AMERICAN SOCIETY FOR MICROBIOLOGY | mBio®

RESEARCH ARTICLE
Ecological and Evolutionary Science

Salmonella enterica Phylogeny Based on Whole-Genome Sequencing Reveals Two New Clades and Novel Patterns of Horizontally Acquired Genetic Elements

Jay Worley,[a]* Jianghong Meng,[a,b] Marc W. Allard,[c] Eric W. Brown,[c] Ruth E. Timme[c]

[a]Joint Institute for Food Safety and Applied Nutrition, University of Maryland, College Park, Maryland, USA
[b]Department of Nutrition and Food Science, University of Maryland, College Park, Maryland, USA
[c]Center for Food Safety & Applied Nutrition, U.S. Food & Drug Administration, College Park, Maryland, USA

ABSTRACT Using whole-genome sequence (WGS) data from the GenomeTrakr network, a globally distributed network of laboratories sequencing foodborne pathogens, we present a new phylogeny of *Salmonella enterica* comprising 445 isolates from 266 distinct serovars and originating from 52 countries. This phylogeny includes two previously unidentified *S. enterica* subsp. *enterica* clades. Serovar Typhi is shown to be nested within clade A. Our findings are supported by both phylogenetic support, based on a core genome alignment, and Bayesian approaches, based on single-nucleotide polymorphisms. Serovar assignments were refined by *in silico* analysis using SeqSero. More than 10% of serovars were either polyphyletic or paraphyletic. We found variable genetic content in these isolates relating to gene mobilization and virulence factors which have different distributions within clades. Gifsy-1- and Gifsy-2-like phages appear more prevalent in clade A; other viruses are more evenly distributed. Our analyses reveal IncFII is the predominant plasmid replicon in *S. enterica*. Few core or clade-defining virulence genes are observed, and their distributions appear probabilistic in nature. Together, these patterns demonstrate that genetic exchange within *S. enterica* is more extensive and frequent than previously realized, which significantly alters how we view the genetic structure of the bacterial species.

IMPORTANCE Rapid improvements in nucleotide sequencing access and affordability have led to a drastic increase in availability of genetic information. This information will improve the accuracy of molecular descriptions, including serovars, within *S. enterica*. Although the concept of serovars continues to be useful, it may have more significant limitations than previously understood. Furthermore, the discrete absence or presence of specific genes can be an unstable indicator of phylogenetic identity. Whole-genome sequencing provides more rigorous tools for assessing the distributions of these genes. Our phylogenetic and genetic content analyses reveal how active genetic elements are dynamically distributed within a species, allowing us to better understand genetic reservoirs and underlying bacterial evolution.

KEYWORDS GenomeTrakr, phylogeny, plasmids, *Salmonella*, virulence, whole-genome sequencing

Salmonella enterica is a leading cause of bacterial foodborne illness worldwide, and within the United States, it is the organism highest in incidence and total morbidity (1). The vast majority of salmonellosis is caused by a small set of the over 2,600 serovars found in *S. enterica* subsp. *enterica* (2). These serovars are named based on the antigens presented on the outer membrane and flagella. Although serotypes are often thought of as genetically exclusive groups, some cases of polyphyly have been documented

Received 22 October 2018 **Accepted** 25 October 2018 **Published** 27 November 2018

Citation Worley J, Meng J, Allard MW, Brown EW, Timme RE. 2018. *Salmonella enterica* phylogeny based on whole-genome sequencing reveals two new clades and novel patterns of horizontally acquired genetic elements. mBio 9:e02303-18. https://doi.org/10.1128/mBio.02303-18.

Editor Qijing Zhang, Iowa State University

Address correspondence to Jay Worley, worley.jn@gmail.com, or Ruth E. Timme, ruth.timme@fda.hhs.gov.

* Present address: Jay Worley, Department of Pathology, Brigham and Women's Hospital, Boston, Massachusetts, USA.

This article is a direct contribution from a Fellow of the American Academy of Microbiology. Solicited external reviewers: John McQuiston, Centers for Disease Control and Prevention; David Rasko, University of Maryland School of Medicine.

TABLE 1 Genome assembly statistics

Parameter	Avg	SD	Minimum	Maximum
Genome size (Mb)	4.73	0.13	4.41	5.17
No. of contigs	57.9	77.7	18	704
Coverage	41.9	19.0	20.1	292.1

where genetically distinct *S. enterica* bacteria contain a mixture of named serovars (*S. enterica* subsp. *enterica* serovars Typhimurium and O-4,[5],12:i:−) or multiple independent lineages of a single serovar (*S. enterica* serovars Newport and Bareilly) (3, 4).

There are currently six subspecies recognized within *S. enterica*: *enterica* (I), *salamae* (II), *arizonae* (IIIa), *diarizonae* (IIIb), *houtenae* (IV), and *indica* (VI) (5, 6). Several DNA-based phylogenetic investigations established *S. enterica* subsp. *arizonae* as the earliest diverging (5, 6). In addition, more recently, phylogenetic resolution between the other five subspecies has emerged using whole-genome sequencing (WGS) (7) with evidence for additional unnamed subspecies (8). Within *S. enterica* subsp. *enterica*, previous molecular phylogenies revealed a major split resulting in two major lineages: clades A and B (4, 9). Clade A included most strains associated with disease in humans, including *S. enterica* serovars such as Typhi, Typhimurium, and Enteritidis. Within clade A, there were two well-supported lineages known as clades A1 and A2; within the second subclade, there were several notable sections: Typhi, Typhimurium, and Newport (4). Clade B was phylogenetically distinct and recognized as having a unique profile of virulence and metabolic genes (9).

The potential for and documentation of horizontal gene transfer in *S. enterica* subsp. *enterica* are extensive both within and between these clades. Even among genes not associated with virulence, such as those considered to be "core," there are at least three known examples of horizontal gene transfer (HGT) within *S. enterica* subsp. *enterica* (10–12), focused mainly on virulence genes, phages, and plasmids. One of the first reports of HGT of specific genes in *S. enterica* subsp. *enterica* focused on the fimbrial operons, which were shown to have variable patterns between serovars (13). Early genomic analyses highlighted the flexibility of these genomes due to the prevalence of horizontally acquired DNA, particularly in regard to virulence genes (14–16). The increased availability of sequencing information led to investigations revealing patterns of inheritance within the virulence genes (9) and, more generally, patterns and timing of inheritance (17, 18). With expanded taxon sampling, it should be possible to add detail and nuance to these patterns and describe new ones.

GenomeTrakr is a distributed laboratory network that collects and publicizes foodborne pathogen genome sequences from public health laboratories, academic laboratories, and other U.S. and non-U.S. partners (19). As of September 2018, the database for *S. enterica* contained more than 155,000 genome sequences and associated metadata, all collected from food, facilities, animals, the environment, and human clinical patients, making it the largest, most diverse, curated database of its kind. Understanding the broader evolutionary context of this pathogen helps public health officials understand why certain lineages appear to be more problematic than others. For this reason, we sought to leverage the GenomeTrakr database to expand our understanding of the genetic diversity in *S. enterica* subsp. *enterica* and to find new phylogenetic features and patterns. The presence and absence of individual phages, plasmid replicons, and virulence genes reveal dynamics of genetic acquisition, loss, and maintenance that contribute to *S. enterica* subsp. *enterica* evolution. From this, several new perspectives emerge about pathogen genomics.

RESULTS

Taxon sampling and associated sequence data. Our WGS data set of 445 *Salmonella* isolates included representatives from all six subspecies of *S. enterica and* 260 different serovars concentrated in *S. enterica* subsp. *enterica*. These isolates originated from 52 countries across the six agriculturally productive continents (Table 1; see Data

Set S1 in the supplemental material). A total of 166 isolates are from the United States, 51 are from Mexico, and 19 are from other countries within North America. Of the 153 isolates originating from Asia, 51 were from India. Among the rest of the isolates, 27 originated from Africa, 12 from South America, 7 from Europe, and 4 from Oceania, and 6 had an unknown origin.

The multigene alignment, representing a core genome, totaled 2,278 genes that were present only once in each genome and of the same length, without indels. The data set resulting from this approach, meant to reduce phylogenetically confusing genetic information that could be created by homologous recombination and sequences problematic for assembly, is 2,036,954 bases in length compared to the 4,857,450 present in the reference genome of *S. enterica* serovar Typhimurium LT2.

The broad phylogenetic structure of *Salmonella enterica*. This core genome was used to infer a phylogenetic tree representing the genomic diversity within *S. enterica* subsp. *enterica* (Fig. 1; see Data Set S2 in the supplemental material). The topology of this tree is consistent with previous results from which we assigned known clade identities (4, 9); it reveals a deep split in the subspecies, forming the primary clades A and B, with section Typhi embedded within clade A. However, our tree also reveals two new clades, here called C and D (Fig. 1). Clade D, which contains eight isolates, is sister to clades A, B, and C. Clade C, containing four isolates, is sister to clades A and B. Based on the STRUCTURE results and relative branch lengths, we consider them members of *S. enterica* subsp. *enterica,* not new subspecies (20).

Phylogenetic support for previously named lineages varied in our expanded topology. The new clades, C and D, both have 100% bootstrap support (BS) (see Fig. S1 in the supplemental material). The monophyly of the lineage comprising clades A and B was strongly supported with 100% BS (Fig. S1). Clades A and B themselves were also strongly supported (100 and 88% BS, respectively). Within clade A, there was moderate support for two large subclades: A1, which includes section Typhi, and A2 (69 and 63% BS, respectively). Finally, section Typhi was strongly supported with 95% BS.

A single-nucleotide polymorphism (SNP) analysis of core SNPs using kSNP 3.0 revealed no unique SNPs defining clades A and B (Fig. S1C). However, clades C and D have 62 and 36 unique SNPs that define their identity, respectively. The largest number of unique SNPs found for any serovar represented more than once was for *S. enterica* serovar Typhi, which had 74 unique SNPs.

A Bayesian approach using the program STRUCTURE revealed support for six populations within this data set (Fig. 1). Clade A1 appears genetically distinct from A2. Despite the moderate bootstrap support for the section Typhi + clade A2, STRUCTURE uncovers mostly one population underlying the two lineages, strengthening its placement here in the phylogeny. This analysis also clearly distinguished clade D from the rest of *S. enterica* subsp. *enterica*, whereas clade C, while distinct, appears to contain a mixture of three sequence populations. Similarly, earlier-diverging strains in clade B appear to contain a mixture of sequence populations. Population mixtures with this pattern could potentially arise from past recombination events, or perhaps more likely, there was an ancestral blue population that gave rise to clades A, B, and C: thus, earlier-diverging lineages within these clades carry that genetic signature.

Although we recognized that our data set overrepresents isolates collected from North America and Asia, the continent-colored barcode shows no broad patterns of geographic localization at the clade level (Fig. 1).

Serovar assignment and diversity. SeqSero, a program developed to make serovar predictions based on WGS information, was used to check the classically determined serovar assignment (21). Most serovar calls were confirmed, but there were a significant number of sequences where SeqSero and the classically determined serovar assignment disagreed (Table 2; Data Set S1). Approximately 78% of the assignments were in agreement with what SeqSero determined, with many of the others being close in observed/predicted antigenic formula, but there were a set of 22 where all three antigens used for serotyping, two flagellar and one lipopolysaccharide, were consid-

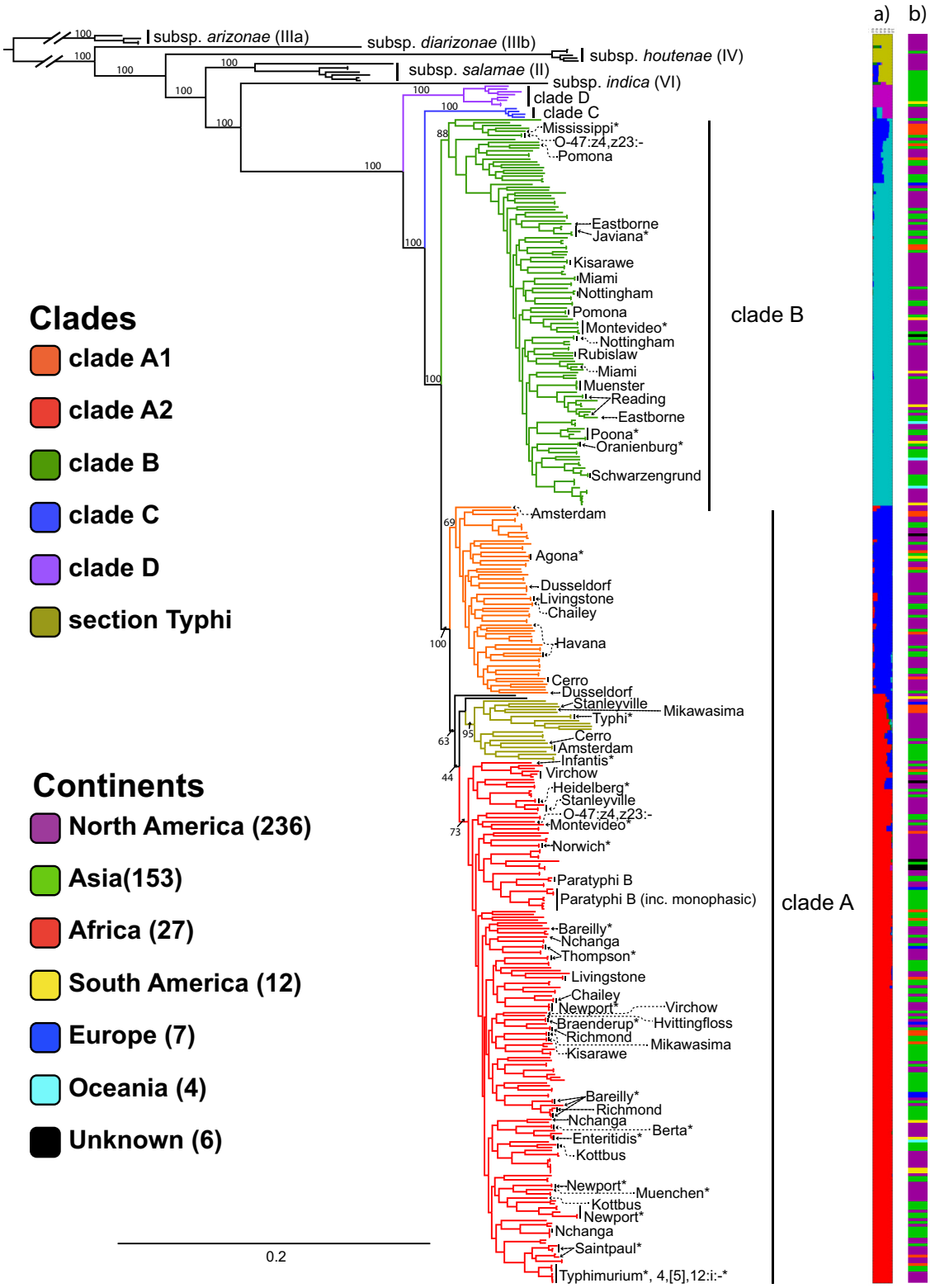

FIG 1 phylogenetic analysis of *S. enterica*. Shown is a maximum-likelihood phylogeny of *S. enterica* calculated using a core genome alignment with RAxML. Bootstrap support is included for the major lineages. Named clades within *Salmonella enterica* subsp. *enterica* are indicated by color. The top 20 serovars found in the United States in 2015, plus serovar Typhi, are noted with asterisks. The remaining labeled serovars are nonmonophyletic. (a) Bayesian clustering for 6 assumed populations calculated using core SNPs found using kSNP with STRUCTURE. (b) Continent from which each isolate originated, indicated by color: Africa, red; Asia, green; Europe, blue; North America, purple; South America, yellow; Oceania, cyan; unknown, black.

TABLE 2 Serotype changes

Category[a]	Count	%
No database call	2	0.4
Antigens in disagreement		
3 antigens	22	4.9
2 antigens	28	6.3
1 antigen	36	8.1
Antigens in agreement	357	78.4
Total	445	100.0
Preferred database call	13	25.6
Preferred SeqSero call	38	74.5

[a]The categories represent the different types of serotype call corrections made in this study. "No database call" indicates there was no serotype call in the original database. "Antigens in disagreement" indicates the number of antigens where disagreement was found between the serotype formulas indicated. The "Preferred database/SeqSero call" indicates there was a change made where one method was clearly preferred over the other, usually by phylogenetic context. In some cases, there is no reason for preference, so the number of calls here is less than the number of changes called above. For these, the percentage listed is for those in which one was preferred.

ered in disagreement. For cases in which a serotype could be corrected with high confidence based on phylogenetic relationship to other serotypes or any other evidence stated in Data Set S1, SeqSero was used nearly 3 to 1.

The top 20 most prevalent serovars are labeled in Fig. 1 (22). Of these, three-quarters are located in clade A, with several sharing a common ancestor with *S. enterica* serovar Typhimurium. From these prevalent serovars, *S. enterica* serovar Newport is polyphyletic, with three distinct clades, and *S. enterica* serovar Saintpaul appears to be paraphyletic, with an isolate from *S. enterica* serovar Haifa nested within.

While most serovars of *S. enterica* subsp. *enterica* sampled in this study appear to be monophyletic (4), we see a significant number that are not (Data Set S2). As in a previous report, two lineages of *S. enterica* serovar Bareilly have been found, one of them with nested isolates of *S. enterica* serovar Richmond. *S. enterica* serovar Richmond was likewise polyphyletic. *S. enterica* serovar Newport comprises three independent clades, two of which appear to be closely related.

Twenty-four serovars appear to be polyphyletic within *S. enterica* subsp. *enterica*: *S. enterica* serovars O-47:z4,z23:−, Amsterdam, Bareilly, Cerro, Chailey, Duesseldorf, Eastbourne, Havana, Hvittingfoss, Kisarwe, Kottbus, Livingstone, Miami, Mikawasima, Montevideo, Nchanga, Newport, Nottingham, Pomona, Reading, Richmond, Stanelyville, Thompson, and Virchow (Data Set S2). Missing from this report, but present in a previous report, is *S. enterica* serovar Muenchen (4). Six serovars are paraphyletic: *S. enterica* serovars Bareilly, Bredeney, Kibusi, Onderstepoort, Paratyphi B, and Saintpaul. These appear to be instances in which one serovar is nested within another serovar, both sharing a recent common ancestor. Combined, these represent 29 different serovars that are not monophyletic—over 10% of the sampled serovar diversity within *S. enterica* subsp. *enterica*, which has 247 serovars represented. In the other subspecies, the serovar *S. enterica* subsp. *houtenae* IV O-43:z4:z23 was paraphyletic.

Phage and plasmid populations. Intact lysogenic phages were detected using PHASTER (23). A diverse set of phages, including 46 different putatively identified phages (greater than 50% of the found region's coding DNA sequences [CDSs] matching a reference phage genome), were found in the genome sequences. PHASTER identified 1,122 phages' genome sequences as "intact," 586 (52%) of which were putatively identified by gene content. Of these 586 phages, nine were found more than once in a given genome, and of these five were related to Fels-2 and two to Gifsy-1. Genomic regions from these 586 putatively identified phages were found in 320 of 455 isolates, with a maximum of six intact phages in one isolate. Of the 46 different putatively identified phages, 12 were present in more than 10 of the isolates (Fig. 2; see Fig. S2 in the supplemental material). Gifsy-1 and Fels-2 were the most common phages

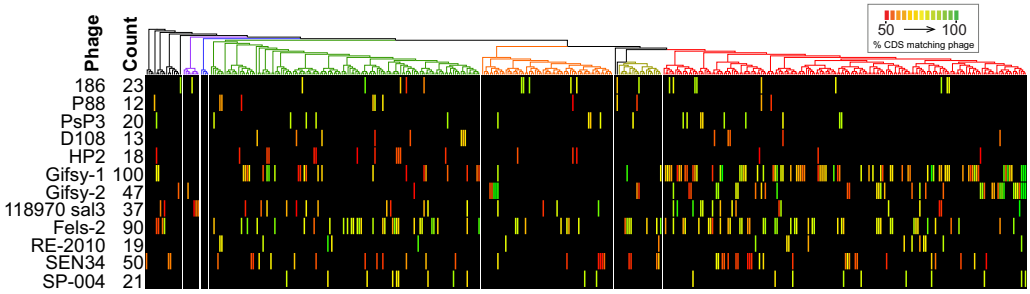

FIG 2 Distribution of prophages. Distribution of prophages for which more than 10 related examples were found scored "intact" by PHASTER, with a cutoff of 50% of CDSs assigned to the indicated phage type used for phage assignment. Color indicates the percentage of CDSs within the phage region that matches the indicated phage on a linear scale from red at 50% to green at 100%. The phylogeny at the top uses the same colors as Fig. 1 for the different clades.

found with intact phages found in 100 and 90 strains, respectively (Fig. 2). Phages related to Gifsy-2, a relative of Gifsy-1, were found 47 times.

Different phages have different patterns of inheritance within *S. enterica* subsp. *enterica*. Both Gifsy-1 and Gifsy-2 seem to be more prevalent among clade A isolates, while other phages, like those related to Fels-2, seem to be evenly distributed throughout the subspecies (Fig. 2; see Fig. S3 in the supplemental material). In the case of Gifsy-1 and Gifsy-2, very strong hits with high identity are seen in strains that are or are closely related to *S. enterica* serovar Typhimurium, the origin of these reference sequences, and so the presence of many phages with different levels of sequence identification indicates significant genetic heterogeneity within these phages.

However, and unsurprisingly given the often repetitive structure of phage genomic sequence, many of these intact phage sequences presented trouble for genome assembly. Of the 458 hits displayed in Fig. 2, 47% of them are located within 2 kb of a contig end, 23% within 100 bases. The phage classification metric used here, percentage of CDSs matching a reference sequence, means that the hits should not be influenced by contig ends since missing CDSs are not counted.

Plasmid replicons were widely present among the strains sampled, but less prevalent than lysogenic phages. 204 replicons were found among 160 strains, with a maximum of three observed in a single isolate (Fig. 3; see Fig. S4 in the supplemental material). Plasmid replicons of IncFII were the most common found, totaling 108. Of the IncFII subgroups, the most common type was IncFII (24), with 59 examples found. Most of the plasmid replicons, however, were represented by only a few examples, and only 5 of the 28 specific replicon types were seen more than 10 times. The overall patterns of plasmid presence, though, seem to be similar to that of phages, but without any obvious clade preferences (Fig. S4).

Type III effector gene presence and absence. The Virulence Factor Database (VFDB) was used to screen a selection of type III secretion effector-encoding genes from *S. enterica* subsp. *enterica* against the assembled genomes (25). These are classified as being secreted by one or both of the pathogenicity island-encoded type III secretion systems (T3SSs), the SPI-1-encoded T3SS and the SPI-2-encoded T3SS.

Of the SPI-1 T3SS-secreted effector-encoding genes screened, *sipA*, *sipB*, *sipC*, and *sptP* were present in all strains (Fig. 4; see Data Set S3 in the supplemental material), and *sopE2* was present in all strains except one. *sopD* was present in all clades, except for clade C. *sopB*/*sigD* is core to *S. enterica*, but appears to have been lost in *S. enterica* subsp. *salamae*. *sopD* and *sopA* are missing from one clade each (C and D, respectively) and may be defining absences for these clades. The effectors *sopE* and *avrA* show greater variability across *S. enterica*, although *avrA* shows more stability within clade A2 and *sopE* is absent from nearly all taxa in clade A1.

Of the SPI-2 T3SS-secreted effector-encoding genes, only *spiC*/*ssaB* appears to be

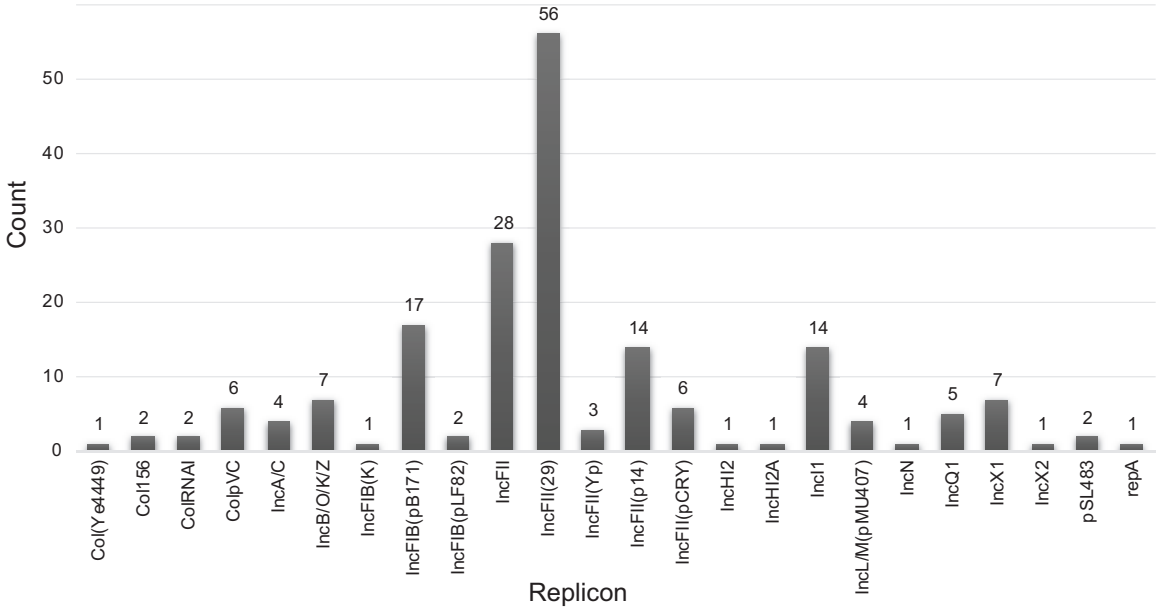

FIG 3 Plasmid replicon abundance in *Salmonella enterica*. Shown is the number of times the indicated replicon was identified within the isolates using the PlasmidFinder database and an 80% nucleotide identity cutoff threshold.

fully core to *S. enterica*. *sifA* was found in all strains of *S. enterica* subsp. *enterica*. Several effectors were found in nearly all strains (*sifB*, *spiC/ssaB*, *sseF*, *sseG*, *pipB*, *sopD2*, and *sseL*) and might be considered core genes. *pipB2* is found in most clade A strains and all clade C and D strains, but is absent from many clade B strains. The other effectors—*gogB*,

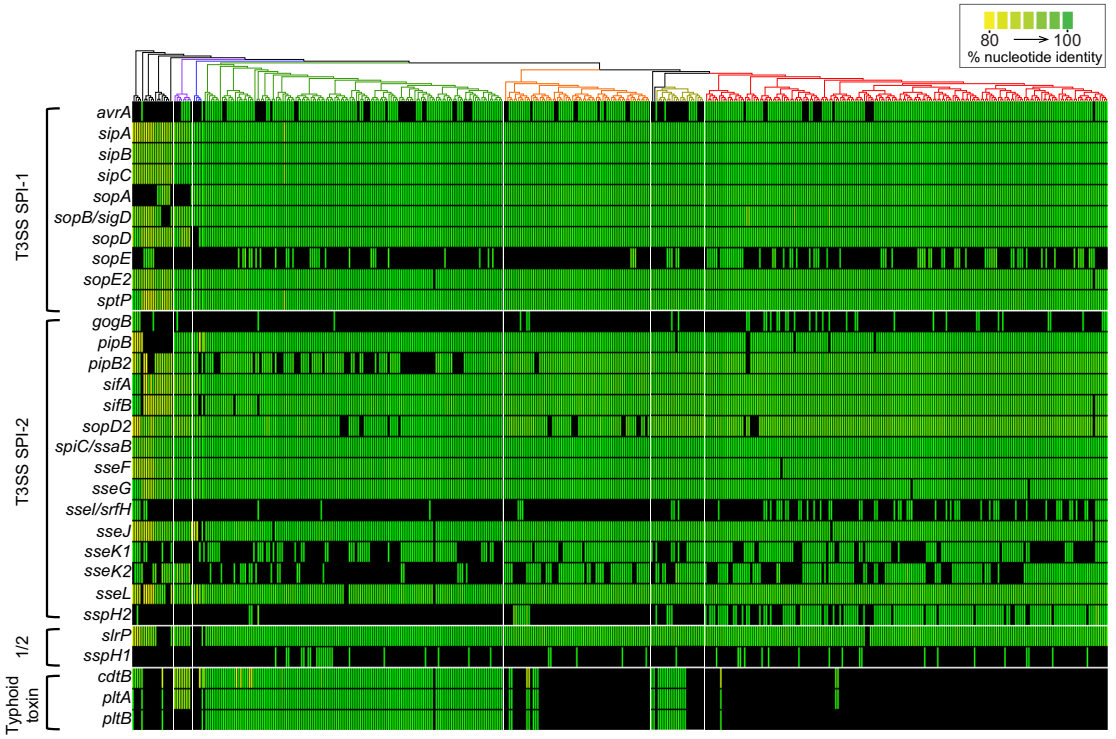

FIG 4 Distribution of genes encoding T3SS-delivered proteinaceous effectors and typhoid toxins. Shown is the presence or absence of genes coding for proteinaceous effectors delivered by the T3SS encoded in SPI-1, SPI-2, or delivered by both ("1/2"), as well as typhoid toxin subunit-encoding genes. The phylogeny at the top uses the same colors as Fig. 1 for the different clades. Color indicates the hit percentage of nucleotide identity on a linear scale from yellow at 80% to green at 100%. *sspH1* and *sspH2* were only shown above the 90% threshold to account for the high similarity of the alleles.

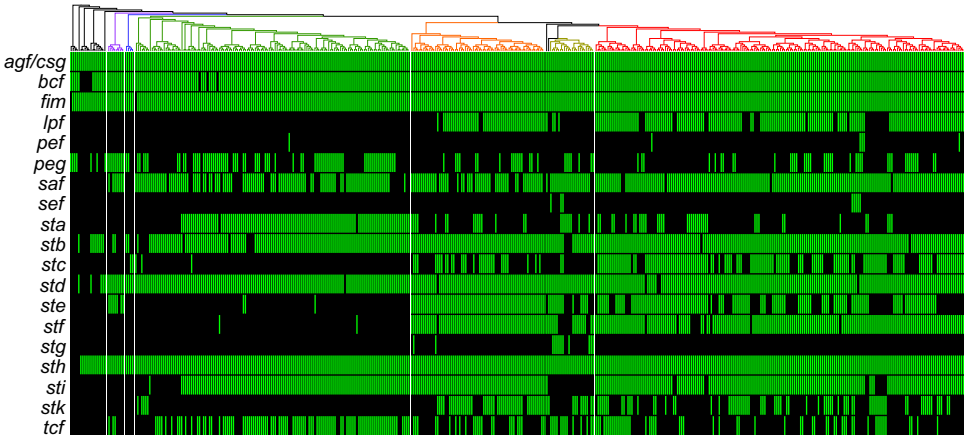

FIG 5 Distribution of fimbrial operons. Shown is the presence or absence of the fimbria-encoding operons. The phylogeny at the top uses the same colors as Fig. 1 for the different subclades. Positive hits had at least 80% nucleotide identity for greater than 75% of the genes in the operon, using the VFDB as a reference. A positive hit is indicated in green with no respect to sequence identity.

sopD2, *sseI*/*srfH*, *sseK1*, *sseK2*, and *sspH2*—are variably present in *S. enterica* subsp. *enterica*, although *sspH2* is heavily concentrated in clade A2.

Of the effectors classified as secreted from both secretion systems, the *sspH1* and *sspH2* genes reveal opposite patterns across the phylogeny: when one is present, the other is not (most clearly seen in clade A2). *slrP* is nearly a core effector within the subspecies, but is absent from clade C as well as one small lineage in clade A.

Typhoid toxin. The typhoid toxin genes *ctdB*, *pltA*, and *pltB* are found in almost all clade B strains, a little more than half of section Typhi strains, and infrequently in the rest of clade A (Fig. 4). The proteins encoded by these genes form a holotoxin, with PltA and CtdB being the enzymatically active units and PltB forming a pore. An interesting pattern was found in clade D and two strains from clade A, where only the subunits *cdtB* and *pltA* were found at a lower level of sequence homology, which is unexpected since deletions in *pltB* have eliminated the virulence activity of this toxin (26).

Additionally, we found different variants of the typhoid toxin genes within *S. enterica* subsp. *enterica*. An aligned concatenation of the *pltA*, *pltB*, and *ctdB* gene protein-encoding sequences revealed three distinct typhoid toxin sequence variants (see Fig. S5 in the supplemental material). One of these variants includes most clade B, section Typhi, and A1 sequences. The second type is more related to the typhoid toxin-encoding genes found in *S. enterica* subspecies *arizonae* and *diarizonae* and contains five sequences: three from clade B and two from clade A2. Three of the sequences represent duplicate copies of the typhoid toxin-encoding genes that are located elsewhere in the genome: NY_FSL S10-1092, FNE0139, and FDA00004119. The third group includes four clade B sequences, two clade A1 sequences, one clade A2 sequence, and a sequence found in one of the *S. enterica* subsp. *salamae* isolates. This indicates communication of the toxin across subspecies.

Fimbrial adherence gene operons indicate different selective pressures. The variability within fimbrial repertoires encoded by each strain is underscored because, out of the 19 operons screened for, only two, *agf*/*csg* and *fim*, appear to be core to *S. enterica*, and only two, *bcf* and *sth*, appear to be core to *S. enterica* subsp. *enterica* (Fig. 5). Others appear to be widespread. For example, *stb* was found throughout the subspecies, but was absent within a few small clades. The new clades C and D show some interesting patterns here: clade C has lost *saf*, and clade D appears to have independently acquired *ste*.

Interesting patterns within clades A and B were also found. Most notably, *lpf*, *stc*, *ste*, *stf*, and *stk* tend to be widely present among subclades A1 and A2, with section Typhi having a lower proportion of strains carrying *lpf*, *stc*, and *stf*. In addition, section Typhi

appears to have lost *sti*, which is otherwise widespread across the subspecies. Section Typhi, conversely, appears to have acquired *stg*, only found in two other isolates. Clade B has a higher proportion of strains carrying *sta*, *tcf*, and *peg*. Some of the early diverging lineages in clade B are missing *sta* and *sti*, as are clades C and D.

DISCUSSION

This data set comprises a broad sampling of *S. enterica* subsp. *enterica* diversity (19), and despite sampling only one-tenth of the 2,600 described serovars, new clades and genetic patterns were revealed using WGS, underscoring the potential for additional refinement and detail within *S. enterica* phylogenies. Additionally, large-scale WGS made possible detailed phylogenetic analyses regarding distribution patterns of genetic elements within clades.

Two new *S. enterica* subsp. *enterica* clades were uncovered in this investigation: clades C and D. Their relative rarity within the GenomeTrakr database may explain why these groups have not been previously described and assigned. According to our STRUCTURE results (Fig. 1), isolates from these two lineages appear to belong to *S. enterica* subsp. *enterica* and do not represent new subspecies. Although with a relatively limited sample size, we see the following set of synapomorphies that appear to define the new lineages: clade C isolates lack the secreted effector-encoding genes *sopD* and *slrP*, while clade D isolates have variants of the typhoid toxin genes *cdtB* and *pltA*, while lacking *pltB*, and lack *sopA*. Relative to clades B and C, clade D strains are more likely to carry the fimbrial operons *peg* and *ste* and the secreted effector protein gene *sseK2*.

The overall phylogeny agreed with previous phylogenetic assessments for both the relationship between the six subspecies and for the major clades within *S. enterica* subsp. *enterica* (clade A, clade B, and section Typhi) (4, 7, 9). We see strong support for the major clades (A, B, C, and D), but more moderate support for the relationships within each clade.

In most cases, we found serovars to be monophyletic. However, a significant number of polyphyletic serovars exist within *S. enterica* subsp. *enterica*, and in the case of *S. enterica* serovar Fulica, between subspecies. Most of these are supported independently of SeqSero, and over three-fourths of our calls were in complete agreement between SeqSero and the user-reported serotype. It is important to note that we used phylogenetic context and antigenic formula relatedness as important pieces of evidence in serotype calling, not relying on any piece of information in isolation. It is not possible to directly compare methods here since different labs that submitted strains may use different methods to serotype their *Salmonella enterica* isolates, but we find large agreement nonetheless. Of those with disagreements, 41% (36/88) were with one antigen alone. Often, these would take the form of well-known variants, such as potential "rough" variants that lack an O antigen, or the acquisition of a plasmid carrying a new flagellar antigen. Serotypes continue to provide invaluable context to microbiological investigations, but bioinformatic tools can provide an important validation to existing typing steps and add phylogenetic context.

In contrast to previous reports, and perhaps due to the expanded diversity of isolates selected for this study, polyphyly was found in approximately one-tenth of the serovar diversity, as well as 7 examples of paraphyletic serovars (4). The true number of polyphyletic serovars is likely higher and reinforces the utility of using WGS for routine pathogen surveillance. A previous report using multilocus sequence typing (MLST) among a large set of isolates suggested that over half of the sampled serovars within were polyphyletic (27). This study has increased depth per genome (ca. 2 Mb versus 7 gene fragments), but less depth in the number of representatives per serotype (445 isolates in 266 serovars versus 4,257 in 554), which could account for the difference.

Among the serovars, *S. enterica* serovar Typhi had the most unique genetic information, with the highest number of unique core SNPs relative to its neighbors. This obligate human pathogen has been noted for its highly clonal phylogenetic structure and high number of pseudogenes compared to its nearest relatives (15, 24, 28, 29).

This study addressed two major contributors to horizontal gene transfer in *Salmonella enterica*. Sequence, database, and tool limitations mean that the results do not represent an authoritative assessment of the plasmid and phage populations within *S. enterica* subsp. *enterica*. However, several evolutionarily important patterns of horizontal transfer within the subspecies were observed that shed light on *S. enterica* subsp. *enterica* evolution. The diverse temperate phage populations in *Salmonella enterica* subsp. *enterica* are important sources of horizontal gene transfer, including virulence and virulence-related genes (30–32). It is intriguing that the data suggest phages similar to Gifsy-1 and Gifsy-2 may contribute more to HGT within clade A2 relative to other clades. Other phages, such as those similar to Fels-2, seem to contribute more evenly across the subspecies.

High diversity but lower prevalence was found for plasmid replicons. In this preliminary work, the focus was put on the replicons as a first step toward understanding plasmid dynamics on the population scale. The replicon classes IncFI and IncFII are some of the most common in *S. enterica* subsp. *enterica*, although our results do show a greater proportion of IncFII type replicons than a previous report suggests (33). As in the phage populations, there is evidence here for the frequent exchange of plasmids since different replicon subtypes of the same incompatibility group appear in closely related strains. Previous reports have found highly variable plasmid content within *S. enterica* serovar Typhimurium alone (33, 34).

Virulence genes provided a greater contrast of inheritance patterns, from those that are core to *S. enterica* to those that are randomly distributed. There were several instances in which major clades showed a preference for specific virulence genes, such as *sseI*/*srfH* being found more often in clade A than clade B. This effector-encoding gene is found on phage Gifsy-2, which shows a similar distribution pattern, which makes this a potential example of a phage genetic repertoire being crafted for the specific niche its host bacterium occupies (35, 36). This, among other patterns detailed here, hints that there are significant differences in the evolutionary pressures regarding pathogenicity for the different clades of *S. enterica* subsp. *enterica*.

For the SPI-2-secreted effectors alone, an *S. enterica* subsp. *enterica* effectorome has been described with 7 core (encoded in all genomes) and 21 non-core effectors (37). Even within those effectors proposed to be core, most of those represented in the VFDB had at least one strain in which the gene sequence was not found. For those that are variably present, some clade preferences were observed, such as the previously mentioned *sseI*. One pattern worth noting is that of *sspH2*, carried by a prophage remnant, which may have been deactivated early in clade A2 formation, an event that seems evolutionarily beneficial given *sspH2*'s frequency in the subclade (38). The study has omitted several effector-encoding genes not in the source database, including *steA* and *steD*, which are considered core effectors by at least one review and for which there is recent data on their role (37, 39, 40). Because the database takes into account several alleles, and these effectors were scored for closeness to any within, we considered but ultimately declined to include them and other alleles in our screen, but acknowledge their potential importance.

An interesting pattern of typhoid toxin gene presence within clade D was observed in which two putatively active subunit homologs, but not the porin, were present. This toxin is expected to be nonfunctional if this prediction is correct, but it remains unclear why this allele would be persistent in clade D if it is nonfunctional (26). A similar pattern was found in two additional strains within clade A2. It is possible that a separate, more degenerate porin-encoding sequence exists in these genomes. It may also be significant that the typhoid toxin-encoding genes appear to have been independently acquired within section Typhi.

Similar to some patterns seen in the T3SS-secreted effector proteins, high variability in the presence of fimbrial adherence gene operons was observed. The fimbrial adherence factor-encoding genes range from rarely present to ubiquitous, with some patterns of presence or absence within clades being revealed with expanded taxon sampling within clades, such as *tcf*, which is widely distributed across *S. enterica* subsp.

enterica, but more commonly found in clade B. There are several common fimbrial operons that appear to have been lost in section Typhi. Additionally, perhaps reflecting *S. enterica* serovar Typhi's uniqueness within section Typhi, the two *S. enterica* serovar Typhi strains appear to have aquired *sef* and *stc*, which were absent in all but one of the other section Typhi strains. The *sef* operon in strain CT18, however, has been reported to contain pseudogenes (41). Altogether, our findings underscore the uniqueness of *S. enterica* serovar Typhi, even juxtaposed with its closest phylogenetic relatives.

The rapid increase in the amount of bacterial whole-genome sequences will continue to precipitate improvements in the understanding of bacterial genomics. The impact is beginning to be felt beyond individual genomes: here, variable genetic information was used to gain insight broadly into the dynamics of horizontal gene transfer and gene populations within *S. enterica*. This census has revealed new information on phage, plasmid, and virulence factor inheritance and how to evaluate them as individual phylogenetic actors. The perspective gained through massive WGS will become increasingly important for the study of bacterial phylogenies and the genes that populate them.

MATERIALS AND METHODS

Taxon selection and genome assembly. Our goal was to capture both the serovar and underlying genomic diversity present in the *Salmonella* GenomeTrakr database while minimizing the overall size of our phylogenetic tree. To do this, we imposed a set of filters for initial inclusion into our data set: (i) serovar name or antigenic formula present in metadata and (ii) raw genome data collected on an Illumina sequencing machine (Illumina, San Diego, CA). From this initial pass, we chose a maximum of five isolates from each serovar. We then preformed *de novo* assemblies on each genome using SPAdes v.3.8.0 (42), discarding contigs shorter than 500 bases. Isolates that had less than 20× average coverage were removed. We then checked the assemblies and raw reads for contamination or mislabeling with Kraken v.0.10.5-beta (43) and removed sequences that returned less than 80% *Salmonella* hits under either condition. We performed an *in silico* check for the traditional serovar determination with the program SeqSero v.1.0 (21). When the original serotype disagreed with the SeqSero call, the phylogenetic context of the isolate was used to make a decision on which serovar call was more accurate. After we were comfortable that our data set contained high-quality sequence data and correctly typed isolates, we performed a quick SNP-based phylogenetic analysis using kSNP 3.0 (44) to further prune redundant taxa, leaving at most two representatives per serovar unless significant genetic diversity remained. We also created an assembly BLAST database from this final taxon set for other downstream analyses.

Core genome. A core genome was identified using tBLASTn (45) to search for each CDS in the *S. enterica* serovar Typhimurium strain LT2 genome (14) (GenBank accession no. NC_003197.2) from our database of assembled genomes. Orthologs were determined under the following thresholds: hits above 75% translated nucleotide sequence compared to the LT2 protein sequence, within 50% of the original length, and with an "expect" value of ≤0.005. Each CDS that was represented exactly once in each genome was used, and only those where each hit was the same length. This was conservatively called to reduce the proportion of highly modified or horizontally acquired genes used to calculate the main phylogeny.

Phylogeny. The genes were aligned with MUSCLE v.3.8.31 using default settings (46). RAxML was used to calculate the main phylogeny using a concatenation of the aligned gene sequences from the core genome with the GTRCAT + GAMMA model and 1,000 bootstraps, with a final maximum-likelihood search (47).

Structure. STRUCTURE v.2.3.3 was run on the core SNP matrix from kSNP using default settings for assumed populations of sizes 2 through 10 (20). A total of 50,000 replicates were run after a 10,000-rep burn-in period.

Phage detection. Assembled genomes were submitted to the PHASTER API (http://phaster.ca/phaster_api) on 29 and 30 April 2017 (23). Results were taken from the files returned by the server.

Plasmid detection. We identified putative plasmid sequences in our data set by BLAST searching the PlasmidFinder database replicons against our data set with BLASTn (accessed 24 April 2017) (33, 45). A sequence identity of 80% was used as a cutoff for calling a specific replicon present or absent.

Virulence factor identification. We identified putative virulence factors by searching for *Salmonella* sp. sequences from the Virulence Factor Database (accessed 18 April 2017) against our data set using tBLASTn (25). Sequences that were listed as belonging to the genus *Salmonella* were used in the query, including, in most cases, several different alleles. Only matches above 80% identity and longer than 60% of the query length were considered good hits. In the case of the effector-encoding genes *sspH1* and *sspH2*, we adjusted the identity cutoff to 90% to account for the high sequence similarity between these two alleles.

Data availability. For brevity, the sequences used in this study are listed in Data Set S1. Code used in this study is available at www.github.com/jnw29/Salmonella2018.

SUPPLEMENTAL MATERIAL

Supplemental material for this article may be found at https://doi.org/10.1128/mBio.02303-18.

FIG S1, PDF file, 0.3 MB.
FIG S2, PDF file, 0.8 MB.
FIG S3, PDF file, 0.3 MB.
FIG S4, PDF file, 0.1 MB.
FIG S5, PDF file, 0.2 MB.
DATA SET S1, XLSX file, 0.1 MB.
DATA SET S2, TXT file, 0.1 MB.
DATA SET S3, XLSX file, 2.6 MB.

ACKNOWLEDGMENTS

The authors thank the sequencing effort made by the GenomeTrakr labs, specifically the following labs whose data were used in this article: Ohio State and the International Congress on Pathogens at the Human Animal Interface (ICOPHAI), The Thakur Molecular Epidemiology Lab at North Carolina State University, Anita Wright's lab at the University of Florida, Hawaii State Department of Health, Minnesota Department of Health, NY Department of Health Wadsworth Center, Texas Department of State Health Services, Virginia Department of Health, Washington State Department of Health Public Health Laboratories, FDA/CFSAN Office of Applied Research and Safety Assessment, the National Microbial Resistance Monitoring System (NARMS), and the FDA/ORA field labs. We also thank Lili Velez for manuscript editing and our two reviewers, whose comments hugely improved our manuscript.

REFERENCES

1. Scallan E, Hoekstra RM, Angulo FJ, Tauxe RV, Widdowson M-A, Roy SL, Jones JL, Griffin PM. 2011. Foodborne illness acquired in the United States–major pathogens. Emerg Infect Dis 17:7–15. https://doi.org/10.3201/eid1701.091101p1.
2. CDC. 2017. National Enteric Disease Surveillance: Salmonella annual report, 2015. https://www.cdc.gov/nationalsurveillance/salmonella-surveillance.html.
3. Sangal V, Harbottle H, Mazzoni CJ, Helmuth R, Guerra B, Didelot X, Paglietti B, Rabsch W, Brisse S, Weill F-X, Roumagnac P, Achtman M. 2010. Evolution and population structure of *Salmonella enterica* serovar Newport. J Bacteriol 192:6465–6476. https://doi.org/10.1128/JB.00969-10.
4. Timme RE, Pettengill J, Allard MW, Strain E, Barrangou R, Wehnes C, Kessel JV, Karns J, 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. https://doi.org/10.1093/gbe/evt159.
5. Verma N, Reeves P. 1989. Identification and sequence of rfbS and rfbE, which determine antigenic specificity of group A and group D salmonellae. J Bacteriol 171:5694–5701. https://doi.org/10.1128/jb.171.10.5694-5701.1989.
6. Boyd EF, Wang FS, Whittam TS, Selander RK. 1996. Molecular genetic relationships of the salmonellae. Appl Environ Microbiol 62:804–808.
7. Desai PT, Porwollik S, Long F, Cheng P, Wollam A, Clifton SW, Weinstock GM, McClelland M. 2013. Evolutionary genomics of *Salmonella enterica* subspecies. mBio 4:e00579-12. https://doi.org/10.1128/mBio.00579-12.
8. Alikhan N-F, Zhou Z, Sergeant MJ, Achtman M. 2018. A genomic overview of the population structure of *Salmonella*. PLoS Genet 14:e1007261. https://doi.org/10.1371/journal.pgen.1007261.
9. den Bakker HC, Moreno Switt AI, Govoni G, Cummings CA, Ranieri ML, Degoricija L, Hoelzer K, Rodriguez-Rivera LD, Brown S, Bolchacova E, Furtado MR, Wiedmann M. 2011. Genome sequencing reveals diversification of virulence factor content and possible host adaptation in distinct subpopulations of *Salmonella enterica*. BMC Genomics 12:425. https://doi.org/10.1186/1471-2164-12-425.
10. McQuiston JR, Herrera-Leon S, Wertheim BC, Doyle J, Fields PI, Tauxe RV, Logsdon JMJ. 2008. Molecular phylogeny of the salmonellae: relationships among *Salmonella* species and subspecies determined from four housekeeping genes and evidence of lateral gene transfer events. J Bacteriol 190:7060–7067. https://doi.org/10.1128/JB.01552-07.
11. Brown EW, Mammel MK, LeClerc JE, Cebula TA. 2003. Limited boundaries for extensive horizontal gene transfer among *Salmonella* pathogens. Proc Natl Acad Sci U S A 100:15676–15681. https://doi.org/10.1073/pnas.2634406100.
12. Octavia S, Lan R. 2006. Frequent recombination and low level of clonality within *Salmonella enterica* subspecies I. Microbiology 152:1099–1108. https://doi.org/10.1099/mic.0.28486-0.
13. Bäumler AJ, Gilde AJ, Tsolis RM, van der Velden AW, Ahmer BM, Heffron F. 1997. Contribution of horizontal gene transfer and deletion events to development of distinctive patterns of fimbrial operons during evolution of *Salmonella* serotypes. J Bacteriol 179:317–322. https://doi.org/10.1128/jb.179.2.317-322.1997.
14. McClelland M, Sanderson K, Spieth J, Clifton S, Latreille P, Courtney L, Porwollik S, Ali J, Dante M, Du F, Hou S, Layman D, Leonard S, Nguyen C, Scott K, Holmes A, Grewal N, Mulvaney E, Ryan E, Sun H, Florea L, Miller W, Stoneking T, Nhan M, Waterston R, Wilson R. 2001. Complete genome sequence of *Salmonella enterica* serovar Typhimurium LT2. Nature 413:852–856. https://doi.org/10.1038/35101614.
15. Parkhill J, Dougan G, James KD, Thomson NR, Pickard D, Wain J, Churcher C, Mungall KL, Bentley SD, Holden MT, Sebaihia M, Baker S, Basham D, Brooks K, Chillingworth T, Connerton P, Cronin A, Davis P, Davies RM, Dowd L, White N, Farrar J, Feltwell T, Hamlin N, Haque A, Hien TT, Holroyd S, Jagels K, Krogh A, Larsen TS, Leather S, Moule S, O'Gaora P, Parry C, Quail M, Rutherford K, Simmonds M, Skelton J, Stevens K, Whitehead S, Barrell BG. 2001. Complete genome sequence of a multiple drug resistant *Salmonella enterica* serovar Typhi CT18. Nature 413:848–852. https://doi.org/10.1038/35101607.
16. Porwollik S, McClelland M. 2003. Lateral gene transfer in *Salmonella*. Microbes Infect 5:977–989. https://doi.org/10.1016/S1286-4579(03)00186-2.
17. Vernikos GS, Thomson NR, Parkhill J. 2007. Genetic flux over time in the *Salmonella* lineage. Genome Biol 8:R100. https://doi.org/10.1186/gb-2007-8-6-r100.
18. Nieto PA, Pardo-Roa C, Salazar-Echegarai FJ, Tobar HE, Coronado-Arrázola I, Riedel CA, Kalergis AM, Bueno SM. 2016. New insights about excisable pathogenicity islands in *Salmonella* and their contribution to

virulence. Microbes Infect 18:302–309. https://doi.org/10.1016/j.micinf.2016.02.001.
19. Allard MW, Strain E, Melka D, Bunning K, Musser SM, Brown EW, Timme R. 2016. Practical value of food pathogen traceability through building a whole-genome sequencing network and database. J Clin Microbiol 54:1975–1983. https://doi.org/10.1128/JCM.00081-16.
20. Pritchard JK, Stephens M, Donnelly P. 2000. Inference of population structure using multilocus genotype data. Genetics 155:945–959.
21. Zhang S, Yin Y, Jones MB, Zhang Z, Deatherage Kaiser BL, Dinsmore BA, Fitzgerald C, Fields PI, Deng X. 2015. *Salmonella* serotype determination utilizing high-throughput genome sequencing data. J Clin Microbiol 53:1685–1692. https://doi.org/10.1128/JCM.00323-15.
22. CDC. 2017. CDC FoodNet surveillance report (final data). Foodborne Diseases Active Surveillance Network FoodNet, CDC, Atlanta, GA. https://www.cdc.gov/foodnet/reports/index.html.
23. Arndt D, Grant JR, Marcu A, Sajed T, Pon A, Liang Y, Wishart DS. 2016. PHASTER: a better, faster version of the PHAST phage search tool. Nucleic Acids Res 44:W16–W21. https://doi.org/10.1093/nar/gkw387.
24. Yap K-P, Ho WS, Gan HM, Chai LC, Thong KL. 2016. Global MLST of *Salmonella* Typhi revisited in post-genomic era: genetic conservation, population structure, and comparative genomics of rare sequence types. Front Microbiol 7:270. https://doi.org/10.3389/fmicb.2016.00270.
25. Chen L, Zheng D, Liu B, Yang J, Jin Q. 2016. VFDB 2016: hierarchical and refined dataset for big data analysis—10 years on. Nucleic Acids Res 44:D694–D697. https://doi.org/10.1093/nar/gkv1239.
26. Spanò S, Ugalde JE, Galán JE. 2008. Delivery of a *Salmonella* Typhi exotoxin from a host intracellular compartment. Cell Host Microbe 3:30–38. https://doi.org/10.1016/j.chom.2007.11.001.
27. Achtman M, Wain J, Weill F-X, Nair S, Zhou Z, Sangal V, Krauland MG, Hale JL, Harbottle H, Uesbeck A, Dougan G, Harrison LH, Brisse S, S. enterica MLST Study Group. 2012. Multilocus sequence typing as a replacement for serotyping in *Salmonella enterica*. PLoS Pathog 8:e1002776. https://doi.org/10.1371/journal.ppat.1002776.
28. Roumagnac P, Weill F-X, Dolecek C, Baker S, Brisse S, Chinh NT, Le TAH, Acosta CJ, Farrar J, Dougan G, Achtman M. 2006. Evolutionary history of *Salmonella* Typhi. Science 314:1301–1304. https://doi.org/10.1126/science.1134933.
29. Holt KE, Parkhill J, Mazzoni CJ, Roumagnac P, Weill F-X, Goodhead I, Rance R, Baker S, Maskell DJ, Wain J, Dolecek C, Achtman M, Dougan G. 2008. High-throughput sequencing provides insights into genome variation and evolution in *Salmonella* Typhi. Nat Genet 40:987–993. https://doi.org/10.1038/ng.195.
30. Figueroa-Bossi N, Uzzau S, Maloriol D, Bossi L. 2001. Variable assortment of prophages provides a transferable repertoire of pathogenic determinants in *Salmonella*. Mol Microbiol 39:260–271. https://doi.org/10.1046/j.1365-2958.2001.02234.x.
31. Hardt WD, Urlaub H, Galán JE. 1998. A substrate of the centisome 63 type III protein secretion system of *Salmonella* Typhimurium is encoded by a cryptic bacteriophage. Proc Natl Acad Sci U S A 95:2574–2579. https://doi.org/10.1073/pnas.95.5.2574.
32. Switt AIM, Sulakvelidze A, Wiedmann M, Kropinski AM, Wishart DS, Poppe C, Liang Y. 2015. *Salmonella* phages and prophages: genomics, taxonomy, and applied aspects, p 237–287. *In* Eisenstark A, Schatten H (ed), Salmonella: methods and protocols. Springer-Humana Press, New York, NY.
33. Carattoli A, Zankari E, García-Fernández A, Voldby Larsen M, Lund O, Villa L, Møller Aarestrup F, Hasman H. 2014. In silico detection and typing of plasmids using PlasmidFinder and plasmid multilocus sequence typing. Antimicrob Agents Chemother 58:3895–3903. https://doi.org/10.1128/AAC.02412-14.
34. García P, Hopkins KL, García V, Beutlich J, Mendoza MC, Threlfall J, Mevius D, Helmuth R, Rodicio MR, Guerra B, Med-Vet-Net WP21 Project Group. 2014. Diversity of plasmids encoding virulence and resistance functions in Salmonella enterica subsp. enterica serovar Typhimurium monophasic variant 4,[5],12:i:− strains circulating in Europe. PLoS One 9:e89635. https://doi.org/10.1371/journal.pone.0089635.
35. Miao EA, Miller SI. 2000. A conserved amino acid sequence directing intracellular type III secretion by *Salmonella* Typhimurium. Proc Natl Acad Sci U S A 97:7539–7544. https://doi.org/10.1073/pnas.97.13.7539.
36. Worley MJ, Ching KHL, Heffron F. 2000. *Salmonella* SsrB activates a global regulon of horizontally acquired genes. Mol Microbiol 36:749–761.
37. Jennings E, Thurston TLM, Holden DW. 2017. *Salmonella* SPI-2 type III secretion system effectors: molecular mechanisms and physiological consequences. Cell Host Microbe 22:217–231. https://doi.org/10.1016/j.chom.2017.07.009.
38. Brussow H, Canchaya C, Hardt WD. 2004. Phages and the evolution of bacterial pathogens: from genomic rearrangements to lysogenic conversion. Microbiol Mol Biol Rev 68:560–602. https://doi.org/10.1128/MMBR.68.3.560-602.2004.
39. McQuate SE, Young AM, Silva-Herzog E, Bunker E, Hernandez M, de Chaumont F, Liu X, Detweiler CS, Palmer AE. 2017. Long-term live-cell imaging reveals new roles for *Salmonella* effector proteins SseG and SteA. Cell Microbiol 19:e12641. https://doi.org/10.1111/cmi.12641.
40. Bayer-Santos E, Durkin CH, Rigano LA, Kupz A, Alix E, Cerny O, Jennings E, Liu M, Ryan AS, Lapaque N, Kaufmann SHE, Holden DW. 2016. The *Salmonella* effector SteD mediates MARCH8-dependent ubiquitination of MHC II molecules and inhibits T cell activation. Cell Host Microbe 20:584–595. https://doi.org/10.1016/j.chom.2016.10.007.
41. Townsend SM, Kramer NE, Edwards R, Baker S, Hamlin N, Simmonds M, Stevens K, Maloy S, Parkhill J, Dougan G, Bäumler AJ. 2001. *Salmonella enterica* serovar Typhi possesses a unique repertoire of fimbrial gene sequences. Infect Immun 69:2894–2901. https://doi.org/10.1128/IAI.69.5.2894-2901.2001.
42. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol 19:455–477. https://doi.org/10.1089/cmb.2012.0021.
43. Wood DE, Salzberg SL. 2014. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol 15:R46. https://doi.org/10.1186/gb-2014-15-3-r46.
44. Gardner SN, Hall BG. 2013. When whole-genome alignments just won't work: kSNP v2 software for alignment-free SNP discovery and phylogenetics of hundreds of microbial genomes. PLoS One 8:e81760. https://doi.org/10.1371/journal.pone.0081760.
45. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. BMC Bioinformatics 10:421. https://doi.org/10.1186/1471-2105-10-421.
46. Edgar RC. 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5:113. https://doi.org/10.1186/1471-2105-5-113.
47. Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312–1313. https://doi.org/10.1093/bioinformatics/btu033.