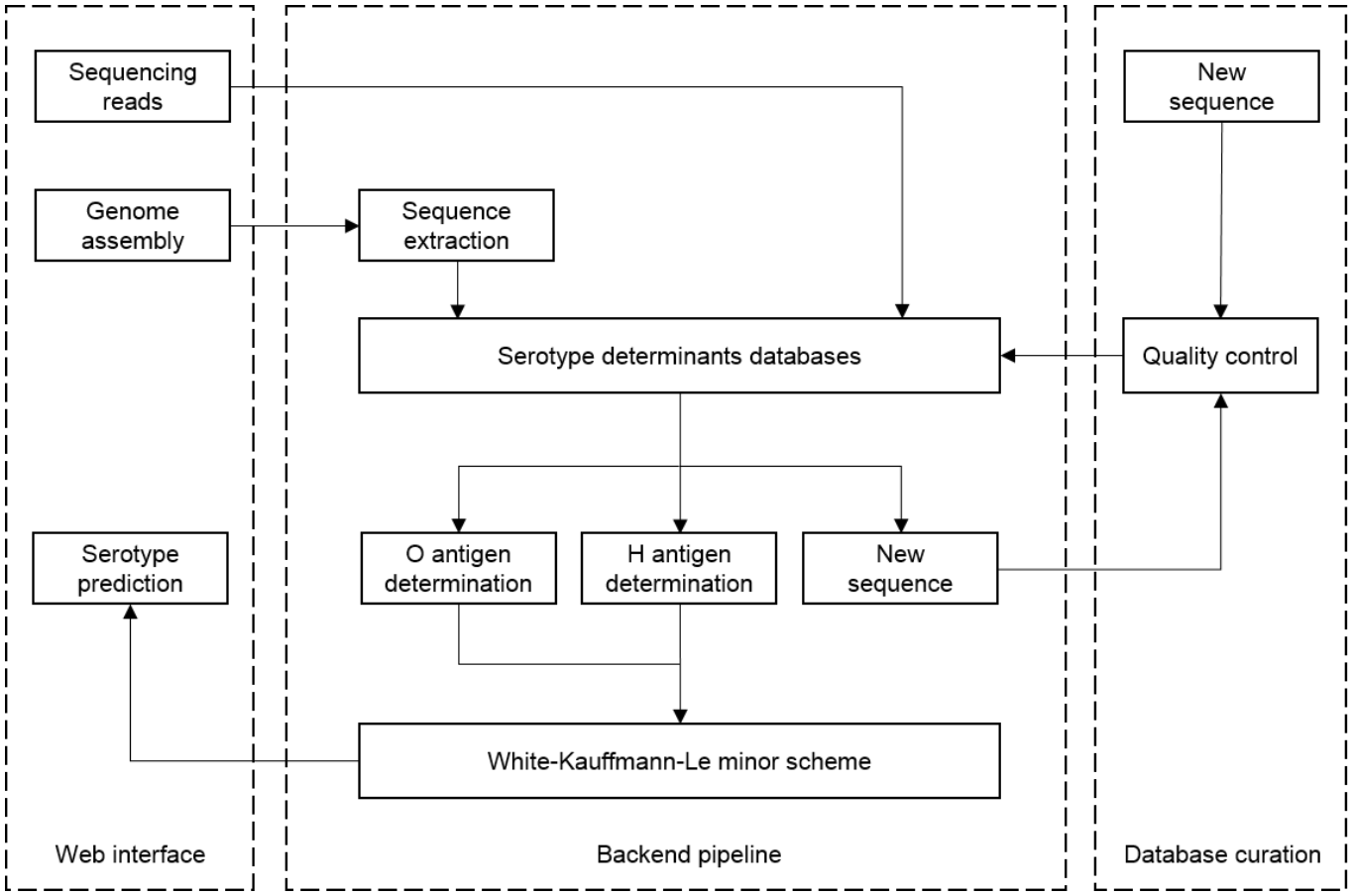
1. **Herikstad H, Motarjemi Y, Tauxe RV.** 2002. *Salmonella* surveillance: a global survey of public health serotyping. Epidemiology and infection **129:**1-8.
2. **Weinberger M, Keller N.** 2005. Recent trends in the epidemiology of non-typhoid Salmonella and antimicrobial resistance: the Israeli experience and worldwide review. Current opinion in infectious diseases **18:**513-521.
3. **Ran L, Wu S, Gao Y, Zhang X, Feng Z, Wang Z, Kan B, Klena JD, Lo Fo Wong DM, Angulo FJ, Varma JK.** 2011. Laboratory-based surveillance of nontyphoidal *Salmonella* infections in China. Foodborne pathogens and disease **8:**921-927.
4. **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. Emerging infectious diseases **17:**7-15.
5. **Grimont P, Weill F.** 2007. Antigenic formulae of the *Salmonella* serovars, 9th ed. WHO Collaborating Centre for Reference and Research on Salmonella.
6. **Guibourdenche M, Roggentin P, Mikoleit M, Fields PI, Bockemuhl J, Grimont PA, Weill FX.** 2010. Supplement 2003-2007 (No. 47) to the White-Kauffmann-Le Minor scheme. Research in microbiology **161:**26-29.

7. **Wattiau P, Boland C, Bertrand S.** 2011. Methodologies for *Salmonella enterica* subsp. enterica subtyping: gold standards and alternatives. Applied and environmental microbiology **77:**7877-7885.
8. **Shi C, Singh P, Ranieri ML, Wiedmann M, Moreno Switt AI.** 2013. Molecular methods for serovar determination of Salmonella. Critical reviews in microbiology.
9. **Samuel G, Reeves P.** 2003. Biosynthesis of O-antigens: genes and pathways involved in nucleotide sugar precursor synthesis and O-antigen assembly. Carbohydrate research **338:**2503-2519.
10. **Jiang XM, Neal B, Santiago F, Lee SJ, Romana LK, Reeves PR.** 1991. Structure and sequence of the rfb (O antigen) gene cluster of *Salmonella* serovar typhimurium (strain LT2). Molecular microbiology **5:**695-713.
11. **Smith NH, Selander RK.** 1990. Sequence invariance of the antigen-coding central region of the phase 1 flagellar filament gene (fliC) among strains of *Salmonella* typhimurium. Journal of bacteriology **172:**603-609.
12. **Vanegas RA, Joys TM.** 1995. Molecular analyses of the phase-2 antigen complex 1,2,.. of *Salmonella spp*. Journal of bacteriology **177:**3863-3864.
13. **Fitzgerald C, Collins M, van Duyne S, Mikoleit M, Brown T, Fields P.** 2007. Multiplex, bead-based suspension array for molecular determination of common *Salmonella* serogroups. Journal of clinical microbiology **45:**3323-3334.
14. **McQuiston JR, Waters RJ, Dinsmore BA, Mikoleit ML, Fields PI.** 2011. Molecular determination of H antigens of *Salmonella* by use of a microsphere-based liquid array. Journal of clinical microbiology **49:**565-573.
15. **Koser CU, Ellington MJ, Cartwright EJ, Gillespie SH, Brown NM, Farrington M, Holden MT, Dougan G, Bentley SD, Parkhill J, Peacock SJ.** 2012. Routine use of microbial whole genome sequencing in diagnostic and public health microbiology. PLoS pathogens **8:**e1002824.
16. **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.* Journal of clinical microbiology **52:**1501-1510.
17. **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. Journal of clinical microbiology **50:**1355-1361.
18. **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.
19. **Zankari E, Hasman H, Cosentino S, Vestergaard M, Rasmussen S, Lund O, Aarestrup FM, Larsen MV.** 2012. Identification of acquired antimicrobial resistance genes. The Journal of antimicrobial chemotherapy **67:**2640-2644.
20. **Yu G, Greninger AL, Isa P, Phan TG, Martinez MA, de la Luz Sanchez M, Contreras JF, Santos-Preciado JI, Parsonnet J, Miller S, DeRisi JL, Delwart E, Arias CF, Chiu CY.** 2012. Discovery of a novel polyomavirus in acute diarrheal samples from children. PloS one **7:**e49449.
21. **Loman NJ, Constantinidou C, Christner M, Rohde H, Chan JZ, Quick J, Weir JC, Quince C, Smith GP, Betley JR, Aepfelbacher M, Pallen MJ.** 2013.

A culture-independent sequence-based metagenomics approach to the investigation of an outbreak of Shiga-toxigenic *Escherichia coli* O104:H4. Jama **309:**1502-1510.

22. **Eisenstein TK.** 1975. Evidence for O antigens as the antigenic determinants in "ribosomal" vaccines prepared from *Salmonella*. Infection and immunity **12:**364-377.
23. **Hong Y, Cunneen MM, Reeves PR.** 2012. The Wzx translocases for *Salmonella enterica* O-antigen processing have unexpected serotype specificity. Molecular microbiology **84:**620-630.
24. **Brenner FW, McWhorter-Murlin AC.** 1998. Identfication and serotyping of *Salmonella*. Centers for Diseases Control and Prevention. Atlanta, GA.
25. **Deatherage Kaiser BL, Li J, Sanford JA, Kim YM, Kronewitter SR, Jones MB, Peterson CT, Peterson SN, Frank BC, Purvine SO, Brown JN, Metz TO, Smith RD, Heffron F, Adkins JN.** 2013. A Multi-Omic View of Host-Pathogen-Commensal Interplay in -Mediated Intestinal Infection. PloS one **8:**e67155.
26. **Liu B, Knirel YA, Feng L, Perepelov AV, Senchenkova SN, Reeves PR, Wang L.** 2014. Structural diversity in *Salmonella* O antigens and its genetic basis. FEMS microbiology reviews **38:**56-89.
27. **Fitzgerald C, Sherwood R, Gheesling LL, Brenner FW, Fields PI.** 2003. Molecular analysis of the rfb O antigen gene cluster of *Salmonella enterica* serogroup O:6,14 and development of a serogroup-specific PCR assay. Applied and environmental microbiology **69:**6099-6105.
28. **McQuiston JR, Parrenas R, Ortiz-Rivera M, Gheesling L, Brenner F, Fields PI.** 2004. Sequencing and comparative analysis of flagellin genes fliC, fljB, and flpA from *Salmonella*. Journal of clinical microbiology **42:**1923-1932.
29. **Li H, Durbin R.** 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics **26:**589-595.
30. **Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL.** 2009. BLAST+: architecture and applications. BMC bioinformatics **10:**421.
31. **Miller JR, Koren S, Sutton G.** 2010. Assembly algorithms for next-generation sequencing data. Genomics **95:**315-327.
32. **Ottesen AR, Gonzalez A, Bell R, Arce C, Rideout S, Allard M, Evans P, Strain E, Musser S, Knight R, Brown E, Pettengill JB.** 2013. Co-enriching microflora associated with culture based methods to detect *Salmonella* from tomato phyllosphere. PloS one **8:**e73079.
33. **Smith NH, Selander RK.** 1991. Molecular genetic basis for complex flagellar antigen expression in a triphasic serovar of *Salmonella*. Proceedings of the National Academy of Sciences of the United States of America **88:**956-960.
34. **Achtman M, Wain J, Weill FX, Nair S, Zhou Z, Sangal V, Krauland MG, Hale JL, Harbottle H, Uesbeck A, Dougan G, Harrison LH, Brisse S, Group SEMS.** 2012. Multilocus sequence typing as a replacement for serotyping in *Salmonella enterica*. PLoS pathogens **8:**e1002776.
35. **Kupferschmidt K.** 2011. Epidemiology. Outbreak detectives embrace the genome era. Science **333:**1818-1819.

36. **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. Genomic Epidemiology of *Salmonella enterica* Serotype Enteritidis based on Population Structure of Prevalent Lineages. Emerging infectious diseases **20:**1481-1489.



Journal of Clinical Microbiology
JCM
Journal of Clinical Microbiology