Genomic Epidemiology of *Salmonella enterica* Serotype Enteritidis based on Population Structure of Prevalent Lineages

Xiangyu Deng, Prerak T. Desai, Henk C. den Bakker, Matthew Mikoleit, Beth Tolar, Eija Trees, Rene S. Hendriksen, Jonathan G. Frye, Steffen Porwollik, Bart C. Weimer, Martin Wiedmann, George M. Weinstock, Patricia I. Fields,[1] and Michael McClelland[1]

Salmonella enterica serotype Enteritidis is one of the most commonly reported causes of human salmonellosis. Its low genetic diversity, measured by fingerprinting methods, has made subtyping a challenge. We used whole-genome sequencing to characterize 125 *S*. enterica Enteritidis and 3 *S. enterica* serotype Nitra strains. Single-nucleotide polymorphisms were filtered to identify 4,887 reliable loci that distinguished all isolates from each other. Our whole-genome single-nucleotide polymorphism typing approach was robust for *S. enterica* Enteritidis subtyping with combined data for different strains from 2 different sequencing platforms. Five major genetic lineages were recognized, which revealed possible patterns of geographic and epidemiologic distribution. Analyses on the population dynamics and evolutionary history estimated that major lineages emerged during the 17th–18th centuries and diversified during the 1920s and 1950s.

Salmonella enterica causes ≈1 million illnesses and >350 deaths annually in the United States (*1*). Among >2,500 known serotypes, *S. enterica* serotype Enteritidis is one of the most commonly reported causes of human salmonellosis in most industrialized countries (*2*). From the 1970s through the mid-1990s, the incidence of serotype Enteritidis infection increased dramatically; shelled eggs were a major vehicle for transmission. Despite a decrease in serotype Enteritidis infection since 1996 in the United States, outbreaks resulting from contaminated eggs continue to occur (*3*), and Enteritidis remains among the most common serotypes isolated from humans worldwide (*2*). Epidemiologic surveillance and outbreak investigation of microbial pathogens require subtyping that provides sufficient resolution to discriminate closely related isolates. Differentiation of *S. enterica* Enteritidis challenges traditional subtyping methods, such as pulsed-field gel electrophoresis (PFGE), because isolates of serotype Enteritidis are more genetically homogeneous than are isolates of many other serotypes (*4*,*5*). Among the serotype Enteritidis isolates reported to PulseNet, ≈45% display a single PFGE *Xba*I pattern (JEGX01.0004), which renders PFGE ineffective in some investigations (*5*). Of the second-generation methods evaluated for *S. enterica* Enteritidis subtyping, multilocus variable number–tandem repeat analysis offers slightly better discrimination, but differentiating common patterns remains a substantial problem (*6*). Therefore, new methods are needed to better subtype and differentiate this serotype. Recent applications of whole-genome sequencing (WGS) have demonstrated exceptional resolution that enables fine delineation of infectious disease outbreaks (*7*–*10*).

In addition to sufficient subtyping resolution, accurately ascribing isolates to epidemiologically meaningful clusters, i.e., grouping isolates associated with an outbreak while discriminating unrelated strains, is critical for pathogen subtyping. Outbreak and epidemiologically unrelated isolates might not be differentiated by using current methods. Despite the high incidence of *S. enterica* Enteritidis

Author affiliations: University of Georgia, Griffin, Georgia, USA (X. Deng); University of California Irvine, Irvine, California, USA (P.T. Desai, S. Porwollik, M. McClelland); Cornell University, Ithaca, New York, USA (H.C. den Bakker, M. Wiedmann); Centers for Disease Control and Prevention, Atlanta, Georgia, USA (M. Mikoleit, B. Tolar, E. Trees, P.I. Fields); Technical University of Denmark, Lyngby, Denmark (R.S. Hendriksen); US Department of Agriculture, Athens, Georgia, USA (J.G. Frye); University of California Davis, Davis, California, USA (B.C. Weimer); Washington University School of Medicine, St. Louis, Missouri, USA (G.M. Weinstock);

DOI: http://dx.doi.org/10.3201/eid2009.131095

[1]These authors contributed equally to this article.

infection in humans, genome sequencing of this serotype has lagged behind sequencing of other major foodborne pathogens. To our knowledge, only 1 finished *S. enterica* Enteritidis genome is publicly available (*11*). Recent sequencing of *S. enterica* Enteritidis genomes of the common PFGE *Xba*I pattern JEGX01.0004 has provided a valuable resource on the *S. enterica* Enteritidis genome (*12*). Here we present a broad sampling of WGS to include diversity of other major lineages.

We expanded the genomic population structure of *S. enterica* Enteritidis by sequencing a collection of 81 *S. enterica* Enteritidis genomes and 3 *S. enterica* serotype Nitra genomes selected to capture epidemiologic and phylogenetic diversity in current domestic and international serotype Enteritidis populations. We included serotype Nitra in the study because it is thought to be a variant of serotype Enteritidis with its O antigen (serogroup O2) being a minor genetic variant of serogroup O9 found in serotype Enteritidis (*13*). These genomes, along with 44 draft genomes of *S. enterica* Enteritidis (14 historical strains and 30 isolates selected from the 2010 egg outbreak investigation [http://www.cdc.gov/salmonella/enteritidis/]), provided a phylogenetic framework of diverse circulating serotype Enteritidis lineages. Model-based Bayesian estimation of age and effective population size of major *S. enterica* Enteritidis lineages showed that the spreading of *S. enterica* Enteritidis coincided with 2 periods: the 18th century period of colonial trade and the 20th century period of agricultural industrialization. A single-nucleotide polymorphism (SNP) pipeline was developed for high-throughput whole-genome SNP typing and was robust for combining data from different sequencing platforms in the same analysis. This enabled retrospective investigation of recent clinical cases in Thailand and the shelled eggs outbreak in the United States. The ability of whole-genome SNP typing to infer the polyclonal genomic nature of at least some *S. enterica* Enteritidis strains causing outbreaks, despite high genetic homogeneity among *S. enterica* Enteritidis genomes, demonstrates the utility and sensitivity of whole-genome SNP typing in epidemiologic surveillance and outbreak investigations. Potential challenges of whole-genome SNP typing, such as ways to accurately define individual outbreaks, were discussed.

Methods

Isolates

We obtained 125 serotype Enteritidis and 3 serotype Nitra isolates from Centers for Disease Control and Prevention, US Department of Agriculture, and University of California Davis (online Technical Appendix Tables 1, 2, http://wwwnc.cdc.gov/EID/article/20/9/13-1095-Techapp1.pdf). *S. enterica* Enteritidis isolates of diverse PFGE subtypes (18 *Xba*I patterns accounting for >90% of all *S. enterica* Enteritidis isolates reported to PulseNet [online Technical Appendix Figure 1]), spatiotemporal origins, and sources were sampled to span a broad epidemiologic and phylogenetic diversity of prevalent lineages of which we were aware.

WGS

Bacterial strains were grown in Luria broth at 37°C to stationary phase. Genomic DNA was prepared by using the GenElute Genomic DNA isolation kit (Sigma-Aldrich, St. Louis, MO, USA). Eighty-one isolates were sequenced by using Illumina (San Diego, CA, USA) technology (100-bp paired-end reads) at Washington University (St. Louis, MO, USA). Another 44 isolates were sequenced by using Roche (Indianapolis, IN, USA) 454 technology (single-end reads) as described previously (*12*).

SNP Detection

We developed a bioinformatics pipeline to detect high-quality SNPs from raw sequencing reads. The design of the workflow was geared toward a customizable and robust solution for whole-genome SNP typing of many isolates. It enables user-defined parameters for SNP quality filters and provides additional functions, such as assembly of unmapped reads and functional annotation of SNPs (online Technical Appendix Figure 2). The program snp-sites was then used to code missing data and SNP sites from ambiguous sites within the consensus sequences and create an alignment containing variable sites (https://github.com/andrewjpage/snp_sites).

Phylogenetic Analyses

We used BratNextGen (*14*) to detect recombination events in the genomes. The consensus sequences were used as input with 100 replicates (10 iterations each) to infer the significance of detected recombination events. Regions with a significant signal of recombination were excluded, as were highly homoplastic sites (as inferred in PAUP 4.0b10 [Sinauer Associates, Inc., Sunderland, MA, USA]; rescaled consistency index <1) indicative of non-neutral evolution, recombination, or ambiguous SNP calls. The remaining SNP sites were used only for further analysis when unambiguously called for at least 95% of the isolates. We performed maximum-likelihood (ML) analyses in MEGA5.1 (*15*). The resulting ML trees were used to test for a temporal signal by using Path-O-Gen v1.3 (http://tree.bio.ed.ac.uk/software/pathogen/). Bayesian phylogenetic analyses were performed by using BEAST v. 1.7.5 (*16*). The isolation year of each isolate was used to establish a temporal framework for constructing phylogenetic relationship among the isolates and estimating parameters to describe the evolutionary dynamics of the population (*17*).

Comparisons of different molecular clock models and tree priors were performed similarly to a method of Bakker et al. (*18*), except that we used the path sampling method (*19*) to estimate the marginal likelihood (online Technical Appendix Table 4).

Results

Divergent Isolates and Serotypes

Comparison of 84 newly sequenced genomes to the reference genome showed that all but 4 were closely related to the reference strain, differing by no more than 950 SNPs. The 4 divergent genomes (77–0915, 07–0056, SARB17, and SARB19) contained 19,800–43,544 SNPs, comparable to the number of SNPs between phylogenetically distinct serotypes. They also lacked *sdf*, a characteristic marker of commonly circulating serotype Enteritidis organisms, and were phylogenetically apart from the main serotype Enteritidis lineage (online Technical Appendix Figure 3). We did not include these divergent genomovars (genetic lineages) in subsequent SNP and phylogenetic analyses. The 3 *S. enterica* Nitra genomes were highly similar to the reference genome, with numbers of SNPs comparable to those of other *S. enterica* Enteritidis strains.

High-Quality Core Genome SNPs and Phylogeny

We observed 6,542 SNP loci in the remaining strains after we excluded other genomovars. The pairwise homoplasy index test (*20*) found no evidence of recombination for the SNP data of both the 81-isolate set (Illumina sequenced isolates and the reference) and the 125-isolate set (Illumina and 454 sequenced isolates plus the reference). However, putative regions (14 in total, online Technical Appendix Table 3) involved in homologous recombination were detected by BratNextGen (*14*) in the 125-isolate set, comprising 1,519 SNPs. After exclusion of these regions encompassing recombination, and homoplastic sites (136 SNPs, identified by using PAUP 4.0), 4,887 core genome SNP loci were left to be included in the analysis.

The general time-reversible model of nucleotide substitution was the best fit model for the dataset and was subsequently used in phylogenetic analyses. ML analysis based on high-quality core genome SNPs yielded highly congruent phylogenies between the 81-isolate (Illumina data only) and the 125-isolate (Illumina and 454 data combined) datasets (Figure 1). All 454 sequenced isolates clustered in 1 lineage, including the 30 selected for the shelled eggs outbreak investigation. These isolates represented 8 of the 9 clades defined by Allard et al. (*12*). For the Illumina-sequenced isolates in both datasets, the inferred phylogenies were highly congruent (Figure 1). Five major genetic lineages were identified (Figure 1): LI, LII, LIII, LIV, and LV. Isolates from clinical cases in Thailand and associated with a shell egg outbreak in the United States were found predominately in LIII and LV, respectively.

Population Dynamics

The 125-isolate set displayed a temporal signal, as demonstrated by a positive correlation between distances to the most recent common ancestor (MRCA) and dates of sampling. Although this correlation was weak when measured for the whole dataset (correlation coefficient 0.3, $R^2 = 0.09$, $p<0.001$), exclusion of LII from the dataset led to an increased correlation (correlation coefficient 0.6, $R^2 = 0.40$, $p<0.0001$). Both the Path-O-Gen analysis and molecular clock–based analyses indicate that LII evolved at a higher mutation rate than other clades.

We compared 4 population genetics models through Bayes factors (BF) (*21*). For this analysis, we excluded redundant isolates that were derived from the same outbreak and differed by only 1 or 2 SNPs. The final dataset comprised 99 isolates. A relaxed log-normal molecular clock was strongly favored over a strict clock rate ($\log_{10}$ BF>100), suggesting that mutation rates vary significantly among branches (online Technical Appendix Table 4). We found strong evidence ($\log_{10}$ BF>100) in favor of a constant effective population size model over an effective population size model that enables fluctuations of effective population size through time (Gaussian Markov random field skyride model [*22*]). The results of the analyses assuming a constant effective population size model are thus discussed here.

The mean mutation rate across all lineages was inferred to be 2.2×10^{-7} substitutions per site per year or 1.01 SNPs per genome per year. The MRCA of the whole population (Table) was estimated to date to 1549 CE (95% highest posterior probability density, from 1351 to 1704) (*23*). Although the inferred ages of the MRCA differ because of the model of choice, the estimates for the younger nodes appeared to converge between models, with overlapping highest posterior probability densities. We constructed a lineage through time plot (Figure 2) to show the change of inferred number of lineages over time on the basis of a constant effective population size model using BEAST (http://tree.bio.ed.ac.uk/).

Discussion

WGS of the 125 *S. enterica* isolates of serotypes Enteritidis and Nitra enabled us to probe the population dynamics and evolutionary history of prevalent serotype Enteritidis lineages. The inferred mean mutation rates of serotype Enteritidis are comparable to those of short-term evolution in several other pathogens (*24–26*). The prediction that the MRCA possibly emerged between 1351 and 1704 CE is in line with the historical fact that serotype Enteritidis was one of the first recognized *Salmonella* serotypes in 1888 CE

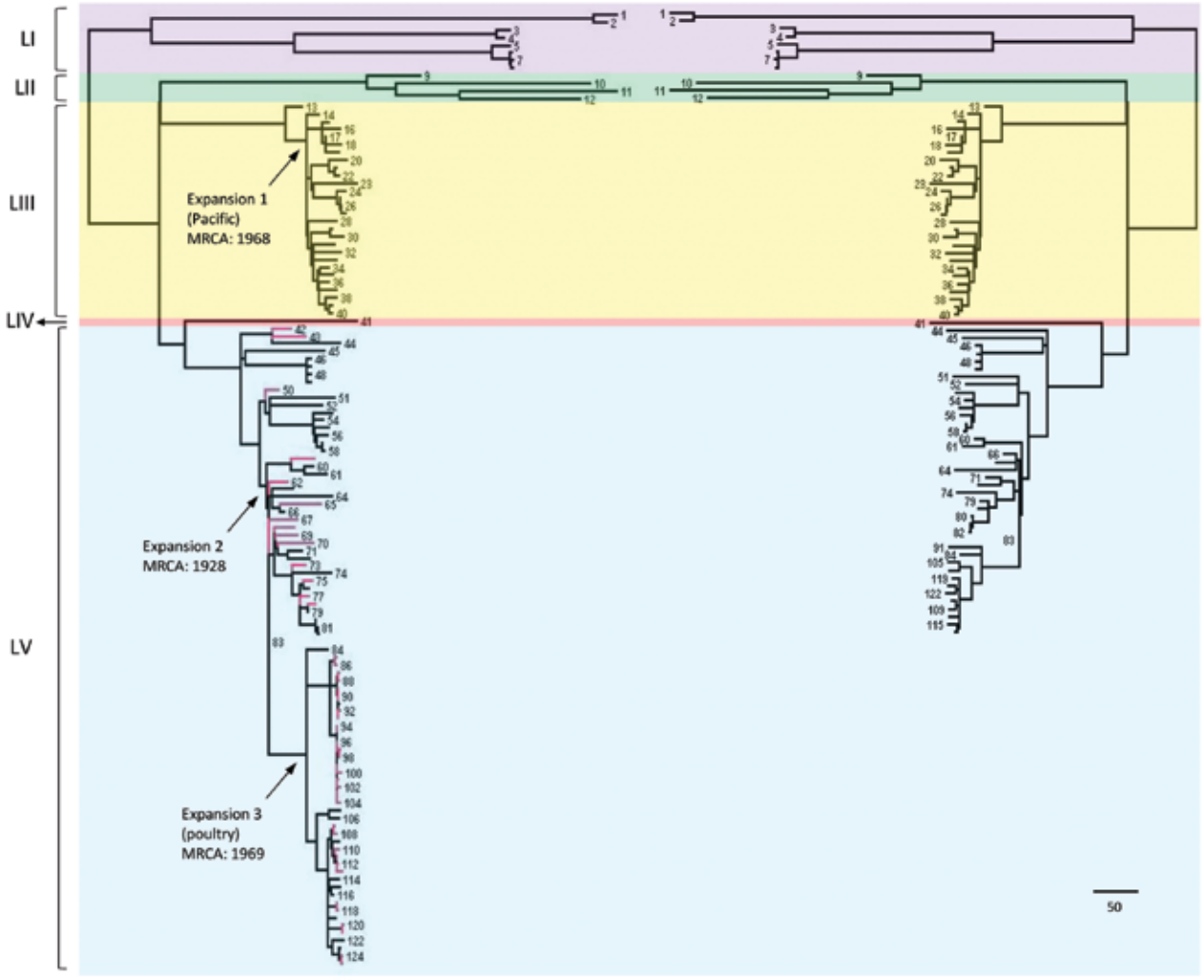


Figure 1. Comparison of *Salmonella enterica* serotype Enteritidis phylogenies inferred from Illumina data and combined data of Illumina (San Diego, CA, USA) and Roche 454 (Indianapolis, IN, USA). The tree on the right incudes 80 Illumina sequenced isolates and the reference genome (PT4). The tree on the left includes both the 80 Illumina and the 44 454 sequenced isolates in addition to the reference. Isolates were numbered (online Technical Appendix Table 1, http://wwwnc.cdc.gov/EID/article/20/9/13-1095-Techapp1.pdf). Lineages I, II, III, IV, and V are highlighted in purple, green, yellow, red, and blue, respectively. Branches representing 454 sequenced isolates are labeled in red. Arrows on the left tree indicate the 3 serotype Nitra isolates. MRCA, most recent common ancestor. Scale bar indicates single-nucleotide polymorphisms.

(*27*). According to the same model, the 5 lineages identified in this study originated during 1608–1788 CE (Table). This initial diversification event could possibly be related to the emergence of colonial trade around that time, facilitating effective global dispersal of serotype Enteritidis, as in the case of the intercontinental transmission of the yellow fever virus (*28*).

On the basis of the molecular clock and other assumptions made in the Bayesian analysis, the number of lineages of *S. enterica* Enteritidis is estimated to have increased sharply during 1925–1950 CE, which indicated rapid diversification of serotype Enteritidis (Figure 2). This period coincides with the Green Revolution, a period of increased global agricultural production. We speculate that the start of better poultry farming practices in the United Kingdom and United States might have created poultry farms as a niche for *S. enterica* Enteritidis by reducing closely related traditional avian *Salmonella* serotypes, such as Gallinarum (*29*).

The genes responsible for serotype (primarily *rfb* region, *fliC*, and *fljB*) are commonly subject to horizontal gene transfer, resulting in very similar/the same alleles present in distinct genetic backgrounds. This phenomenon has contributed to the huge diversity of *Salmonella* serotypes

Table. Dating of nodes in the maximum-likelihood tree of *Salmonella enterica* serotype Enteritidis using BEAST and strict and relaxed mutation rates*

	Median MRCA (95% CI)		
	Relaxed clock		
Node, cluster	Constant population size	GMRF	Strict clock, constant population size
MRCA, node N01	1549 (1351–1704)	1896 (1858–1933)	1520 (1455–1576)
MRCA II, III, IV, V	1709 (1608–1788)	NA	1680 (1636–1716)
MRCA II	1831 (1760–1897)	1937 (1903–1966)	1824 (1799–1846)
MRCA III	1941 (1919–1957)	1950 (1933–1962)	1934 (1924–1944)
MRCA V	1888 (1858–1911)	1917 (1900–1936)	1883 (1869–1897)
Expansion 1, Pacific	1968 (1958–1976)	1975 (1964–1982)	1964 (1958–1970)
Expansion 2	1928 (1916–1937)	1934 (1926–1943)	1926 (1919–1933)
Expansion 3, poultry	1969 (1956–1981)	1984 (1974–1995)	1970 (1964–1976)

*BEAST (*16*). MRCA, most recent common ancestor; GMRF, Gaussian Markov random field; NA, not available.

and dictates that the serotyping system sometimes poorly reflects true phylogeny. Multiple genomovars have been noted for some serotypes (*30*) and most likely arose from horizontal gene transfer of the same assortment of serotype antigen genes into distinct genetic lineages by coincidence. Four isolates serotyped as *S. enterica* Enteritidis but lacking *sdf*, a characteristic marker of commonly circulating *S. enterica* Enteritidis, were divergent by WGS; they putatively represent different genomovars. The divergent *S. enterica* Enteritidis lineages are rarely encountered in the United States.

In contrast to the observation of different genomovars within a single serotype, *S. enterica* Nitra represents a separate serotype that is embedded within serotype Enteritidis lineages. Three serotype Nitra isolates clustered with LIII (CDC-STK-1280 and 96–0186) and LIV (2010K-0860), indicating that they are members of these *S. enterica* Enteritidis lineages. This finding is consistent with the finding that serogroup O2 is a minor genetic variant of serogroup O9. The 3 *S. enterica* Nitra isolates originated from different geographic locations and decades apart, suggesting that rare, independent, and distinct mutational events caused the switch of serotype from Enteritidis to Nitra. In this instance, it seems appropriate to consider *S. enterica* Nitra as part of the *S. enterica* Enteritidis lineage. As scientists move toward genetic determination of serotype, *S. enterica* Nitra is likely to be identified as *S. enterica* Enteritidis (*13*).

As shown in this and previous studies, whole-genome SNP typing provides both superior subtyping resolution and phylogenetic accuracy without compromising either, which is rarely possible with traditional molecular subtyping methods. Whole-genome SNP typing achieves this resolution by surveying point mutations across entire genomes rather than by targeting relatively few polymorphic sites characteristic of either high mutation rates for sufficient discriminatory power (e.g., variable number tandem repeats [*31*] and clustered regularly interspaced short palindromic repeats [*32*]) or reliable phylogenetic signals for accurate phylogeny (e.g., housekeeping genes). We found

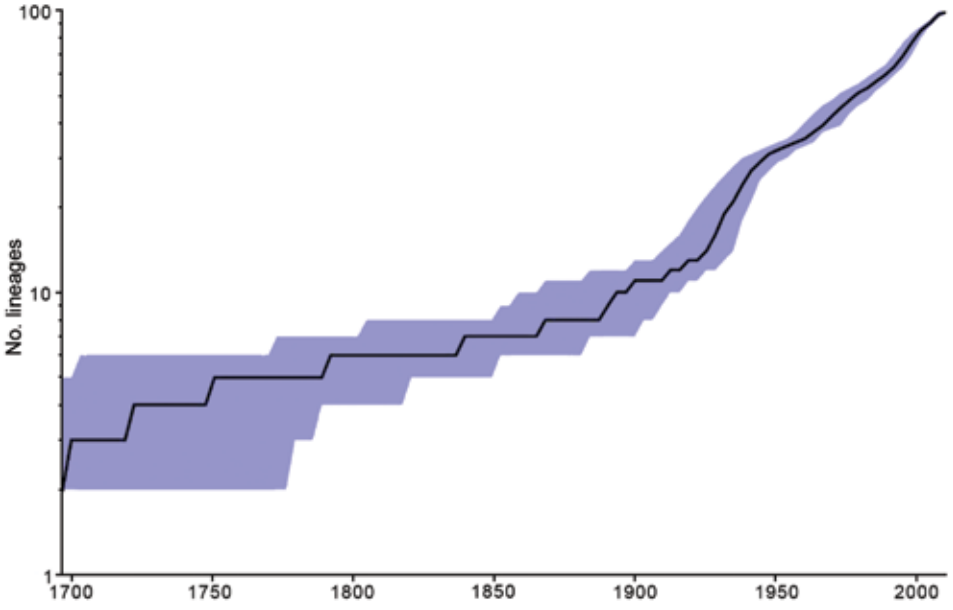


Figure 2. Inferred number of *Salmonella enterica* serotype Enteritidis lineages over time based on a constant effective population size model using BEAST (*16*). Blue shading indicates 95% CIs.

that the 4,887 SNPs resulting from combining data from 2 platforms were sufficient to resolve every serotype Enteritidis isolate in a phylogeny highly congruent with the tree based only on 1 platform (Figure 1). Such robustness facilitates the use of different sequencing platforms in actual surveillance and investigations to achieve repeatable results in subtype differentiation. Specifically, the same analytical approach and bioinformatics pipeline was applied to sequencing data with drastically different error patterns by employing strict criteria to search for high-quality SNPs. This strategy was effective and efficient for phylogenetic inference, subtyping, and routine investigation of serotype Enteritidis, especially when multiple instruments, library preparation, and bioinformatics methods are available for whole-genome SNP typing applications in public health laboratories.

WGS of *S. enterica* Enteritidis isolates with a wide range of genetic, spatiotemporal, and source attributes broadened the understanding of phylogenetic diversity of this genetically homogeneous pathogen. WGS enabled us to build a phylogenetic framework of prevalent serotype Enteritidis lineages that will be highly valuable for ongoing and future surveillance and investigation.

We recognized lineages and clades with geographic and epidemiologic characteristics. Of the 3 seemingly rare or potentially undersampled lineages (LI, LII, and LIV), LI and LII appeared to be associated with specific geographic locations and have diverged earlier than other lineages. LI was further divided into 2 clades, of which one had 2 isolates from Africa (isolates 1 and 2 in Figure 1) and the other had 6 isolates from the western United States and predominately associated with wildlife and environmental sources (isolates 3–8). LII consisted of 3 isolates associated with marine mammals in California and 1 isolate linked to a human in the same state (isolates 9–12). The 3 marine mammal isolates formed a highly supported clade and came from 2 different host species in a 10-year span, suggesting a stable *S. enterica* Enteritidis lineage circulating among those animals. The fact that a human case was linked to this clade suggests possible transmission between marine animals and humans. The marine animal community on and around the Channel Islands off the coast of southern California is extremely rich and diverse and displays the highest known prevalence of *S. enterica* in the world (*33*). Free-ranging and migratory animals and birds from this natural reservoir of *S. enterica* could play a role in long distance dispersal of this pathogen. LIII and LV appeared to be the more prevalent lineages on the international and US domestic scales, respectively. LIII contained isolates from 4 continents; its major clade (isolates 14–40) originated primarily from the Pacific region, including Thailand and California. LIV represented a common lineage in the United States, including a clade predominately associated with poultry products (isolates 84–125).

Our retrospective investigation of clinical cases from Thailand and a shelled eggs outbreak in the United States demonstrated the utility of our method on the basis of the robust bioinformatics pipeline and the broad phylogenetic framework. Fourteen of the 15 isolates from clinical cases in Thailand clustered in LIII and fell into 4 different clusters with high bootstrapping support (Figure 3, clusters A–D), suggesting multiple genetic origins consistent with a previous study (*34*). Isolates from blood and fecal samples clustered with other isolates from the United States and Europe, including the reference strain P125109 from United Kingdom. Previous observation of a disproportionately high percentage of bloodstream infections of *S. enterica* Enteritidis in Thailand (*35*) may be due to host factors (e.g., underlying health conditions, concurrent infections), underreported or unreported gastrointestinal infections less detectable than invasive ones, and/or random invasive infections (e.g., high-dosage exposure). Similarly, compromised immunity from the high percentage of HIV cases in sub-Saharan Africa was proposed as a contributing factor to the perceived high invasiveness of serotype Enteritidis infections in the region (*36*). We recommend caution when interpreting extraintestinal infections of *Salmonella* as evidence for high virulence and that newly proposed

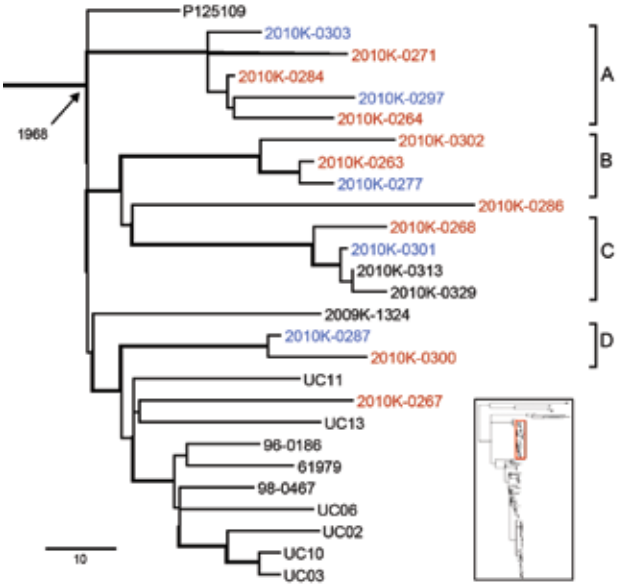


Figure 3. Salmonella enterica serotype Enteritidis genetic lineage III. Isolates from human blood and stool samples are indicated by red and blue, respectively. Four clades are highlighted and designated A–D, representing individual outbreak clades identified from clinical cases in Thailand in 2008. Branches with bootstrap value >0.9 are indicated by thickened lines. Age of the ancestral node (median most recent common ancestor) is labeled. Inset indicates the location of the highlighted lineage on the global phylogenetic tree.

hypervirulent lineages (*37*) be evaluated within a global phylogenetic context for their evolutionary identity and origin.

Concerning the shelled eggs outbreak, 2 definite (Figure 4, clusters A and B, both including isolates traced to the implicated facility) and 3 putative outbreak clades (clusters C–E, none of which had a direct epidemiologic link to the outbreak), i.e., individual clusters each containing isolates originated from a single source of contamination, were identified with phylogenetic and/or epidemiologic support on the basis of 2 criteria: 1) the clade is phylogenetically highly supported and 2) the isolation dates of all the isolates in the clade correspond to the outbreak period. Three isolates from sporadic cases in 2009 and 2010 might be attributed to the outbreak because they fell into the outbreak clades A and E (Figure 4, underlined), suggesting that the outbreak strains might have been circulating before the outbreak. Although A corresponded to a major clade defined by Allard et al. (*12*), B, C, and D clustered and thus were designated as a single clade in that study, possibly because of the absence of reference strains to separate them. Among the 4 isolates associated with poultry (Figure 4, blue highlighted), 60277 and 85366 were respectively isolated in 2002 and 2007 and therefore unlikely to be associated with each other and with the outbreak in 2010. These and other isolates unrelated to the outbreak helped delineate the individual outbreak clades, corroborating the likely polyclonal nature of the outbreak, which also was recognized by Allard et al. (*12*), and emphasizing the importance of incorporating multiple proper background references into outbreak investigations.

During outbreak investigations, putative outbreak isolates are analyzed with epidemiologically unrelated strains to determine whether they are related and thus likely to be associated with the same source. The availability and selection of background references can be critical (online Technical Appendix Figure 4). To maximize the availability of suitable background reference datasets, researchers desire phylogenetic frameworks with sufficient coverage of the genetic diversity in major pathogens, which is an aim of the ongoing 100K pathogen genome project (http://100kgenome.vetmed.ucdavis.edu/index.cfm).

Phylogenetic data alone are insufficient for defining an outbreak. Determining the polyclonal nature and scope of an outbreak remains a challenge, and no definitive criteria yet exist. For example, investigations of a recent *S. enterica* serotype Montevideo outbreak associated with red and black peppers reached discrepant conclusions. A proposed domestic source isolate from the implicated food processing facility (*38*) was excluded from the

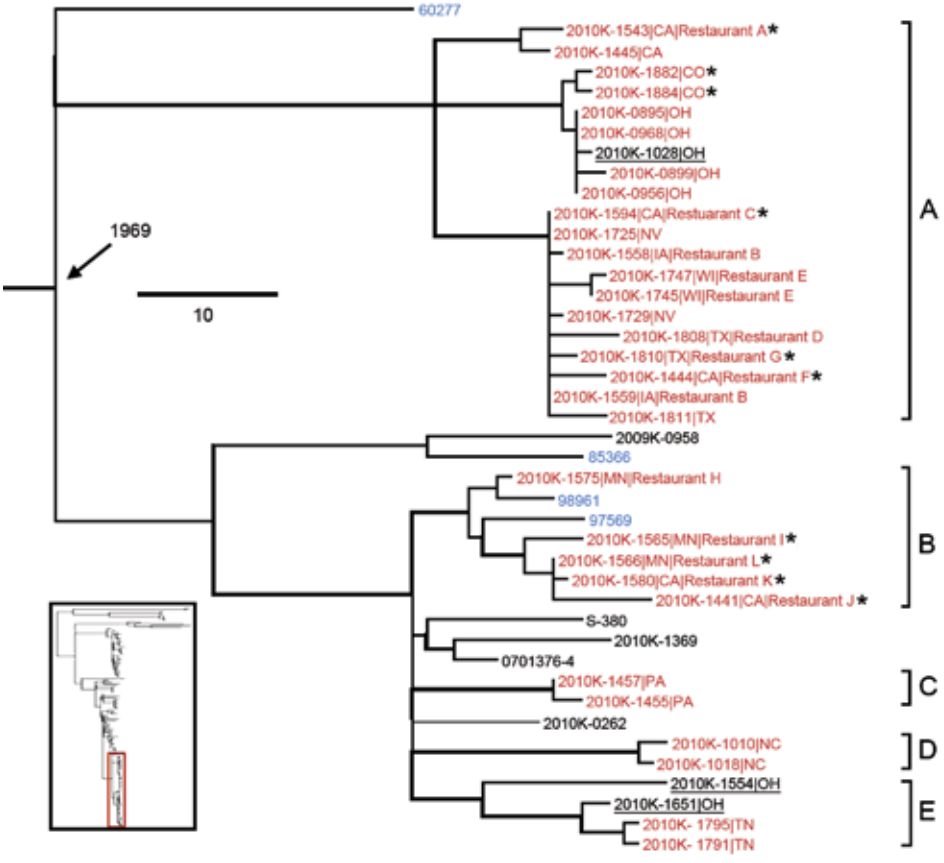


Figure 4. Salmonella enterica serotype Enteritidis clades associated with the 2010 US shelled eggs outbreak. Red indicates isolates sequenced as part of the outbreak investigation using Roche 454 technology (Indianapolis, IN, USA); blue indicates isolates associated with chicken or chicken products; asterisk (*) indicates the isolate was traced back to the implicated egg production facility. Five outbreak clades are highlighted and designated as A–E, of which A and B are definite and C, D, and E are putative. Isolates ascribed to the egg outbreak in this study are underlined. Branches with bootstrap values >0.9 are shown by thickened lines. Age of the ancestral node (median most recent common ancestor) is labeled. Scale bar indicates single-nucleotide polymorphisms. Inset indicates the location of the highlighted lineage on the global phylogenetic tree..

outbreak clade defined in another study (*18*), presumably because of differences in the definition of the scope of the outbreak between the 2 studies. Although a polyclonal outbreak is among the explanations for this discrepancy, the possibility of other scenarios cannot be rejected by available phylogenomic and epidemiologic evidence. For instance, in a case of continuous or intermittent outbreak, a single persistent founder strain can shed divergent descendants that contaminate food and cause disease over an extended time, as reported by Orsi et al. (*39*). Such microevolution events give rise to clones that might or might not be considered as distinct genotypes or separate outbreak clades depending on levels of mutation, epidemiologic information, or even subjective interpretation. Therefore, case-by-case understanding of evolutionary dynamics and population structure of major foodborne pathogens (*40*), which might vary among different species and be affected by particular food-processing environments and outbreak vehicles, is necessary for elucidating population genetics and transmission dynamics of foodborne pathogens and lays the groundwork for the increasing application of genomic epidemiology in pathogen surveillance and outbreak investigation. Ultimately, if some strains in an outbreak have continued to evolve quickly, then the ability to cluster strains and identify outbreaks will rely even more on a suitable set of outgroups.

Acknowledgments

We thank the PulseNet participating laboratories, Thailand Ministry of Public Health, and the Central Health Laboratory Mauritius for contributing strains used in this study. We thank Mark Allard, Errol Strain, and Eric Brown for providing the 454 sequencing data and editorial suggestions on the manuscript. We thank Jean Guard for many helpful discussions.

X.D. was supported in part by an American Society for Microbiology/Centers for Disease Control and Prevention Fellowship and startup funds from University of Georgia. M.M., S.P., and P.D. were supported in part by National Institutes of Health grant nos. AI039557 AI052237, AI073971, AI075093, AI077645 AI083646, and HHSN272200900040C; US Department of Agriculture (USDA) grant nos. 2009-03579 and 2011-67017-30127; the Binational Agricultural Research and Development Fund; and a grant from the Center for Produce Safety. J.G.F was supported by USDA Agricultural Research Services project no. 6612-32000-006-00. H.dB. and M.W. were supported in part by USDA grant no. 2010-34459-20756

Dr Deng is an assistant professor at the Center for Food Safety, University of Georgia, and a guest researcher at the Enteric Diseases Laboratory Branch, Centers for Disease Control and Prevention. His research interests focus on using genomic and molecular biology approaches to better understand the biology, transmission, and evolution of foodborne pathogens.

References

1. Scallan E, Hoekstra RM, Angulo FJ, Tauxe RV, Widdowson MA, Roy SL, et al. Foodborne illness acquired in the United States—major pathogens. Emerg Infect Dis. 2011;17:7–15. http://dx.doi.org/10.3201/eid1701.P11101
2. Hendriksen RS, Vieira AR, Karlsmose S, Lo Fo Wong DM, Jensen AB, Wegener HC, et al. Global monitoring of *Salmonella* serovar distribution from the World Health Organization Global Foodborne Infections Network Country Data Bank: results of quality assured laboratories from 2001 to 2007. Foodborne Pathog Dis. 2011;8:887–900. http://dx.doi.org/10.1089/fpd.2010.0787
3. Braden CR. *Salmonella enterica* serotype Enteritidis and eggs: a national epidemic in the United States. Clin Infect Dis. 2006;43:512–7. http://dx.doi.org/10.1086/505973
4. Olson AB, Andrysiak AK, Tracz DM, Guard-Bouldin J, Demczuk W, Ng LK, et al. Limited genetic diversity in *Salmonella enterica* serovar Enteritidis PT13. BMC Microbiol. 2007;7:87. http://dx.doi.org/10.1186/1471-2180-7-87
5. Zheng J, Keys CE, Zhao S, Meng J, Brown EW. Enhanced subtyping scheme for *Salmonella* Enteritidis. Emerg Infect Dis. 2007;13:1932–5. http://dx.doi.org/10.3201/eid1312.070185
6. Boxrud D, Pederson-Gulrud K, Wotton J, Medus C, Lyszkowicz E, Besser J, et al. Comparison of multiple-locus variable-number tandem repeat analysis, pulsed-field gel electrophoresis, and phage typing for subtype analysis of *Salmonella enterica* serotype Enteritidis. J Clin Microbiol. 2007;45:536–43. http://dx.doi.org/10.1128/JCM.01595-06
7. Harris SR, Feil EJ, Holden MT, Quail MA, Nickerson EK, Chantratita N, et al. Evolution of MRSA during hospital transmission and intercontinental spread. Science. 2010;327:469–74. http://dx.doi.org/10.1126/science.1182395
8. Rasko DA, Webster DR, Sahl JW, Bashir A, Boisen N, Scheutz F, et al. Origins of the *E. coli* strain causing an outbreak of hemolytic-uremic syndrome in Germany. N Engl J Med. 2011;365:709–17. http://dx.doi.org/10.1056/NEJMoa1106920
9. Eppinger M, Mammel MK, Leclerc JE, Ravel J, Cebula TA. Genomic anatomy of *Escherichia coli* O157:H7 outbreaks. Proc Natl Acad Sci U S A. 2011;108:20142–7. http://dx.doi.org/10.1073/pnas.1107176108
10. Chin CS, Sorenson J, Harris JB, Robins WP, Charles RC, Jean-Charles RR, et al. The origin of the Haitian cholera outbreak strain. N Engl J Med. 2011;364:33–42. http://dx.doi.org/10.1056/NEJMoa1012928
11. Thomson NR, Clayton DJ, Windhorst D, Vernikos G, Davidson S, Churcher C, et al. Comparative genome analysis of *Salmonella* Enteritidis PT4 and *Salmonella* Gallinarum 287/91 provides insights into evolutionary and host adaptation pathways. Genome Res. 2008;18:1624–37. http://dx.doi.org/10.1101/gr.077404.108
12. Allard MW, Luo Y, Strain E, Pettengill J, Timme R, Wang C, et al. On the evolutionary history, population genetics and diversity among isolates of *Salmonella* Enteritidis PFGE pattern JEGX01.0004. PLoS ONE. 2013;8:e55254. http://dx.doi.org/10.1371/journal.pone.0055254
13. Fitzgerald C, Collins M, van Duyne S, Mikoleit M, Brown T, Fields P. Multiplex, bead-based suspension array for molecular determination of common *Salmonella* serogroups. J Clin Microbiol. 2007;45:3323–34. http://dx.doi.org/10.1128/JCM.00025-07
14. Marttinen P, Hanage WP, Croucher NJ, Connor TR, Harris SR, Bentley SD, et al. Detection of recombination events in bacterial genomes from large population samples. Nucleic Acids Res. 2012;40:e6. http://dx.doi.org/10.1093/nar/gkr928
15. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol. 2011;28:2731–9. http://dx.doi.org/10.1093/molbev/msr121

16. Drummond AJ, Rambaut A. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol Biol. 2007;7:214. http://dx.doi.org/10.1186/1471-2148-7-214
17. Okoro CK, Kingsley RA, Connor TR, Harris SR, Parry CM, Al-Mashhadani MN, et al. Intracontinental spread of human invasive *Salmonella* Typhimurium pathovariants in sub-Saharan Africa. Nat Genet. 2012;44:1215–21. http://dx.doi.org/10.1038/ng.2423
18. Bakker HC, Switt AI, Cummings CA, Hoelzer K, Degoricija L, Rodriguez-Rivera LD, et al. A whole-genome single nucleotide polymorphism–based approach to trace and identify outbreaks linked to a common *Salmonella enterica* subsp. *enterica* serovar Montevideo pulsed-field gel electrophoresis type. Appl Environ Microbiol. 2011;77:8648–55. http://dx.doi.org/10.1128/AEM.06538-11
19. Baele G, Li WL, Drummond AJ, Suchard MA, Lemey P. Accurate model selection of relaxed molecular clocks in Bayesian phylogenetics. Mol Biol Evol. 2013;30:239–43. http://dx.doi.org/10.1093/molbev/mss243
20. Bruen TC, Philippe H, Bryant D. A simple and robust statistical test for detecting the presence of recombination. Genetics. 2006;172:2665–81. http://dx.doi.org/10.1534/genetics.105.048975
21. Suchard MA, Weiss RE, Sinsheimer JS. Bayesian selection of continuous-time Markov chain evolutionary models. Mol Biol Evol. 2001;18:1001–13. http://dx.doi.org/10.1093/oxfordjournals.molbev.a003872
22. Gill MS, Lemey P, Faria NR, Rambaut A, Shapiro B, Suchard MA. Improving Bayesian population dynamics inference: a coalescent-based model for multiple loci. Mol Biol Evol. 2013;30:713–24. http://dx.doi.org/10.1093/molbev/mss265
23. Drummond AJ, Ho SY, Phillips MJ, Rambaut A. Relaxed phylogenetics and dating with confidence. PLoS Biol. 2006;4:e88. http://dx.doi.org/10.1371/journal.pbio.0040088
24. Morelli G, Didelot X, Kusecek B, Schwarz S, Bahlawane C, Falush D, et al. Microevolution of *Helicobacter pylori* during prolonged infection of single hosts and within families. PLoS Genet. 2010;6:e1001036. http://dx.doi.org/10.1371/journal.pgen.1001036
25. Didelot X, Eyre DW, Cule M, Ip CL, Ansari MA, Griffiths D, et al. Microevolutionary analysis of *Clostridium difficile* genomes to investigate transmission. Genome Biol. 2012;13:R118. http://dx.doi.org/10.1186/gb-2012-13-12-r118
26. Zhou Z, McCann A, Litrup E, Murphy R, Cormican M, Fanning S, et al. Neutral genomic microevolution of a recently emerged pathogen, *Salmonella enterica* serovar Agona. PLoS Genet. 2013;9:e1003471. http://dx.doi.org/10.1371/journal.pgen.1003471
27. Salmonella Subcommittee of the Nomenclature Committee of the International Society for Microbiology. The genus *Salmonella* Lignieres. 1900. J Hyg (Lond). 1934;34:333–50.
28. Bryant JE, Holmes EC, Barrett AD. Out of Africa: a molecular perspective on the introduction of yellow fever virus into the Americas. PLoS Pathog. 2007;3:e75. http://dx.doi.org/10.1371/journal.ppat.0030075
29. Bäumler AJ, Hargis BM, Tsolis RM. Tracing the origins of *Salmonella* outbreaks. Science. 2000;287:50–2. http://dx.doi.org/10.1126/science.287.5450.50
30. Porwollik S, Boyd EF, Choy C, Cheng P, Florea L, Proctor E, et al. Characterization of *Salmonella enterica* subspecies I genovars by use of microarrays. J Bacteriol. 2004;186:5883–98. http://dx.doi.org/10.1128/JB.186.17.5883-5898.2004
31. Hyytiä-Trees E, Smole SC, Fields PA, Swaminathan B, Ribot EM. Second generation subtyping: a proposed PulseNet protocol for multiple-locus variable-number tandem repeat analysis of Shiga toxin–producing *Escherichia coli* O157 (STEC O157). Foodborne Pathog Dis. 2006;3:118–31. http://dx.doi.org/10.1089/fpd.2006.3.118
32. Liu F, Barrangou R, Gerner-Smidt P, Ribot EM, Knabel SJ, Dudley EG. Novel virulence gene and clustered regularly interspaced short palindromic repeat (CRISPR) multilocus sequence typing scheme for subtyping of the major serovars of *Salmonella enterica* subsp. *enterica.* Appl Environ Microbiol. 2011;77:1946–56. http://dx.doi.org/10.1128/AEM.02625-10
33. Stoddard RA, DeLong RL, Byrne BA, Jang S, Gulland FM. Prevalence and characterization of *Salmonella* spp. among marine animals in the Channel Islands, California. Dis Aquat Organ. 2008;81:5–11. http://dx.doi.org/10.3354/dao01905
34. Hendriksen RS, Hyytia-Trees E, Pulsrikarn C, Pornruangwong S, Chaichana P, Svendsen CA, et al. Characterization of *Salmonella enterica* serovar Enteritidis isolates recovered from blood and stool specimens in Thailand. BMC Microbiol. 2012;12:92. http://dx.doi.org/10.1186/1471-2180-12-92
35. Hendriksen RS, Bangtrakulnonth A, Pulsrikarn C, Pornruangwong S, Noppornphan G, Emborg HD, et al. Risk factors and epidemiology of the ten most common *Salmonella* serovars from patients in Thailand: 2002–2007. Foodborne Pathog Dis. 2009;6:1009–19. http://dx.doi.org/10.1089/fpd.2008.0245
36. Graham SM. Nontyphoidal salmonellosis in Africa. Curr Opin Infect Dis. 2010;23:409–14. http://dx.doi.org/10.1097/QCO.0b013e32833dd25d
37. Heithoff DM, Shimp WR, House JK, Xie Y, Weimer BC, Sinsheimer RL, et al. Intraspecies variation in the emergence of hyperinfectious bacterial strains in nature. PLoS Pathog. 2012;8:e1002647. http://dx.doi.org/10.1371/journal.ppat.1002647
38. Lienau EK, Strain E, Wang C, Zheng J, Ottesen AR, Keys CE, et al. Identification of a salmonellosis outbreak by means of molecular sequencing. N Engl J Med. 2011;364:981–2. http://dx.doi.org/10.1056/NEJMc1100443
39. Orsi RH, Borowsky ML, Lauer P, Young SK, Nusbaum C, Galagan JE, et al. Short-term genome evolution of *Listeria monocytogenes* in a non-controlled environment. BMC Genomics. 2008;9:539. http://dx.doi.org/10.1186/1471-2164-9-539
40. Wilson MR, Allard MW, Brown EW. The forensic analysis of foodborne bacterial pathogens in the age of whole-genome sequencing. Cladistics. 2013;29:449–61. http://dx.doi.org/10.1111/cla.12012

Address for correspondence: Xiangyu Deng, Center for Food Safety, University of Georgia, 1109 Experiment St, Griffin, GA 30223-1797, USA; email: xdeng@uga.edu

Article DOI: http://dx.doi.org/10.3201/eid2009.131095

Genomic Epidemiology of *Salmonella enterica* Serotype Enteritidis based on Population Structure of Prevalent Lineages

Technical Appendix

Technical Appendix Table 1. Isolates used in this study*

Isolate†	Serotype	Location	PFGE-*Xba*I‡	Year	Source	Sequencer
2009K-0477 (1)	Enteritidis	Uganda	JEGX01.0767	2009	Human	Illumina
2009K-0479 (2)	Enteritidis	Uganda	JEGX01.0767	2009	Human	Illumina
77320 (3)	Enteritidis	Western USA	JEGX01.0031	2004	Animal (deer)	Illumina
81748 (4)	Enteritidis	Western, USA	JEGX01.0048	2005	Animal (wild mammal)	Illumina
04–0307 (5)	Enteritidis	Arizona, USA	NA	2004	Almond processing plant	Illumina
J0915 (6)	Enteritidis	California, USA	NA	2001	Environment (drag swab)	Illumina
J0903 (7)	Enteritidis	California, USA	NA	2001	Environment (drag swab)	Illumina
J0828 (8)	Enteritidis	NA	NA	2001	Human	Illumina
UC12 (9)	Enteritidis	California, USA	NA	2002	Human blood	Illumina
0811210F (10)	Enteritidis	California, USA	NA	2008	Animal (sea lion)	Illumina
9810102B (11)	Enteritidis	California, USA	NA	1998	Animal (sea otter)	Illumina
502571 (12)	Enteritidis	California, USA	NA	2005	Animal (sea lion)	Illumina
CDC-STK-1280 (13)	Nitra	Slovakia	NA	1965	NA	Illumina
P125109 (14)	Enteritidis	United Kingdom	NA	1988	Human	Reference
2010K-0303 (15)	Enteritidis	Thailand	JEGX01.0158	2008	Human feces	Illumina
2010K-0271 (16)	Enteritidis	Thailand	JEGX01.0008	2008	Human blood	Illumina
2010K-0284 (17)	Enteritidis	Thailand	JEGX01.0167	2008	Human blood	Illumina
2010K-0297 (18)	Enteritidis	Thailand	JEGX01.0002	2008	Human feces	Illumina
2010K-0264 (19)	Enteritidis	Thailand	JEGX01.0019	2008	Human blood	Illumina
2010K-0302 (20)	Enteritidis	Thailand	JEGX01.0158	2008	Human blood	Illumina
2010K-0263 (21)	Enteritidis	Thailand	JEGX01.0002	2008	Human blood	Illumina
2010K-0277 (22)	Enteritidis	Thailand	JEGX01.0002	2008	Human feces	Illumina
2010K-0286 (23)	Enteritidis	Thailand	JEGX01.0019	2008	Human blood	Illumina
2010K-0268 (24)	Enteritidis	Thailand	JEGX01.0158	2008	Human blood	Illumina
2010K-0301 (25)	Enteritidis	Thailand	JEGX01.0158	2008	Human feces	Illumina
2010K-0313 (26)	Enteritidis	Mauritius	JEGX01.0004	2010	Human blood	Illumina
2010K-0329 (27)	Enteritidis	Mauritius	JEGX01.0004	2010	Human feces	Illumina
2009K-1324 (28)	Enteritidis	Kansas, USA	NA	2009	Human	Illumina
2010K-0287 (29)	Enteritidis	Thailand	JEGX01.0167	2008	Human feces	Illumina
2010K-0300 (30)	Enteritidis	Thailand	JEGX01.0019	2008	Human blood	Illumina
UC11 (31)	Enteritidis	California, USA	NA	2001	Human blood	Illumina
2010K-0267 (32)	Enteritidis	Thailand	JEGX01.0019	2008	Human blood	Illumina
UC13 (33)	Enteritidis	California, USA	NA	2009	Human blood	Illumina
96–0186 (34)	Nitra	California, USA	NA	1996	Human urine	Illumina
61979 (35)	Enteritidis	Northern plains, USA	JEGX01.0002	2002	Food ready-to-eat	Illumina
98–0467 (36)	Enteritidis	Wyoming, USA	NA	1998	Human blood	Illumina
UC06 (37)	Enteritidis	NA	NA	2004	Human blood	Illumina
UC02 (38)	Enteritidis	California, USA	NA	2004	Human feces	Illumina
UC10 (39)	Enteritidis	California, USA	NA	2004	Human	Illumina
UC03 (40)	Enteritidis	California, USA	NA	2005	Human	Illumina
2009K-1726 (41)	Enteritidis	Maryland, USA	JEGX01.0009	2009	Human	Illumina
56–3991 (42) [C8]	Enteritidis	Tennessee, USA	NA	1956	Human	Roche 454
62–1976 (43) [C8]	Enteritidis	Massachusetts, USA	NA	1962	Human	Roche 454
2010K-0860 (44)	Nitra	Maryland, USA	NA	2010	Human feces	Illumina
93–0063 (45)	Enteritidis	Puerto Rico	NA	1993	Human feces	Illumina
2010K-2029 (46)	Enteritidis	Illinois, USA	JEGX01.0002	2010	Human	Illumina
2010K-0345 (47)	Enteritidis	Mauritius	JEGX01.0004	2010	Human brain abscess	Illumina
2010K-1923 (48)	Enteritidis	Maryland, USA	JEGX01.0004	2010	Human	Illumina
2010K-0351 (49)	Enteritidis	Mauritius	JEGX01.0004	2010	Human feces	Illumina

Isolate†	Serotype	Location	PFGE-*Xba*I‡	Year	Source	Sequencer
50–3079 (50) [C7]	Enteritidis	New Jersey, USA	NA	1950	NA	Roche 454
61080 (51)	Enteritidis	Northern plains, USA	JEGX01.0004	2002	Animal (cow)	Illumina
75788 (52)	Enteritidis	Northern plains, USA	JEGX01.0003	2004	Animal (reptile)	Illumina
08–1080 (53)	Enteritidis	New Mexico, USA	JEGX01.0030	2008	Human feces	Illumina
82631 (54)	Enteritidis	Southeastern, USA	JEGX01.0111	2006	Food (chicken)	Illumina
2010K-1832 (55)	Enteritidis	Oregon, USA	JEGX01.0005	2010	Human	Illumina
0804789B (56)	Enteritidis	California, USA	NA	2008	Animal (elephant seal)	Illumina
78296 (57)	Enteritidis	Northern plains, USA	JEGX01.0030	2005	Food (chicken)	Illumina
S-277 (58)	Enteritidis	NA	JEGX01.0005	2008	Environment (scald water)	Illumina
53–0407 (59) [C6]	Enteritidis	California, USA	NA	1953	Food (turkey)	Roche 454
98–9534 (60)	Enteritidis	California, USA	NA	1998	Animal (dog)	Illumina
1102933A (61)	Enteritidis	California, USA	NA	2011	Animal (elephant seal)	Illumina
77–2659 (62) [C3]	Enteritidis	South Dakota, USA	NA	1977	Human	Roche 454
93–7741 (63)	Enteritidis	California, USA	NA	1993	Animal (horse)	Illumina
02–2966 (64)	Enteritidis	California, USA	NA	2002	Animal (rodent)	Illumina
76–2651 (65) [C4]	Enteritidis	Maryland, USA	NA	1976	Human	Roche 454
54–2220 (66)	Enteritidis	Illinois, USA	NA	1954	Human feces	Illumina
78–1757 (67) [C5]	Enteritidis	Nebraska, USA	NA	1978	Human	Roche 454
50–5646 (68) [C3]	Enteritidis	Texas, USA	NA	1950	NA	Roche 454
77–0424 (69) [C3]	Enteritidis	Arizona, USA	NA	1977	Human	Roche 454
81–2490 (70) [C3]	Enteritidis	New Jersey, USA	NA	1981	Human	Roche 454
84–1226 (71)	Enteritidis	New York, USA	NA	1984	Human urine	Illumina
35986 (72)	Enteritidis	NA	JEGX01.0021	1998	Food (chicken)	Illumina
81–2625 (73) [C3]	Enteritidis	New Mexico, USA	NA	1981	Human	Roche 454
39997 (74)	Enteritidis	Southeastern, USA	JEGX01.0037	1999	Food (chicken)	Illumina
76–3618 (75) [C3]	Enteritidis	Arizona, USA	NA	1976	Human	Roche 454
76–0331 (76)	Enteritidis	Saudi Arabia	NA	1976	Human feces	Illumina
69–4941 (77) [C3]	Enteritidis	New Hampshire, USA	NA	1969	Human	Roche 454
77–1427 (78) [C3]	Enteritidis	Rhode Island, USA	NA	1977	Human	Roche 454
75–2732 (79)	Enteritidis	New Hampshire, USA	NA	1975	Human feces	Illumina
93–2836A (80)	Enteritidis	California, USA	NA	1993	Animal (elephant seal)	Illumina
93–6175B (81)	Enteritidis	California, USA	NA	1993	Animal (sea lion)	Illumina
93–7922A (82)	Enteritidis	California, USA	NA	1993	Animal (elephant seal)	Illumina
55795(83)	Enteritidis	Southeastern, USA	JEGX01.0005	2001	Food (chicken)	Illumina
60277 (84)	Enteritidis	Northeastern USA	JEGX01.0004	2002	Food (egg yolks)	Illumina
2010K-1543 (85) [C1]	Enteritidis	California, USA	JEGX01.0004	2010	Human	Roche 454
2010K-1445 (86) [C1]	Enteritidis	California, USA	JEGX01.0004	2010	Human	Roche 454
2010K-1882 (87) [C1]	Enteritidis	Colorado, USA	JEGX01.0005	2010	Human	Roche 454
2010K-1884 (88) [C1]	Enteritidis	Colorado, USA	JEGX01.0004	2010	Food ready-to-eat	Roche 454
2010K-0895 (89) [C1]	Enteritidis	Ohio, USA	JEGX01.0004	2010	Human	Roche 454
2010K-0968 (90) [C1]	Enteritidis	Ohio, USA	JEGX01.0004	2010	Human	Roche 454
2010K-1028 (91)	Enteritidis	Ohio, USA	JEGX01.0004	2010	Human	Illumina
2010K-0899 (92) [C1]	Enteritidis	Ohio, USA	JEGX01.0004	2010	Food ready-to-eat	Roche 454
2010K-0956 (93) [C1]	Enteritidis	Ohio, USA	JEGX01.0004	2010	Human	Roche 454
2010K-1594 (94) [C1]	Enteritidis	California, USA	JEGX01.0004	2010	Human	Roche 454
2010K-1725 (95) [C1]	Enteritidis	Nevada, USA	JEGX01.0004	2010	Human	Roche 454
2010K-1558 (96) [C1]	Enteritidis	Iowa, USA	JEGX01.0004	2010	Human	Roche 454
2010K-1747 (97) [C1]	Enteritidis	Wisconsin, USA	JEGX01.0004	2010	Human	Roche 454
2010K-1745 (98) [C1]	Enteritidis	Wisconsin, USA	JEGX01.0004	2010	Human	Roche 454
2010K-1729 (99) [C1]	Enteritidis	Nevada, USA	JEGX01.0004	2010	Human	Roche 454
2010K-1808 (100) [C1]	Enteritidis	Texas, USA	JEGX01.0004	2010	Human	Roche 454
2010K-1810 (101) [C1]	Enteritidis	Texas, USA	JEGX01.0004	2010	Human	Roche 454
2010K-1444 (102) [C1]	Enteritidis	California, USA	JEGX01.0004	2010	Human	Roche 454
2010K-1559 (103) [C1]	Enteritidis	Iowa, USA	JEGX01.0004	2010	Human	Roche 454
2010K-1811 (104) [C1]	Enteritidis	Texas, USA	JEGX01.0004	2010	Human	Roche 454
2009K-0958 (105)	Enteritidis	Massachusetts, USA	JEGX01.0021	2009	Human	Illumina
85366 (106)	Enteritidis	Western USA	JEGX01.0021	2007	Food (ground chicken)	Illumina
2010K-1575 (107) [C2]	Enteritidis	Minnesota, USA	JEGX01.0004	2010	Human	Roche 454
98961 (108)	Enteritidis	Southeastern USA	JEGX01.0004	2010	Food (chicken	Illumina
97569 (109)	Enteritidis	Northern plains, USA	JEGX01.0004	2010	Food (ground chicken)	Illumina
2010K-1565 (110) [C2]	Enteritidis	Minnesota, USA	JEGX01.0004	2010	Human	Roche 454
2010K-1566 (111) [C2]	Enteritidis	Minnesota, USA	JEGX01.0004	2010	Human	Roche 454
2010K-1580 (112) [C2]	Enteritidis	Minnesota, USA	JEGX01.0004	2010	Human	Roche 454
2010K-1441 (113) [C2]	Enteritidis	California, USA	JEGX01.0004	2010	Human	Roche 454
S-380 (114)	Enteritidis	NA	JEGX01.0003	2008	Environment (foam)	Illumina
2010K-1369 (115)	Enteritidis	Iowa, USA	JEGX01.0034	2010	Human feces	Illumina

Isolate†	Serotype	Location	PFGE-*Xba*I‡	Year	Source	Sequencer
0701376–4 (116)	Enteritidis	California, USA	NA	2007	Animal (sea otter)	Illumina
2010K-1457 (117) [C2]	Enteritidis	Pennsylvania, USA	JEGX01.0004	2010	Human	Roche 454
2010K-1455 (118) [C2]	Enteritidis	Pennsylvania, USA	JEGX01.0004	2010	Human	Roche 454
2010K-0262 (119)	Enteritidis	Thailand	JEGX01.0019	2008	Human feces	Illumina
2010K-1010 (120) [C2]	Enteritidis	North Carolina, USA	JEGX01.0108	2010	Food ready-to-eat	Roche 454
2010K-1018 (121) [C2]	Enteritidis	North Carolina, USA	JEGX01.0004	2010	Human	Roche 454
2010K-1554 (122)	Enteritidis	Iowa, USA	JEGX01.0004	2010	Human	Illumina
2009K-1651 (123)	Enteritidis	California, USA	JEGX01.0034	2009	Human	Illumina
2010K-1795 (124) [C2]	Enteritidis	Tennessee, USA	JEGX01.0004	2010	Human	Roche 454
2010K-1791 (125) [C2]	Enteritidis	Tennessee, USA	JEGX01.0004	2010	Human	Roche 454
77–0915§	Enteritidis	New Zealand	NA	1977	NA	Illumina
07–0056§	Enteritidis	Minnesota, USA	NA	2006	Human blood	Illumina
SARB17§	Enteritidis	Brazil	NA	NA	NA	454/Sanger¶
SARB19§	Enteritidis	Switzerland	NA	NA	NA	Illumina¶

*PFGE, pulsed-field gel electrophoresis; NA, not available,
†Isolates are numbered (in parenthesis) from the top down as they appear on the 125-isolate tree in Figure 1. For each 454 sequenced isolate also analyzed in (*12*), its cluster affiliation is noted in brackets as defined in (*12*).‡PulseNet PFGE pattern names might be subject to change.
§Genetically divergent isolates that are excluded from phylogenetic analysis.
¶Desai PT, Porwollik S, Long F, Cheng P, Wollam A, Bhonagiri-Palsikar V, et al. Evolutionary genomics of *Salmonella enterica* subspecies. mBio. 2013;4:e00198–13.

Technical Appendix Table 2. New genomes sequenced for this study

Isolate	GenBank accession no.	Sequencing coverage
UC07	PRJNA168935	64.5×
07–0056	PRJNA168933	61.2×
60277	PRJNA168942	64.8×
97569	PRJNA168772	63.6×
85366	PRJNA168767	55.6×
61979	PRJNA168769	64.3×
98961	PRJNA168768	97.6×
78296	PRJNA168766	57.4×
82631	PRJNA168792	66.7×
81748	PRJNA168793	99.0×
77320	PRJNA168794	90.1×
61080	PRJNA168770	75.2×
39997	PRJNA168791	64.0×
55795	PRJNA168764	77.0×
34986	PRJNA168771	88.8×
2010K-0313	PRJNA168735	56.0×
2010K-0329	PRJNA168736	56.5×
2010K-0345	PRJNA168753	64.8×
2010K-0351	PRJNA168754	50.7×
77–0915	PRJNA168760	60.7×
76–0331	PRJNA168759	65.2×
2010K-0284	PRJNA168728	79.8×
2010K-0287	PRJNA168730	60.2×
2010K-0297	PRJNA168731	56.2×
2010K-0300	PRJNA168732	55.6×
2010K-0302	PRJNA168733	95.5×
2010K-0303	PRJNA168734	60.5×
2010K-0268	PRJNA168750	61.0×
2010K-0301	PRJNA168752	98.3×
2010K-0271	PRJNA168727	52.6×
2010K-0286	PRJNA168729	98.2×
2010K-0264	PRJNA168749	59.2×
2010K-0262	PRJNA168725	60.5×
2010K-0267	PRJNA168726	66.2×
2010K-0263	PRJNA168748	55.7×
2010K-0277	PRJNA168751	51.9×
2009K-0477	PRJNA168745	61.9×
2009K-0479	PRJNA168746	57.9×
04–0307	PRJNA168744	57.6×
2009K-1651	PRJNA168723	57.6×
J0903	PRJNA168762	69.2×
J0915	PRJNA168795	56.1×

Isolate	GenBank accession no.	Sequencing coverage
93–6175B	PRJNA168773	95.6×
93–2836A	PRJNA168774	98.7×
93–7741	PRJNA168775	93.9×
93–7922A	PRJNA168776	92.3×
98–9534	PRJNA168777	99.4×
02–2966	PRJNA168778	98.7×
502571	PRJNA168779	98.6×
0701376–4	PRJNA168780	99.0×
0804789B	PRJNA168781	98.6×
0811210F	PRJNA168782	99.1×
1102933A	PRJNA168783	99.4×
9810102B	PRJNA168784	99.4×
54–2220	PRJNA168739	57.2×
2010K-1554	PRJNA168757	99.4×
2009K-1324	PRJNA168747	65.0×
2009K-1726	PRJNA168724	67.6×
2010K-1923	PRJNA168737	64.6×
75–2732	PRJNA168740	57.4×
08–1080	PRJNA168721	96.8×
84–1226	PRJNA168741	56.4×
2010K-1028	PRJNA168755	98.6×
2010K-1832	PRJNA168758	63.3×
93–0063	PRJNA168742	95.5×
UC02	PRJNA168785	49.8×
UC03	PRJNA168786	55.7×
UC10	PRJNA168787	59.5×
UC11	PRJNA168788	62.4×
UC12	PRJNA168789	56.8×
UC13	PRJNA168790	60.8×
2009K-0958	PRJNA168722	56.7×
2010K-1369	PRJNA168756	55.4×
98–0467	PRJNA168743	49.5×
S-380	PRJNA168763	74.1×
S-277	PRJNA168765	69.7×
2010K-2029	PRJNA168738	74.7×
J0828	PRJNA168761	66.2×
CDC-STK −1280	PRJNA168845	47.6×
96–0186	PRJNA168844	54.2×
2010K-0860	PRJNA168843	56.3×
UC06	PRJNA168940	64.2×

Appendix Table 3. Regions flagged as putatively involved in recombination by BratNextGen (Helsinki, Finland) and excluded from analyses

Region	Start	Stop	Locus tags in the reference genome
1	1	31813	SEN0001-SEN0027
2	66159	69974	SEN0057-SEN0061
3	468453	471763	SEN0418-SEN0420
4	570412	590805	SEN0509-SEN0530
5	774386	774422	SEN0697
6	784432	812238	SEN0706-SEN0732
7	924608	926567	SEN0841
8	1000063	1044312	SEN0899-SEN0940
9	1602387	1725731	SEN1501-SEN1621
10	1986102	2093597	SEN1881-SEN2005
11	3402494	3402716	SEN3185
12	4070069	4070244	SEN_r016
13	4232858	4403391	SEN3930-SEN4072
14	4576106	4578296	SEN4254

*P125109**;** NCBI reference sequence: NC_011294.1.

Technical Appendix Table 4. Comparison of 4 models used for phylogenetic analysis*

	Marginal likelihood	Model 1	Model 2	Model 3	Model 4
Model 1	−34183.35	**–**	**79.14**	**–172.99**	**–290.17**
Model 2	−34262.49	**–79.14**	**–**	**–252.13**	**–369.31**
Model 3	−34010.36	**172.99**	**252.13**	**–**	**–117.18**
Model 4	−33893.18	**290.17**	**369.31**	**117.18**	**–**

*Marginal likelihoods and Bayes factors (in bold) are shown. Model 1: strict clock and constant effective populations size; Model 2: strict clock and Gaussian Markov random field (GMRF) skyride; Model 3: relaxed clock and GMRF skride; Model 4: relaxed clock and constant effective population size. For Bayesian phylogenetic analysis, isolation year of each isolate was used as the tip date to calibrate estimated phylogenetic trees. Three different model combinations were tested: 1) strict molecular clock (applying a constant substitution rate across the entire tree) with constant effective populations size; 2) strict molecular clock with constant effective population size; and 3) lognormal relaxed clock (allowing for different substitution rates on different lineages) with GMRF skyride model; and 4) lognormal relaxed clock with constant effective population size. Every combination was run in 3 independent replicates for 100 million generations each, and model parameters and trees were logged every 10,000 generations. The results of the first 15 million generations (roughly the burn-in for most runs) were discarded. If convergence and proper mixing (as confirmed by Tracer 1.5, http://tree.bio.ed.ac.uk/software/tracer/) was achieved for each individual run, the results of the replicate runs were combined for further analysis. For the parameter-rich GMRF model, 5 independent runs of 100 million generations were necessary to obtain proper convergence and mixing. In this case the replicate runs were combined for further analysis. Bayes factors (BF) for each model combination were calculated based on the marginal likelihood values calculated using the path sampling method.

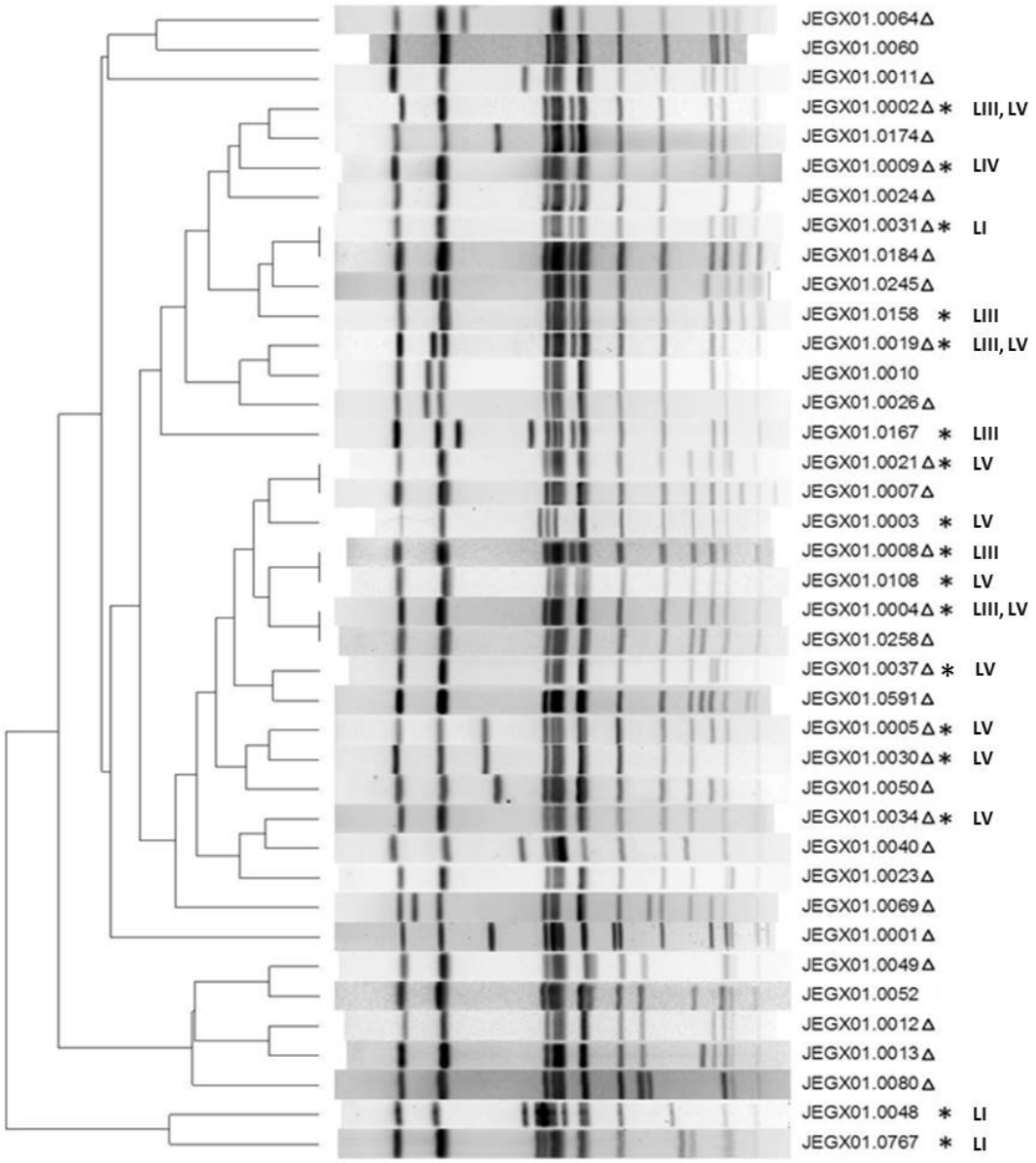
References

1. Kislyuk AO, Katz LS, Agrawal S, Hagen MS, Conley AB, Jayaraman P, et al. A computational genomics pipeline for prokaryotic sequencing projects. Bioinformatics. 2010;26:1819–26. PubMed http://dx.doi.org/10.1093/bioinformatics/btq284

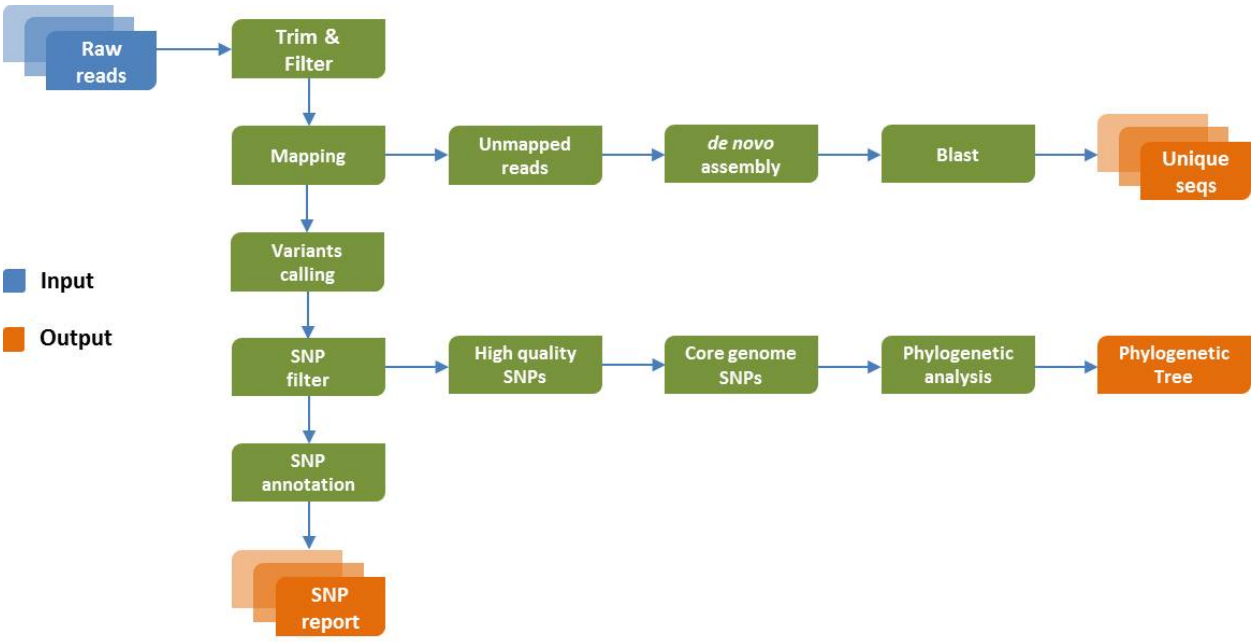
2. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9:357–9. PubMed http://dx.doi.org/10.1038/nmeth.1923

3. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25:2078–9. PubMed http://dx.doi.org/10.1093/bioinformatics/btp352

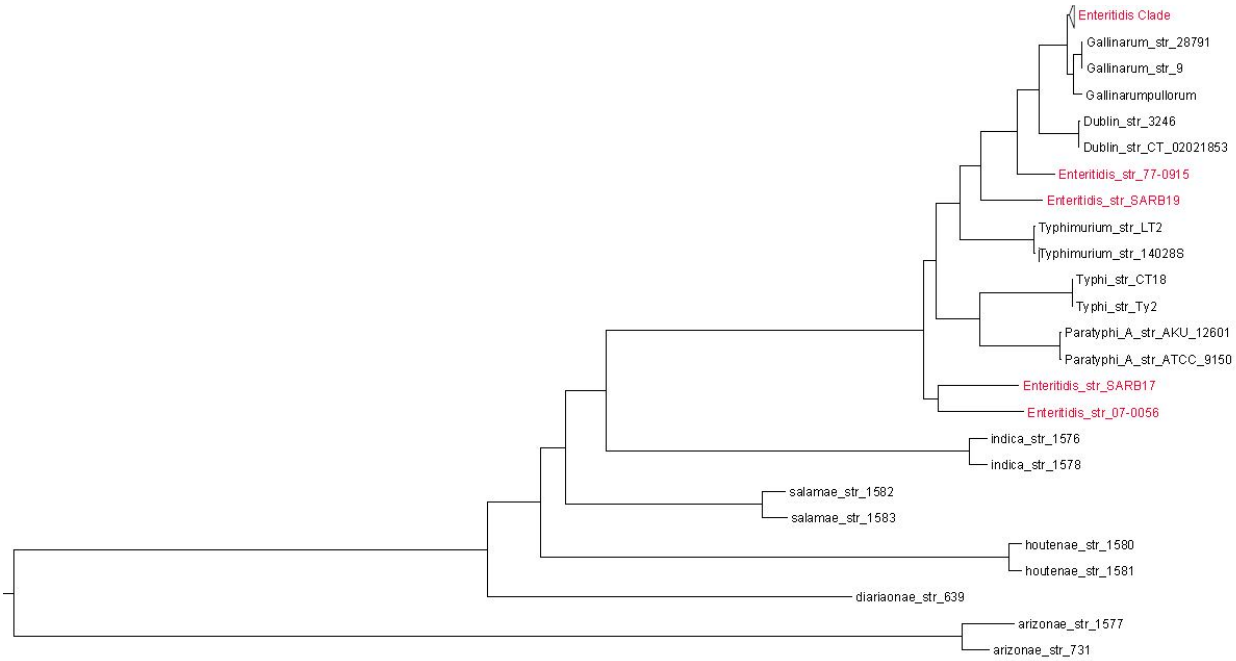
4. Cock PJ, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics. 2009;25:1422–3. PubMed http://dx.doi.org/10.1093/bioinformatics/btp163



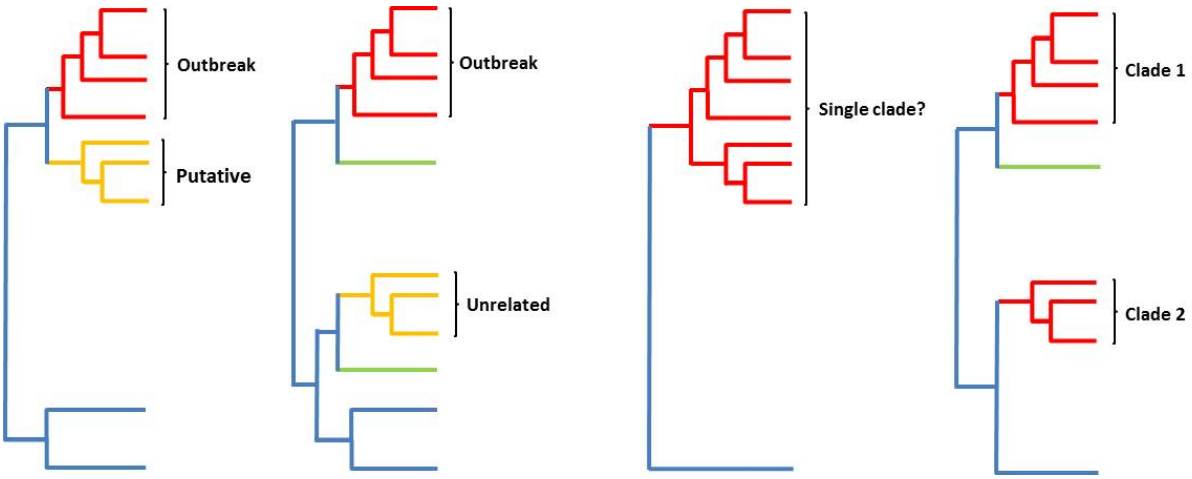
Appendix Figure 1. Dendrogram of top 30 most prevalent *Xba*I pulsed-field gel electrophoresis (PFGE) patterns and patterns selected for sequencing in this study. Δ indicates a top 30 most prevalent pattern. * indicates a sequenced pattern. Corresponding lineage(s) of each PFGE pattern is labeled.



Appendix Figure 2. Bioinformatics pipeline for single-nucleotide polymorphism (SNP) detection. In this study, each raw read in fastq format was trimmed from both ends until whichever of the 2 criteria was met: 1) the end (i.e., the trimmed part) reached an average Phred quality score of 35 or 2) a maximum of 20 bp at each end was removed (*1*). After trimming, any read below an average Phred score of 30 or a minimum length of 62 bp was removed. Trimmed and filtered reads were then mapped to *Salmonella enterica* serotype Enteritidis reference genome (P125109**;** NCBI reference sequence: NC_011294.1) with Bowtie 2 using default settings (*2*). For genomes sequenced by Illumina (n = 81) and 454 (n = 44), the average depth of mapping was 60× and 14×, respectively. Variants calls (SNPs, insertions and deletions) and consensus sequences were generated from BAM files by using SAMtools Mpileup (*3*), after BAM file sorting and removal of potential PCR duplicates. For each genome analyzed, a list of high-quality SNPs was derived by subjecting initial SNP calls to a set of quality filters including a minimum Phred base score of 60, a minimum read mapping score of 20, a mapping depth ranging from 5 to 100 reads per locus, a maximum alternate allele percentage of 25% and SNPs were accepted only when confirmed by reads mapped to both the forward and reverse strands. An optional step of SNP annotation was performed to include functional information such as amino acid change if a nonsynonymous point mutation occurred in a coding region. The pipeline was written in Python by using Biopython modules (*4*). All scripts are available upon request. The program snp-sites was then used to code missing data and SNP sites from ambiguous sites within the consensus sequences and create a multifasta alignment containing variable sites (https://github.com/andrewjpage/snp_sites).



Appendix Figure 3. Core genome single-nucleotide polymorphism maximum likelihood tree of various *Salmonella* serotypes. The main *Salmonella enterica* serotype Enteritidis lineage and the 4 divergent serotype Enteritidis isolates are highlighted in red.



Appendix Figure 4. Importance of reference isolates for whole-genome sequence typing (WGST) outbreak investigation. The outbreak clade is shown in red. Putative and epidemiologically unrelated clades are shown in yellow. Green nodes represent reference isolates. First, the use of proper reference may help exclude isolates that are unrelated to the outbreak under investigation. Without the reference strains highlighted in green in Appendix Figure 4A, the putative outbreak clade highlighted in yellow would cluster with the outbreak clade and appear to be more genetically homogeneous that it actually is. However, with the proper reference strains in place, the yellow clade is most likely a distinct lineage. Second, co-analyzing with a proper set of reference strains can help to separate outbreak isolates into individual clades and uncover multiple sources of contamination in the case of a polyclonal outbreak where multiple strains are simultaneously involved as shown in Appendix Figure 4B. Although phylogeny may help delineate outbreaks, it should not be used as the sole basis for ascribing/excluding isolates to/from an outbreak.