Recent Outbreaks of Shigellosis in California Caused by Two Distinct Populations of *Shigella sonnei* With Increased Virulence or Fluoroquinolone Resistance

Varvara K. Kozyreva[a], Guillaume Jospin[b], Alexander L. Greninger[a], James P. Watt[c] , Jonathan A. Eisen[b], Vishnu Chaturvedi[a]

Microbial Diseases Laboratory[a], Division of Communicable Disease Control[c], California Dept. of Public Health, Richmond, CA; Genome Center[b], Department of Evolution and Ecology, Department of Medical Microbiology and Immunology, University of California, Davis, CA

Running Head: Distinct *Shigella sonnei* populations in California

Corresponding authors: Vishnu Chaturvedi, Vishnu.Chaturvedi@cdph.ca.gov; Varvara K. Kozyreva, Varvara.Kozyreva@cdph.ca.gov.

ABSTRACT

Shigella sonnei has caused unusually large outbreaks of shigellosis in California in 2014 - 2015. Preliminary data indicated the involvement of two distinct yet related bacterial populations, one from San Diego and San Joaquin (SD/SJ) and one from the San Francisco (SF) Bay area. Whole genome sequencing of sixty-eight outbreak and archival isolates of *S. sonnei* was performed to investigate the microbiological factors related to these outbreaks. Both SD/SJ and SF populations, as well as almost all of the archival *S. sonnei* isolates belonged to sequence type 152 (ST152). Genome-wide SNP analysis clustered the majority of California (CA) isolates to an earlier described global Lineage III, which has persisted in CA since 1986. Isolates in the SD/SJ population had a novel Shiga-toxin (STX)-encoding lambdoid bacteriophage, most closely related to that found in an *Escherichia coli* O104:H4 strain responsible for a large outbreak. However, the STX genes (*stx1a* and *stx1b*) from this novel phage had sequences most similar to the phages from *S. flexneri* and *S. dysenteriae*. The isolates in the SF population yielded evidence of fluoroquinolone resistance acquired via the accumulation of point mutations in *gyrA* and *parC* genes. Thus, the CA *S. sonnei* lineage continues to evolve by the acquisition of increased virulence and antibiotic resistance, and enhanced monitoring is advocated for its early detection in future outbreaks.

IMPORTANCE

Shigellosis is an acute diarrheal disease causing nearly a half-million infections, 6,000 hospitalizations, and 70 deaths every year in the United States. *Shigella sonnei* caused two unusually large outbreaks of shigellosis in 2014 – 2015 in California. We applied whole genome sequence analyses to understand the pathogenic potential of bacteria involved in these outbreaks. Our results suggest that a local *S. sonnei* clone has persisted in California since 1986. Recently, a derivative of the original clone acquired the ability to produce Shiga-toxin via exchanges of bacteriophages with other bacteria. Shiga toxin production is connected with more severe disease including bloody diarrhea. A second derivative of the original clone recovered from the San Francisco Bay area showed evidence of gradual acquisition of antimicrobial resistance. These evolutionary changes in *S. sonnei* populations must be monitored to explore the future risks of spread of increasingly virulent and resistant clones.

Introduction

Shigellosis is an acute gastrointestinal infection caused by bacteria belonging to the genus *Shigella*. Shigellosis is the third most common enteric bacterial infection in the United States with 500,000 infections, 6,000 hospitalizations, and 70 deaths each year [1]. There are four *Shigella* species which cause shigellosis: *S. dysenteriae*, *S. flexneri*, *S. boydii*, and *S. sonnei* [2]. *S. dysenteriae* is considered to be the most virulent species especially *S. dysenteriae* type 1 serotype due to its ability to produce a potent cytotoxin called Shiga-toxin. *S. flexneri*, *S. boydii,* and *S. sonnei* generally do not produce Shiga-toxin (STX) and, therefore, cause mild forms of shigellosis [2, 3]. Shiga-toxins can also be produced by Shiga toxin-producing *Escherichia coli* (STEC). Two types of STX are known in STEC: STX 1 which differs by a single amino acid from STX in *S. dysenteriae* type 1, and STX 2 which shares only about 55% amino-acid similarity with STX 1 [4]. STX-operons for both type 1 and 2 STX consist of the *stxA* and *stxB* subunit genes, which encode the AB_5 holotoxin [5]. Rarely, Shiga-toxin genes *stx* can be transferred to non-STX-producing *S.flexneri* and *S. sonnei* by means of lambdoid bacteriophage either from STEC or *S. dysenteriae* [3, 6, 7], providing those strains with the ability to cause more severe disease. Infections caused by bacterial species that produce STX often lead to hemorrhagic colitis and may cause serious complications, like hemolytic uremic syndrome (HUS) [8].

S. sonnei has caused two large outbreaks in California (CA) in 2014–2015. The first outbreak with distinct clusters in the San Diego and San Joaquin (SD/SJ) areas was caused by STX-producing isolates [9]. Another CA *S. sonnei* outbreak occurred within the same time frame as the two clusters described above, but was caused by a STX-negative *S. sonnei* strain and was confined primarily to the San Francisco Bay (SF) area [10]. We performed whole genome sequencing of sixty-eight outbreak and archival isolates to gain further insights into the microbiological factors associated with these shigellosis outbreaks. We examined the phylogeny of local *S. sonnei* isolates and also performed comparison of CA isolates with global *S. sonnei* clones. The results provided new insight into likely origin of current CA *S. sonnei* isolates and evolutionary changes to acquire more virulence and antibiotic resistance.

RESULTS

In silico MLST and hq SNP Analysis

Sequences of 7 house-keeping genes were extracted from genomic sequences of all isolates and their allele numbers were identified using the CGE MLST database [11]. The sequence type (ST) was assigned to each isolate based on combination of identified alleles. All outbreak and archival CA *S. sonnei* isolates had identical sequences for 7 MLST loci and were assigned to the same sequence type 152 (ST152), with the exception of one historical isolate C97 identified as ST1502. ST1502 differs from ST152 by a single allele. *S. sonnei* ST 152 was previously found in Germany in 1997 and China in 2009 [12]. In the MLST database, there is also a record of ST152 isolated in the United Kingdom in 2012, though it is not clear if this ST was assigned to *E. coli* or *Shigella* species.

Based on genome-wide SNP analysis, isolates from CA clustered into several clades (Figure 1A). All STX-positive isolates from two recent outbreak-related San Diego (SD) and San Joaquin (SJ) epidemiological groups clustered together and differed from each other by 0-11 SNPs. The member of SD epidemiological group C24 had 0 SNP differences with several SJ isolates (C101, C114-C119, C122, C123, C125, C128, C129), indicating direct relatedness of the isolates from the SD and SJ epidemiological clusters. The only STX-negative isolate C130 from SJ had 48 SNPs differences from the closest member of STX-positive SD/SJ group.

Isolates from the contemporary outbreak in San Francisco (SF) formed a separate cluster (Figure 1A), and had more SNP differences with the SD/SJ cluster (251 SNPs) than with the historical isolate C96 from Contra Costa County, isolated in 1990 (219 SNPs). Isolate C42 from Santa Clara County (2014) was assigned to the SF outbreak based on PFGE (Table S1), but differed by 62 SNPs from the closest SF outbreak member. C42 clustered closer to the SF outbreak than to any other contemporary or historical isolate, therefore was considered to be a part of the SF population.

The majority of historical isolates were genetically distinct from recent isolates except C84 from Imperial County (2008), which differed by 32 SNPs from closest member of SD/SJ STX-positive cluster and by 28 SNPs from STX-negative SJ isolate C130. This is surprisingly small number of SNPs, when compared with the distance between the SD/SJ cluster and other

historical isolates (100-1474 SNPs) or even to the contemporary SF outbreak cluster (251 SNPs).

Global phylogeny of CA S.sonnei

Genomes of CA *S. sonnei* were compared with global *S. sonnei* clones reported by Holt et al. (Table S2) [13] by genome-wide SNP analysis, which is expected to provide better resolution of lineages than MLST. The clustering of global strains in our analysis showed the same general patterns as in the original publication (Fig. S1). Both modern *S. sonnei* populations and the majority of the historical isolates (1986-2008) clustered with Lineage III strains. The European isolate from the UK belonging to the Global III Lineage isolated in 2008 (the most recent from the analyzed collection by Holt et al.), was the closest global strain to both the SD/SJ and the SF *S. sonnei* populations. Historical isolates C96, C98, and C84 (from 1990, 2007, and 2008, correspondingly), also clustered together with the Global III Lineage. Another clade within Lineage III was formed by an isolate from Mexico (1998) and a group of historical CA *S. sonnei* isolates obtained between 1986 and 2002. Three CA isolates from 1980-1987 clustered together with Lineage II. A single historical isolate C97, which also had a unique MLST pattern ST1502, was grouped with Lineage I strains. No Lineage IV isolates were found in CA (Figure 1A, Fig. S1)

COG and pfam clustering of CA *S.sonnei* genomes

Hierarchical clustering based on COG (the Clusters of Orthologous Groups of proteins) and pfam (protein domains) profiles was used as an alternative method for examining similarity and relatedness of strains. This protein family profile clustering was carried out with representative genomes of CA *S. sonnei* isolates (C7, 113, C123 –SD/SJ population; C39- SF population; C84, C88, C97- historical isolates) and other publicly available genomes of *Shigella* species and different *E. coli* serotypes. Two different clustering methods were used here – one based on presence/absence and one based on abundance (in both cases of COG and pfam groups). The COG-based approach showed all CA *S. sonnei* isolates clustering together with *E. coli* and separate from other *S. sonnei* isolates, with the exception of *S. sonnei* strain 1DT-1, which also clustered with *E.coli* (Fig. S2A). CA *S. sonnei* clustered closer with non-O157 serotypes of *E. coli*, particularly with O104:H4 STEC strain (which caused a large outbreak in Germany and other European countries in 2011) than with *E. coli* O157:H7. Similar results were

seen in the hierarchical clustering based on Pfam profiles (Fig. S2B). Except in the pfam clustering two CA *S. sonnei* isolates (one of the historical isolates and one San Francisco isolate) clustered with other *S. sonnei* from the database, while all other CA isolates clustered with *E. coli,* as previously.

To determine whether specific parts of genome were shared between CA *S.sonnei* and *E.coli* and contributed to CA *S. sonnei* clustering with *E.coli*, we searched for the genes in the genome of CA *S.sonnei* which had homologs in *E. coli* O104:H4 strains, but not in other *S. sonnei* from the database. The genes which were found to be shared between *E. coli* O104:H4 and STX-positive CA *S.sonnei* isolates (representative, C7 and C123) belonged to either the STX-bacteriophage characterized here (for both C7 and C123), or to a $bla_{\text{TEM-1}}$-encoding plasmid (in case of C123) (Table S3). The CA *S.sonnei* isolates lacking STX-bacteriophage didn't have any genes which were shared with *E.coli* O104:H4, but not with other *S.sonnei*.

We also performed a recombination test on alignment of genome-wide hqSNPs using the Recombination Detection Program v. 4 (RDP4) [14]. None of the used recombination detection algorithms included in the software package detected recombination events between CA *S.sonnei* and *E.coli* O104:H4. Thus, no other gene exchange events, except for STX-bacteriophage transfer, were detected between *E.coli* O104:H4 and CA *S.sonnei.*

A phylogenetic tree of genomes was constructed from a concatenation of a set of 37 core phylogenetic marker genes using Phylosift (Fig. S3A) [15]. In this tree, all *S. sonnei* including CA isolates grouped together, and separate from *E. coli* with one exception. The exception was *S. sonnei* strain 1DT-1, which grouped in a clade with *E. coli*. Both nucleotide (Fig. S3A) and amino acid (AA) sequence (data not shown) trees generated with Phylosift showed the same result. Whole genome clustering based on genome-wide SNPs in both coding and non-coding areas of the genome (Fig. S3B) supported the results of Phylosift: all *S. sonnei* isolates, including ones from CA, clustered together, except for *S. sonnei* 1DT-1.

The difference between protein family profile clustering and concatenated marker gene phylogeny is striking. In the protein family profile clustering CA *S.sonnei* isolates group with *E.coli* O104:H4 than with other *S.sonnei* but in the concatenated marker gene phylogeny CA *S.sonnei* and *E.coli* do not

group together. At first, to explain this contrast we considered two plausible biological explanations:1) convergent gene gain or loss could have led to high similarity in terms of protein family profiles between CA *S. sonnei* and *E. coli* even though they are not closely related, or 2) considering the hypothesis of origination of all *Shigella* species from *E.coli* [16], CA *S.sonnei* could be shearing higher similarity of protein family profiles with *E.coli* because it represents a more archaic lineage of *S.sonnei* species, which haven't diverged from *E.coli* as much as the other *Shigella* strains found in the NCBI and IMG databases did.

However, neither of these explained all the results from these analyses. Closer examination of some of the results led us to consider a third possibility – the observed contrast was an artifact of analysis. Upon further review, we found that CA isolates and *S. sonnei* strain 1DT-1 similar to *E. coli* O104:H4 had low gene count for several Transposases (COG2963, COG2801, COG3547, COG3335, COG4584, COG3385) and a DNA replication protein DnaC (COG1484), while other *Shigella* spp. isolates in the database had those genes in abundance (Fig. S4). This indicated that a few transposable elements skewed the clustering of CA *S. sonnei* towards *E .coli*. When only the presence/absence of the COG and pfam elements was taken into the account, COG- and pfam-based clustering appeared to be congruent with nucleotide phylogeny: CA *S. sonnei* have clustered with other *S. sonnei* strains in the database (Figure 2; Fig. S5). Therefore, we would like to emphasize that genome clustering based on gene presence/absence as well as abundance is prone to the issues described above and shouldn't be used to infer phylogenies.

STX-encoding bacteriophage from CA *S.sonnei*

All isolates from SD/SJ outbreak possessed STX-1 subunits A and B genes (*stx1A, stx*1B) (Figure 3). None of the historical isolates or recent isolates related to the San Francisco outbreak had *stx* genes. The STX in the SD/SJ population was encoded by a novel lambdoid bacteriophage. This 62.8kb bacteriophage had identical sequence and was integrated into a chromosomal *wrbA* site in all STX-positive isolates. A modern STX-negative *S. sonnei* isolate C130 from SJ had an intact *wrbA* site; therefore, we conclude that it represents an ancestral STX-negative lineage for the SD/SJ population.

In order to better understand the history of the novel STX-phage in the SD/SJ *S.sonnei* population, we carried out a series of comparative analyses.

First, the Phage Search Tool (PHAST) revealed that the best match for the novel CA STX1-phage was the STX2-encoding P13374 phage found in O104:H4 STEC, an outbreak strain that caused a large outbreak in Germany and other European countries in 2011 [17]. This finding was supported by Blast searcher which revealed that the bacteriophage from the CA isolates was very similar to two phages found in strains from the European *E. coli* O104:H4 outbreak [P13374 phage (NC_018846)] with 97% identity for 88% of the query search coverage and the phage in *E. coli* strain 2011C-3493 (CP003289.1) with 97% identity for 93% of the query coverage. CA STX1-phage also showed very high similarity with stx2-encoding phages from *E. coli* O104:H4 strains 2009EL-2050 (CP003297.1) and 2009EL-2071 (CP003301.1) from Georgia, 2009 [17] with 98% shared identity at 92% query length. A recently characterized STX1-encoding *S. sonnei* phage 75/02_Stx from Hungary (KF766125.2) was also shown to share large co-linear regions with STX2-prophages from *E.coli* O104:H4 and O157:H7 [5]. However, the Hungarian phage only partially resembled CA STX1-phage with 99% of shared identity at 62% query coverage.

In order to better understand the phylogenetic relatedness of CA STX1-phage to other phages, we generated phylogenetic trees of integrase genes. This revealed that the CA STX1-phage grouped evolutionarily with other STX-encoding *E.coli* phages, particularly with phages from *E. coli* O104:H4, but not with ones from *Shigella* species (Fig. S6). Figure S6 shows the phylogeny based on Integrase amino-acid sequences; similar results were seen with nucleotide based trees (not shown). This integrase phylogenetic analysis confirmed the relatedness of the CA STX-phage to the phage found in pandemic *E. coli* O104:H4 from Germany, 2011. It also showed that bacteriophages with integrase genes related to phage from European outbreak *E. coli* O104:H4 were also found earlier in Europe (Norway, 2007 and Georgia, 2009).

We also used progressive Mauve to make and analyze whole phage alignments. This revealed that the STX1-phage in CA isolates more closely resembled STX-coding phages in *E. coli* than phages in *S. sonnei* or *S. flexneri* (Figure 4A, 4B). The stx2-encoding phage from *E. coli* O104:H4 strain 2011C-3493 was the closest to CA STX-phage according to Neighbor-Joining tree built based on the estimate of the shared gene content from progressiveMauve alignment. Notably, *E. coli* strain 2011C-3493 was

isolated in the US from a patient with the history of travel to Germany and shown to be a part of German outbreak [18].

Examination of the distribution of CA STX1-bacteriophage-related genes in different *E. coli* and *Shigella* serovars in the IMG database showed that the CA STX-phage genes were also present in *E. coli* O104:H4 isolates, the majority of which were directly related to the German outbreak. Strain ON2010 of *E. coli* O104:H4 from Canada (2010), which was not linked to the German outbreak [19], showed no presence of CA STX-phage genes (Fig. S7).

In several CA isolates we identified the bacteriophages known to be associated with STX, but which were missing actual *stx* genes. According to PHAST output, the isolate C7 possessed an intact STX2-converting phage 1717 (NC_011357), isolate C113 had an incomplete STX2-converting phage 86 (NC_008464), and isolates C16 harbored a questionable STX2-converting phage I (NC_003525), while none of them contained the *stx* gene sequence (Fig. S8). We propose that the *stx* genes have been gained and lost during the evolution of *S.sonnei* in California.

STX-holotoxin and other virulence determinants CA *S.sonnei*

Even though the phage from German outbreak STEC strain was identified as the most closely related to the CA STX1-phage, it carries the *stx2* gene, while CA phage encodes the STX1 toxin. The STX-operon in CA isolates (holotoxin genes *stx1a* and *stx1b*) was identical to STX operons in *S. flexneri* phage POCJ13 (KJ603229.1) and several *S. dysenteriae* strains, including *S. dysenteriae* type 1 (M19437.1), but differed from STX operon in the majority of *E. coli* strains. The CA *S. sonnei* STX-operon had 1 SNP difference with other STX1-encoding *S. sonnei* bacteriophages (Accession ## KF766125.2 and AJ132761.1). Similarity of different regions of CA STX1-phage to various phages can be explained by the mosaic structure of STX lambdoid phages due to frequent recombinations and modular exchange with other lambdoid phages [20]. Moreover, there is evidence that it could be quite common for lambdoid phages of *S. sonnei* to have an STX genes from *S. dysenteriae* and the rest of the phage genes resembling STEC or EHEC *E. coli* strains [3, 5].

Besides STX, the following virulence genes were detected in *S. sonnei* from CA: Shigella IgA-like protease homologue (*sigA*), glutamate decarboxylase (*gad*), plasmid-encoded enterotoxin (*senB*), P-related fimbriae regulatory gene (*prfB*), long polar fimbriae (*lpfA*), endonuclease colicin E2 (*celb*), invasion protein *S. flexneri* (*ipaD*), VirF transcriptional activator (*virF*), and increased serum survival gene (*iss*). The distribution of these virulence genes in different *S.sonnei* populations from CA is presented in Figure 3. The increased serum survival virulence gene *iss* was mainly found in more recent of the SD/SJ population of isolates, and in a single historical isolate from 1980, however the *iss* alleles in modern and historical isolates were different.

Antimicrobial resistance of CA *S.sonnei*

Acquired antibiotic resistance (ABR) markers to the following classes of antimicrobials were identified: beta-lactams (genes $bla_{\text{TEM-1}}$; $bla_{\text{OXA-2}}$), aminoglycosides (*aph(3'')-Ib; aph(6)-Id; aac(3)-IId; aadA1; aadA2*), macrolides (*mphA*), sulphonamides (*sul1; sul2*), phenicols (*catA1*), trimethoprim (*dfrA1; dfrA8; dfrA12*), and tetracycline (*tetA; tetB*) (Figure 5). The majority of the isolates possessed genes predicted to provide resistance to aminoglycosides, sulphonamides, trimethoprim, and tetracycline. In addition, a phenotypic resistance to penicillins mediated by TEM-1 β-lactamase gene occurred in 16 isolates. The correlation between the genotype and susceptibility phenotype is presented in the Table S4. In 100% of cases, the presence of antibiotic resistance determinants correlated with the expected non-susceptible phenotype. The 92.9% of the isolates lacking the resistance determinant were susceptible to corresponding antimicrobials. A variant of the gene of aminoglycoside-modifying enzyme *aac(3)-IId* was found in a single *S. sonnei* isolate C98 from CA; the *aac(3)-IId*-like gene in isolate C98 shared 99.8% identity with previously described variant of this gene.Acetyltransferase *aac(3)-IId* was shown previously to confer resistance to gentamicin and tobramycin in *E. coli* isolates [21], while a *S. sonnei* isolate harboring a variant of this gene was intermediate to tobramycin and resistant to gentamicin.

In 7% of cases, the isolates without known antibiotic-resistance genes appeared to be non-susceptible (Table S4), suggesting the presence of resistance mechanisms, like loss of porins and increased efflux [22], which are not included in the ResFinder database. For example, two isolates

showed intermediate susceptibility to streptomycin, and 15 isolates were non-susceptible to ampicillin/sulbactam in the absence of the genes which would explain the corresponding resistance phenotypes in those isolates. Among the tetracycline-resistant isolates, 3.3% (n=2) of isolates did not possess any known tetracycline resistance genes, while 86.9% (n=53) had a truncated version of *tet(A)* gene (1172 bp vs. 1200 bp for the full-length reference sequence from the ResFinder database). Nonetheless, the isolates without tetracycline-resistance genes or with incomplete *tet(A)* gene demonstrated a high level resistance to tetracycline (MIC >8μg/mL). This suggests the presence of additional mechanisms of resistance, undetectable by the ResFinder database.

Acquired ABR genes were frequently associated with various mobile elements. Penicillin resistance encoded by $bla_{\text{TEM-1}}$ gene was associated with Tn*3* transposon located on conjugative plasmids of IncB/O/K/Z incompatibility group found in both recent SF and historical isolates (Fig. S9A) or on the putative IncI1 plasmid in several modern SJ isolates (Fig. S9B). Partial plasmid sequences were identified in historical isolates as IncB/O/K/Z and contained plasmid backbone genes encoding following products: transcriptional activator RfaH, plasmid stabilization system protein, ribbon-helix-helix protein CopG, replication initiation protein RepA, and plasmid conjugative transfer protein. This partial IncB/O/K/Z plasmid sequences in historical isolates matched analogous area in IncB/O/K/Z plasmids of recent isolates, suggesting their relatedness to each other. A determinant Tn*3*::$bla_{\text{TEM-1}}$ was integrated into different loci of the IncB/O/K/Z plasmids in historical isolates in comparison with integration locus in IncB/O/K/Z plasmids of modern San Francisco cluster isolates, which suggests that dissemination of the $bla_{\text{TEM-1}}$ gene is associated with the transfer via Tn*3* (Fig. S9A).

Aminoglycoside and trimethoprim resistance genes were often linked with Tn*7* transposon (Fig. S9C). A genetic region, containing an association of Tn*7* with *aadA*1 and *dfrA*1 ABR genes, as well as TDP-fucosamine acetyltransferase, and a partial sequence of Tyrosine recombinase XerC/D, was identical in recent SJ isolates and two historical samples C99 and C92, isolated in Shasta County, 2000 and Imperial County, 2002, correspondingly. However, the area upstream of Tn*7* in novel isolates showed evidence of later integration of additional Tn*7* element, leading to the loss or

383 modification of the area containing tetracycline resistance genes found in
384 historical isolates.
385 One of the historical isolates C80 (Orange County, 1987) was resistant to
386 azithromycin and possessed macrolide resistance gene *mphA*, found in
387 association with Tn*7* transposon on a contig which also encoded
388 trimethoprim, sulfamethoxazole, aminoglycoside resistance, multidrug
389 transporter EmrE, and a Mercuric resistance operon (Figure 6). Resistance to
390 azithromycin is emerging in the United States [23], and the presence of
391 azithromycin resistance gene in one of the isolates dating back to 1987 is
392 remarkable.

393 **Fluoroquinolone resistance**
394 All modern isolates from the SD/SJ population and historical *S. sonnei*
395 isolates were susceptible to fluoroquinolones (FQ), however few historical
396 isolates possessed several types of point mutations leading to AA
397 substitutions in the genes encoding FQ targets- the A subunit of DNA-gyrase
398 (GyrA) and the C subunit of topoisomerase IV (ParC) (Figure 5, Table S4)
399 The quinolone resistance-determining region (QRDR) [24] of the GyrA in two
400 CA historical isolates had a single amino acid substitution S83L. *S.sonnei*
401 isolates with single mutation S83L have been previously reported to exhibit
402 high minimum inhibitory concentration (MIC) of quinolones (nalidixic acid),
403 but low MIC of fluoroquinolones (MIC range of ciprofloxacin 0.125-
404 0.25μg/ml) [25]. In agreement with previous data, historical CA *S.sonnei*
405 isolates with single mutation S83L remained phenotypically susceptible to
406 ciprofloxacin with wild-type MIC values of ≤0.5 μg/mL. Two FQ-susceptible
407 historical isolates had single point mutation V533A in the ParC, and one FQ-
408 susceptible isolate had combination of ParC-V533A and GyrA-D483E
409 mutations; none of the mutations were located in QRDR regions of
410 corresponding proteins and were not associated with a resistant phenotype.

411 In contrast, all SF population isolates implicated in the outbreak
412 expressed high level FQ-resistance. All isolates from the contemporary San
413 Francisco outbreak had a combination of point mutations S83L and D87G in
414 QRDR GyrA and substitution S80I in QRDR ParC, which caused a significant
415 increase in ciprofloxacin MIC to ≥4 μg/mL (Table S4). The modern non-
416 outbreak isolate C42, belonging to the SF population, was phenotypically
417 susceptible to FQ and possessed a single QRDR GyrA mutation S83L, which
418 represents a prerequisite mutation allowing for one-step development of

high-level FQ resistance if combined with another QRDR mutation [26], thus suggesting a concurrent accumulation of FQ resistance mutations in the SF population of *S. sonnei* (Figure 5).

DISCUSSION

From our point of view, the key results of our study were: 1) the first large scale whole genome sequencing and analysis of *S. sonnei* strains from North America, 2) the delineation of two distinct yet related populations of *S. sonnei* that caused large outbreaks of shigellosis in California [SD and SJ population and SF population], 3) linkage of SD/SJ and SF populations to an older lineage of *S. sonnei* documented in California as early as 1986. This historical CA lineage was characterized as sequence type 152 and belonged to a global Lineage III with origin in Europe, 4) evidence of stx1-encoding bacteriophage in the SD/SJ population. The phage was related to the phage from pandemic *E. coli* O104:H4 with the STX-operon identical to those in *S. flexneri* and *S. dysenteriae*, 5) evidence of fluoroquinolone resistance in SF population via accumulation of mutations in the genes of antibiotic targets-GyrA and ParC.

According to the whole-genome hqSNP phylogeny, the isolates from outbreaks of 2014 - 2015 were divided into two distinct SD/SJ and SF populations. The isolates from SD and SJ epidemiological groups were assigned to two separate clusters according to PFGE profiling [9], however based on WGS we showed that SD and SJ clusters were a part of the same SD/SJ outbreak population. Many recent publications have highlighted superior discriminatory power and predictive value of WGS over PFGE for genotyping of bacterial strains [27].

Comparison to global *S. sonnei* clones placed both modern SD/SJ and SF populations as well as majority of historical CA *S. sonnei* isolates dating back as far as 1986 into the Lineage III defined by Holt et al. [13] (Figure 1A, Fig. S1). All STX-positive SD/SJ isolates clustered with Global III clade defined by Holt et al. within Lineage III, which was shown to be particularly successful at global expansion [13]. Remarkably, the isolates from SD/SJ population, as well as from SF population, were closest to the isolate from the UK, 2008, while they did not cluster with geographically proximate isolate from Mexico, which was only related to few historical CA isolates. Accepting the hypothesis that all original *S. sonnei* lineages diversified from Europe and then were introduced to other countries where they underwent

localized clonal expansions [13], our findings suggest that while strain exchange with Mexico had happened in the past, the lineage ancestral to the current SD/SJ and SF *S. sonnei* populations was introduced at some point from Europe, and diversified within CA. The data suggests that the acquisition of stx1 gene happened after the original Lineage III *S. sonnei* clone had already spread in CA. The contemporary C130 and historical C84 STX-negative *S. sonnei* isolates, which both clustered with STX-positive SD/SJ outbreak isolates, most likely represent an ancestral STX-negative lineage for the SD/SJ population, prior to STX-bacteriophage acquisition. It appears that global *S. sonnei* clones belonging to Lineages I, II, and III were introduced to CA independently during the last three decades.

All recent *S. sonnei* outbreak isolates as well as all but one historical isolates belonged to a sequence type ST152. This long-term persistence of a single clone in CA over three decades is remarkable, and could indicate that ST152 is a very successful *S. sonnei* clone. There were previous examples when the same sequence types persisted in the same geographical area for a long time, e.g. ST245 in China was found in 1983 as well as 26 years later in 2009 [12]. This phenomenon, however, is most likely explained by a low resolution of MLST for subtyping of clonal *S. sonnei* population, which has been noted before [28].

Isolates from the SD/SJ *S. sonnei* population, which caused outbreak of severe diarrheal disease [9], carried stx1 genes encoded by a novel lambdoid bacteriophage in all of the isolates. The STX-bacteriophage from *E. coli* O104:H4 strain implicated in an European outbreak of 2011 was found to be the closest known relative to STX-phage of SD/SJ *S. sonnei* isolates based on overall similarity of the whole-phage sequences. This outbreak was one of the world's largest outbreaks of food-borne disease in humans with 855 HUS and 53 fatalities [29, 30]. Considering the particular proximity of CA STX-phage to the phage from *E. coli* O104:H4 strain 2011C-3493 introduced from Germany to the US in 2011, we speculate that introduction of STX-phage into the SD/SJ *S. sonnei* population happened after the outbreak *E. coli* O104:H4 was brought to the US, followed by a local diversification of STX-phage, which involved recombination of STX-operon to yield the final phage with general architecture of *E. coli* phage and STX-operon from *S. flexneri* or *S. dysenteriae*. However, it is also possible that STX-phage from pandemic *E. coli* O104:H4 underwent the modifications via recombination with other European phages, prior to its introduction to CA *S.*

sonnei clone. The transfer of STX-bacteriophage is the only gene exchange event we detected between *E. coli* O104:H4 and *S. sonnei* genomes. It would be pertinent to emphasize that the German outbreak strains of *E. coli* also contained distinct STX operon with *stx2* gene, several virulence and antibiotic-resistance elements [31].

It has been demonstrated previously that phage phylogeny should be inferred from a combination of protein repertoires and phage architecture rather than from a single gene (e.g. integrase) sequence [32, 33]. The mosaic structure of the lambdoid phages poses a limitation for such single locus-phylogeny approach. Thus, even though the integrase phylogeny showed that CA isolate has an integrase gene more closely related to older Georgian than with recent German *E. coli* O104:H4 isolate, this does not contradict the relatedness of CA phage to the outbreak lamboid-phage shown based on the cumulative data derived from the whole phage sequence. This also could be explained by a long-term persistence of phages belonging to a given Integrase family in Europe.

STX-1 encoding bacteriophage was not found in the isolate C130 even though it was closely related to SD/SJ *S. sonnei* population. It is possible that the bacterial population as a whole acquired STX1-encoding bacteriophage more recently, and concurrently. Such a scenario would also suggest that the increase in the virulence took place either by the spread of STX-positive clones or by the horizontal transfer of stx1- genes by bacteriophages. The potential of bacteriophages to transfer STX genes from one *S. sonnei* strain to another has been described previously [3]. Not all the *S. sonnei* strains have a suitable genetic background to be able to express STX efficiently [34]. The modern STX-positive SD/SJ population of *S.sonnei* was shown to be able to express STX and thus is proven to possess the genetic background which is adapted for the increased virulence. The adaptation of SD/SJ *S. sonnei* genomic background to retain and express STX stably by itself is a prerequisite for future emergence of more virulent strains in CA. Another example of SD/SJ population increasing its virulence is an acquisition of increased serum survival virulence gene iss, which occurred seemingly recently in evolution of SD/SJ S.sonnei clone. To our

knowledge, this is the first report of increased serum survival determinant (iss) being found in *Shigella* species.

Even though the treatment of infection caused by *S. sonnei* with antibiotics is not a standard procedure, it is indicated for the treatment of severe cases in order to reduce duration of symptoms [35]. Antibiotic resistance of *S. sonnei* is however on the rise [10, 36, 37]. This highlights the importance of ABR monitoring in *S. sonnei* populations. We detected resistance markers to multiple classes of antimicrobials in CA *S. sonnei*. 91.2% of modern isolates and 81.8% of historical isolates were multi-drug resistant (MDR) according to the definition of MDR microorganisms as demonstrating "non-susceptibility to at least one agent in three or more antimicrobial categories" [38]. Acquired ABR genes in CA *S. sonnei* were frequently associated with various mobile elements and conjugative plasmids, likely contributing to their dissemination. For example, MDR transposon Tn7 has been shown to be a crucial part of ABR evolution in the local populations of *S. sonnei* Lineage III [13]. Once acquired, ABR genes tend to stay in the bacterial population, e.g. traditional antibiotics like cotrimoxazole and tetracycline are no longer in use for patient treatment [12], but the ABR genes still persist in the modern SD/SJ and SF populations of the CA *S. sonnei*. Particularly worrisome was the fluoroquinolone resistance detected in all of the isolates belonging to a modern SF *S.sonnei* population, implicated in the outbreak. FQ-resistance in all SF outbreak isolates was mediated by a combination of double mutation in QRDR GyrA of DNA gyrase and one mutation in QRDR ParC of topoisomerase IV. All historical isolates as well all other modern isolates were susceptible to FQ, however few of those phenotypically susceptible CA *S.sonnei* isolates possessed a prerequisite single AA substitution in QRDR of the GyrA, including two isolate belonging to CA Lineage III: one historical isolate from 2007 and one modern non-outbreak isolate belonging to SF population. Single mutations in QRDR GyrA were shown previously to confer a low level of FQ resistance and to serve a prerequisite for further resistance escalation via stepwise mutations acquisition [26]. Therefore, the data suggests that evolution of SF population from ancestral *S. sonnei* lineage is occurring via increase in fluoroquinolone resistance mediated by the accumulation of FQ-resistance mutations.

There are certain limitations of this study: 1) temporal and spatial representation was sporadic in *S. sonnei* strains in our culture collection, 2) the short read sequencing by synthesis used to generate data might limit complete assembly of *S. sonnei* genomes and enumeration of genome-wide SNPs, 3) no plasmid transformation followed by re-sequencing could be performed to confirm postulated genetic exchanges among *S. sonnei* populations. Although the direct experimental evidence is lacking, Shiga-toxin production and other virulence elements discovered in SD/SJ population appeared to be among the contributors that lead to serious manifestations of gastrointestinal disease in CA shigellosis outbreak including bloody diarrhea in 71% of patients [9]. Fortunately, there were no fatalities and none of the affected patients developed more serious hemolytic uremic syndrome (HUS). One potential clue to the absence of severe manifestations could be that Shiga-toxin positive *S. sonnei* strains contained Stx1. The toxins encoded by Stx1 and Stx2 are known to elicit variable pathology among affected individuals. A number of investigators have demonstrated that *E. coli* O157 Shiga-toxin producing strains carrying Stx2 caused more severe disease including HUS than Stx1 positive strains [39-41].

Conclusion: Two distinct populations of *S. sonnei* (SD/SJ and SF) have been delineated in the recent shigellosis outbreaks in California. These populations evolved from a common lineage of *S. sonnei* likely present in California as early as 1986. The suggested evolutionary pathways were: 1) increased virulence via acquisition of a phage from the *E. coli* O104:H4 German outbreak strain and STX-operon from *S. flexneri* or *S. dysenteriae*, and 2) emergence of fluoroquinolone resistance via the accumulations of point mutations in *gyrA* and *parC* genes. The modern CA *S. sonnei* populations were related to the global Lineage III, which originated in Europe and was known for its successful expansion around the world. Thus, the CA *S. sonnei* lineage continues to evolve by the acquisition of increased virulence and antibiotic resistance, and enhanced surveillance is advocated for its early detection in future shigellosis outbreaks.

MATERIALS AND METHODS

Isolates

Sixty-eight *S.sonnei* human isolates from CA (57 outbreak-related from 2014-2015 and 11 archival isolates from 1980-2008), were identified and serotyped by standard methods [42]. PCR-detection of stx_1 and stx_2 genes, and Vero cell neutralization assay for confirmation of STX-production were performed as previously described [43]. A list of isolates is presented in the Table S1. A map with geographical distribution of the location of origin of the isolates is in Figure 1B.

WGS and data analysis

DNA was extracted with a Wizard Genomic DNA Kits (Promega, Madison, WI). Sequencing libraries were constructed using the Nextera XT (Illumina Inc., San Diego, CA) library preparation kits. Sequencing was performed using 2 x 300 bp sequencing chemistry on an Illumina MiSeq Sequencer as per manufacturer's instructions.

For genome-wide SNP identification, paired-end reads were mapped to the reference genome of *S. sonnei* Ss046 (NC_007384.1) with masking of the mobile and phage elements and phylogenetic tree was built using CLCbio Genomic Workbench 8.0.2 (Qiagen, Aarhus, Denmark). A phylogenetic tree was generated using maximum likelihood phylogeny (under the Jukes-Cantor nucleotide substitution model; with bootstrapping) based on high-quality single nucleotide polymorphisms (hqSNPs). SNPs were called in coding and non-coding genome areas using SAMtools mpileup (v.1.2; [44]) and converted into VCF matrix using bcftools (v0.1.19; http://samtools.github.io/bcftools/). Variants were parsed using vcftools (v.0.1.12b; [45]) to include only high-quality SNPs (hqSNPs) with coverage ≥5x, minimum quality > 200, minimum genotype quality (GQ) 10 (--minDP 5; --minQ 200; --minGQ 10; --remove-indels), with InDels and the heterozygote calls excluded.

Phylosift [15] was used to (1) identify 37 "universal" genes in a set of genomes (including those generated here) (2) generate alignments for each gene family and then concatenate the alignments. A phylogenetic tree was inferred from the concatenated alignments using RAxML 7.2.6. bipartition trees were generated using 1000 bootstraps. COG- and Pfam-based identification and clustering was done using the DOE Joint Genome Institute (JGI) Integrated Microbial Genomes (IMG) system

(https://img.jgi.doe.gov/cgi-bin/mer). The presence or absence matrices were fed to RAxML 7.3.0 [46] to produce a best tree from 50 boostraps using the gamma model rate of heterogeneity for binary input. COG abundance profile and gene homologs search was performed using JGI IMG tools.

De novo assembly for each genome was done on CLCbio GW8.0.2. The assembled genomes of *S. sonnei* isolates had 30-159 x sequencing coverage. Genomes were annotated with prokka v1.1, the JGI IMG database, the Center for Genomic Epidemiology (CGE) (ResFinder, VirulenceFinder, PlasmidFinder) [47], and the Phage Search Tool (PHAST) [48] online resources. *In silico* multi-locus sequence typing (MLST) was performed using the CGE online tool [11] against the MLST database *E.coli*#1 [49]. To estimate similarity between the bacteriophages, sequence alignment and a neighbor-joining tree (based on the estimate of the shared gene content) were generated using the progressiveMauve program [50]. Additionally, sequences of integrases (both the DNA for the genes and the encoded amino acids for the proteins), derived from the sequences of STX-phage belonging to *Shigella* spp. and *E.coli* strains, were aligned using CLCbio GW8.0.2. general aligner and a neighbor-joining tree was generated as above.

Recombination test was performed on genome-wide hqSNP alignment using Recombination Detection Program v. 4.67 (RDP4) [14]. The following algorithms included into the RDP4 package were applied to search for the recombination events: RDP, BOOTSCAN, GENECONV, MAXCHI, CHIMAERA, SISCAN, 3SEQ, PHYLPRO, and VisRD. A window size of 100 and a step size of 30 were used.

Antimicrobial susceptibility and resistance testing

Antimicrobial susceptibility testing of *S. sonnei* isolates was performed using Microscan Dried Gram Negative panels Neg MIC 38 (Beckman Coulter, Brea, CA, USA); the minimum inhibitory concentration (MIC) results were read and interpreted according to the manufacturer's instructions. Streptomycin (10 μg) and azithromycin (15 μg) BBL Sensi-Discs (Becton Dickinson, Franklin Lakes, NJ, USA) were used to determine susceptibility to the corresponding antimicrobials. Standard quality control strains were tested in parallel as required in respective product inserts.

SUPPLEMENTAL MATERIALS

References

1. Mead PS, Slutsker L, Dietz V, McCaig LF, Bresee JS, Shapiro C, Griffin PM, Tauxe RV: **Food-related illness and death in the United States**. *Emerging infectious diseases* 1999, **5**(5):607-625.
2. Niyogi SK: **Shigellosis**. *Journal of microbiology* 2005, **43**(2):133-143.
3. Strauch E, Lurz R, Beutin L: **Characterization of a Shiga toxin-encoding temperate bacteriophage of Shigella sonnei**. *Infection and immunity* 2001, **69**(12):7588-7595.
4. Hamabata T, Tanaka T, Ozawa A, Shima T, Sato T, Takeda Y: **Genetic variation in the flanking regions of Shiga toxin 2 gene in Shiga toxin-producing Escherichia coli O157:H7 isolated in Japan**. *FEMS microbiology letters* 2002, **215**(2):229-236.
5. Toth I, Svab D, Balint B, Brown-Jaque M, Maroti G: **Comparative analysis of the Shiga toxin converting bacteriophage first detected in Shigella sonnei**. *Infection, genetics and evolution : journal of molecular epidemiology and evolutionary genetics in infectious diseases* 2016, **37**:150-157.
6. Nyholm O, Lienemann T, Halkilahti J, Mero S, Rimhanen-Finne R, Lehtinen V, Salmenlinna S, Siitonen A: **Characterization of Shigella sonnei Isolate Carrying Shiga Toxin 2-Producing Gene**. *Emerging infectious diseases* 2015, **21**(5):891-892.
7. Gray MD, Lampel KA, Strockbine NA, Fernandez RE, Melton-Celsa AR, Maurelli AT: **Clinical isolates of Shiga toxin 1a-producing Shigella flexneri with an epidemiological link to recent travel to Hispaniola**. *Emerging infectious diseases* 2014, **20**(10):1669-1677.
8. Mayer CL, Leibowitz CS, Kurosawa S, Stearns-Kurosawa DJ: **Shiga toxins and the pathophysiology of hemolytic uremic syndrome in humans and animals**. *Toxins* 2012, **4**(11):1261-1287.
9. Lamba K, Nelson JA, Kimura AC, Poe A, Collins J, Kao AS, Cruz L, Inami G, Vaishampayan J, Garza A *et al*: **Shiga Toxin 1-Producing Shigella sonnei Infections, California, United States, 2014-2015**. *Emerging infectious diseases* 2016, **22**(4):679-686.
10. Bowen A, Hurd J, Hoover C, Khachadourian Y, Traphagen E, Harvey E, Libby T, Ehlers S, Ongpin M, Norton JC *et al*: **Importation and domestic transmission of Shigella sonnei resistant to ciprofloxacin - United States, May 2014-February 2015**. *MMWR Morbidity and mortality weekly report* 2015, **64**(12):318-320.
11. Larsen MV, Cosentino S, Rasmussen S, Friis C, Hasman H, Marvig RL, Jelsbak L, Sicheritz-Ponten T, Ussery DW, Aarestrup FM *et al*: **Multilocus sequence typing of total-genome-sequenced bacteria**. *Journal of clinical microbiology* 2012, **50**(4):1355-1361.
12. Cao Y, Wei D, Kamara IL, Chen W: **Multi-Locus Sequence Typing (MLST) and Repetitive Extragenic Palindromic Polymerase Chain Reaction (REP-PCR), characterization of shigella spp. over two decades in Tianjin China**. *International journal of molecular epidemiology and genetics* 2012, **3**(4):321-332.
13. Holt KE, Baker S, Weill FX, Holmes EC, Kitchen A, Yu J, Sangal V, Brown DJ, Coia JE, Kim DW *et al*: **Shigella sonnei genome sequencing and phylogenetic analysis indicate recent global dissemination from Europe**. *Nature genetics* 2012, **44**(9):1056-1059.
14. Martin DP, Lemey P, Lott M, Moulton V, Posada D, Lefeuvre P: **RDP3: a flexible and fast computer program for analyzing recombination**. *Bioinformatics* 2010, **26**(19):2462-2463.
15. Darling AE, Jospin G, Lowe E, Matsen FAt, Bik HM, Eisen JA: **PhyloSift: phylogenetic analysis of genomes and metagenomes**. *PeerJ* 2014, **2**:e243.
16. Pupo GM, Lan R, Reeves PR: **Multiple independent origins of Shigella clones of Escherichia coli and convergent evolution of many of their characteristics**. *Proceedings of the National Academy of Sciences of the United States of America* 2000, **97**(19):10567-10572.

17. Beutin L, Hammerl JA, Strauch E, Reetz J, Dieckmann R, Kelner-Burgos Y, Martin A, Miko A, Strockbine NA, Lindstedt BA *et al*: **Spread of a distinct Stx2-encoding phage prototype among Escherichia coli O104:H4 strains from outbreaks in Germany, Norway, and Georgia**. *Journal of virology* 2012, **86**(19):10444-10455.
18. Ahmed SA, Awosika J, Baldwin C, Bishop-Lilly KA, Biswas B, Broomall S, Chain PS, Chertkov O, Chokoshvili O, Coyne S *et al*: **Genomic comparison of Escherichia coli O104:H4 isolates from 2009 and 2011 reveals plasmid, and prophage heterogeneity, including shiga toxin encoding phage stx2**. *PloS one* 2012, **7**(11):e48228.
19. Hao W, Allen VG, Jamieson FB, Low DE, Alexander DC: **Phylogenetic incongruence in E. coli O104: understanding the evolutionary relationships of emerging pathogens in the face of homologous recombination**. *PloS one* 2012, **7**(4):e33971.
20. Johansen BK, Wasteson Y, Granum PE, Brynestad S: **Mosaic structure of Shiga-toxin-2-encoding phages isolated from Escherichia coli O157:H7 indicates frequent gene exchange between lambdoid phage genomes**. *Microbiology* 2001, **147**(Pt 7):1929-1936.
21. Ho PL, Wong RC, Lo SW, Chow KH, Wong SS, Que TL: **Genetic identity of aminoglycoside-resistance genes in Escherichia coli isolates from human and animal sources**. *Journal of medical microbiology* 2010, **59**(Pt 6):702-707.
22. Nikaido H, Pages JM: **Broad-specificity efflux pumps and their role in multidrug resistance of Gram-negative bacteria**. *FEMS microbiology reviews* 2012, **36**(2):340-363.
23. Heiman KE, Grass JE, Sjolund-Karlsson M, Bowen A: **Shigellosis with decreased susceptibility to azithromycin**. *The Pediatric infectious disease journal* 2014, **33**(11):1204-1205.
24. Yoshida H, Bogaki M, Nakamura M, Nakamura S: **Quinolone resistance-determining region in the DNA gyrase gyrA gene of Escherichia coli**. *Antimicrobial agents and chemotherapy* 1990, **34**(6):1271-1272.
25. Hirose K, Terajima J, Izumiya H, Tamura K, Arakawa E, Takai N, Watanabe H: **Antimicrobial susceptibility of Shigella sonnei isolates in Japan and molecular analysis of S. sonnei isolates with reduced susceptibility to fluoroquinolones**. *Antimicrobial agents and chemotherapy* 2005, **49**(3):1203-1205.
26. Bagel S, Hullen V, Wiedemann B, Heisig P: **Impact of gyrA and parC mutations on quinolone resistance, doubling time, and supercoiling degree of Escherichia coli**. *Antimicrobial agents and chemotherapy* 1999, **43**(4):868-875.
27. Salipante SJ, SenGupta DJ, Cummings LA, Land TA, Hoogestraat DR, Cookson BT: **Application of whole-genome sequencing for bacterial strain typing in molecular epidemiology**. *Journal of clinical microbiology* 2015, **53**(4):1072-1079.
28. Fratamico PM, Liu Y, Kathariou S, American Society for Microbiology.: **Genomes of foodborne and waterborne pathogens**. Washington, DC: ASM Press; 2011.
29. Bielaszewska M, Mellmann A, Zhang W, Kock R, Fruth A, Bauwens A, Peters G, Karch H: **Characterisation of the Escherichia coli strain associated with an outbreak of haemolytic uraemic syndrome in Germany, 2011: a microbiological study**. *The Lancet Infectious diseases* 2011, **11**(9):671-676.
30. Frank C, Faber MS, Askar M, Bernard H, Fruth A, Gilsdorf A, Hohle M, Karch H, Krause G, Prager R *et al*: **Large and ongoing outbreak of haemolytic uraemic syndrome, Germany, May 2011**. *Euro surveillance : bulletin Europeen sur les maladies transmissibles = European communicable disease bulletin* 2011, **16**(21).
31. Rasko DA, Webster DR, Sahl JW, Bashir A, Boisen N, Scheutz F, Paxinos EE, Sebra R, Chin CS, Iliopoulos D *et al*: **Origins of the E. coli strain causing an outbreak of hemolytic-uremic syndrome in Germany**. *The New England journal of medicine* 2011, **365**(8):709-717.

32. Deghorain M, Bobay LM, Smeesters PR, Bousbata S, Vermeersch M, Perez-Morga D, Dreze PA, Rocha EP, Touchon M, Van Melderen L: **Characterization of novel phages isolated in coagulase-negative staphylococci reveals evolutionary relationships with Staphylococcus aureus phages**. *Journal of bacteriology* 2012, **194**(21):5829-5839.
33. Rohwer F, Edwards R: **The Phage Proteomic Tree: a genome-based taxonomy for phage**. *Journal of bacteriology* 2002, **184**(16):4529-4535.
34. Yang F, Yang J, Zhang X, Chen L, Jiang Y, Yan Y, Tang X, Wang J, Xiong Z, Dong J *et al*: **Genome dynamics and diversity of Shigella species, the etiologic agents of bacillary dysentery**. *Nucleic acids research* 2005, **33**(19):6445-6458.
35. Christopher PR, David KV, John SM, Sankarapandian V: **Antibiotic therapy for Shigella dysentery**. *The Cochrane database of systematic reviews* 2010(8):CD006784.
36. Boumghar-Bourtchai L, Mariani-Kurkdjian P, Bingen E, Filliol I, Dhalluin A, Ifrane SA, Weill FX, Leclercq R: **Macrolide-resistant Shigella sonnei**. *Emerging infectious diseases* 2008, **14**(8):1297-1299.
37. Sjolund Karlsson M, Bowen A, Reporter R, Folster JP, Grass JE, Howie RL, Taylor J, Whichard JM: **Outbreak of infections caused by Shigella sonnei with reduced susceptibility to azithromycin in the United States**. *Antimicrobial agents and chemotherapy* 2013, **57**(3):1559-1560.
38. Magiorakos AP, Srinivasan A, Carey RB, Carmeli Y, Falagas ME, Giske CG, Harbarth S, Hindler JF, Kahlmeter G, Olsson-Liljequist B *et al*: **Multidrug-resistant, extensively drug-resistant and pandrug-resistant bacteria: an international expert proposal for interim standard definitions for acquired resistance**. *Clinical microbiology and infection : the official publication of the European Society of Clinical Microbiology and Infectious Diseases* 2012, **18**(3):268-281.
39. Boerlin P, McEwen SA, Boerlin-Petzold F, Wilson JB, Johnson RP, Gyles CL: **Associations between virulence factors of Shiga toxin-producing Escherichia coli and disease in humans**. *Journal of clinical microbiology* 1999, **37**(3):497-503.
40. Louise CB, Obrig TG: **Specific interaction of Escherichia coli O157:H7-derived Shiga-like toxin II with human renal endothelial cells**. *The Journal of infectious diseases* 1995, **172**(5):1397-1401.
41. Ogura Y, Mondal SI, Islam MR, Mako T, Arisawa K, Katsura K, Ooka T, Gotoh Y, Murase K, Ohnishi M *et al*: **The Shiga toxin 2 production level in enterohemorrhagic Escherichia coli O157:H7 is correlated with the subtypes of toxin-encoding phage**. *Scientific reports* 2015, **5**:16663.
42. Jorgensen JH, Pfaller MA, Carroll KC, American Society for Microbiology: **Manual of clinical microbiology**, vol. 1, 11th edn; 2015.
43. Probert WS, McQuaid C, Schrader K: **Isolation and identification of an Enterobacter cloacae strain producing a novel subtype of Shiga toxin type 1**. *Journal of clinical microbiology* 2014, **52**(7):2346-2351.
44. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing S: **The Sequence Alignment/Map format and SAMtools**. *Bioinformatics* 2009, **25**(16):2078-2079.
45. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST *et al*: **The variant call format and VCFtools**. *Bioinformatics* 2011, **27**(15):2156-2158.
46. Stamatakis A: **RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models**. *Bioinformatics* 2006, **22**(21):2688-2690.
47. Joensen KG, Scheutz F, Lund O, Hasman H, Kaas RS, Nielsen EM, Aarestrup FM: **Real-time whole-genome sequencing for routine typing, surveillance, and outbreak detection of verotoxigenic Escherichia coli**. *Journal of clinical microbiology* 2014, **52**(5):1501-1510.

48. Zhou Y, Liang Y, Lynch KH, Dennis JJ, Wishart DS: **PHAST: a fast phage search tool**. *Nucleic acids research* 2011, **39**(Web Server issue):W347-352.

49. Wirth T, Falush D, Lan R, Colles F, Mensa P, Wieler LH, Karch H, Reeves PR, Maiden MC, Ochman H *et al*: **Sex and virulence in Escherichia coli: an evolutionary perspective**. *Molecular microbiology* 2006, **60**(5):1136-1151.

50. Darling AE, Mau B, Perna NT: **progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement**. *PloS one* 2010, **5**(6):e11147.

Figure Legends

Figure 1. Local and global phylogeny of CA *S. sonnei*. A. Maximum-likelihood phylogenetic tree based on genome-wide hqSNP differences between STX1-producing *S. sonnei* from San Diego and San Joaquin outbreak, STX-negative *S. sonnei* from San Francisco outbreak and historical isolates from CA. Background color: Red- San Diego/San Joaquin (SD/SJ) outbreak; Green- San Francisco outbreak; Blue- historical isolates; Yellow- STX-positive isolates. Frame line color: Red- designates major lineages I-III by Holt et al.; Purple- sub-lineage Global III by Holt et al. Node labels are displayed as "Isolate ID_Geographic location in California-Date of isolation". Numbers in orange circles correspond to the numbers of SNPs different between the closest isolates of the groups. **B.** Geographical distribution of the CA *S.sonnei* isolates. The international boundary between the US and Mexico is drawn in red. Color of the pins corresponds to the following groups of isolates: red- SD/SJ population, green- SF population, blue- historical isolates, white- modern sporadic isolates

Figure 2. Comparison of CA *S. sonnei* with *E. coli* strains and other publicly-available genomes of *S. sonnei*. Revised Maximum Likelihood clustering of CA representative isolates with *E.coli* and other *Shigella* species from IMG JGI database based on COG profiles (presence/absence). Background color: Red- representative *S. sonnei* from California; Green- other *S. sonnei* from JGI IMG database; Blue- *E. coli* O104:H4 strains from JGI IMG database.

Figure 3. Virulence determinants distribution in California *S. sonnei*

Virulence determinants names: *stx1A*- Shiga toxin 1, subunit A, variant a; *stx1B*- Shiga toxin 1, subunit B, variant a; *sigA*- Shigella IgA-like protease homologue; *gad*- Glutamate decarboxylase; *ipaH9.8*- Invasion plasmid antigen; *senB*- Plasmid-encoded enterotoxin; *prfB*- P-related fimbriae regulatory gene; *lpfA*- Long polar fimbriae; *celb*- Endonuclease colicin E2; *ipaD*- Invasion protein *S. flexneri*; *virF*- VirF transcriptional activator; *iss*- Increased serum survival; *sat*- Secreted autotransporter toxin.

Figure 4. Phylogeny of Shiga-toxin-encoding phage from SD/SJ *S. sonnei*. A. progressiveMauve alignment of phage sequences. Sequences are

853 centered by the *stxA* gene. **B.** A neighbor-joining tree based on an estimate
854 of the shared gene content from the progressiveMauve alignment.

855

856 **Figure 5. Antibiotic resistance determinants distribution in CA *S.***
857 ***sonnei*.** Abbreviations of antimicrobial classes: BL- Beta-lactams; AMG-
858 Aminoglycosides; SA-Sulphonamides; Tet-Tetracycline; TM- Trimethoprim;
859 MCL- Macrolide; PH- Phenicols; FQ- fluoroquinolones.

860

861 **Figure 6. Organization of genetic surrounding of macrolide**
862 **resistance gene in historical *S. sonnei* isolate C80 (Orange County,**
863 **1987).** Abbreviation: hp- hypothetical protein.

864

Figure 1

A.

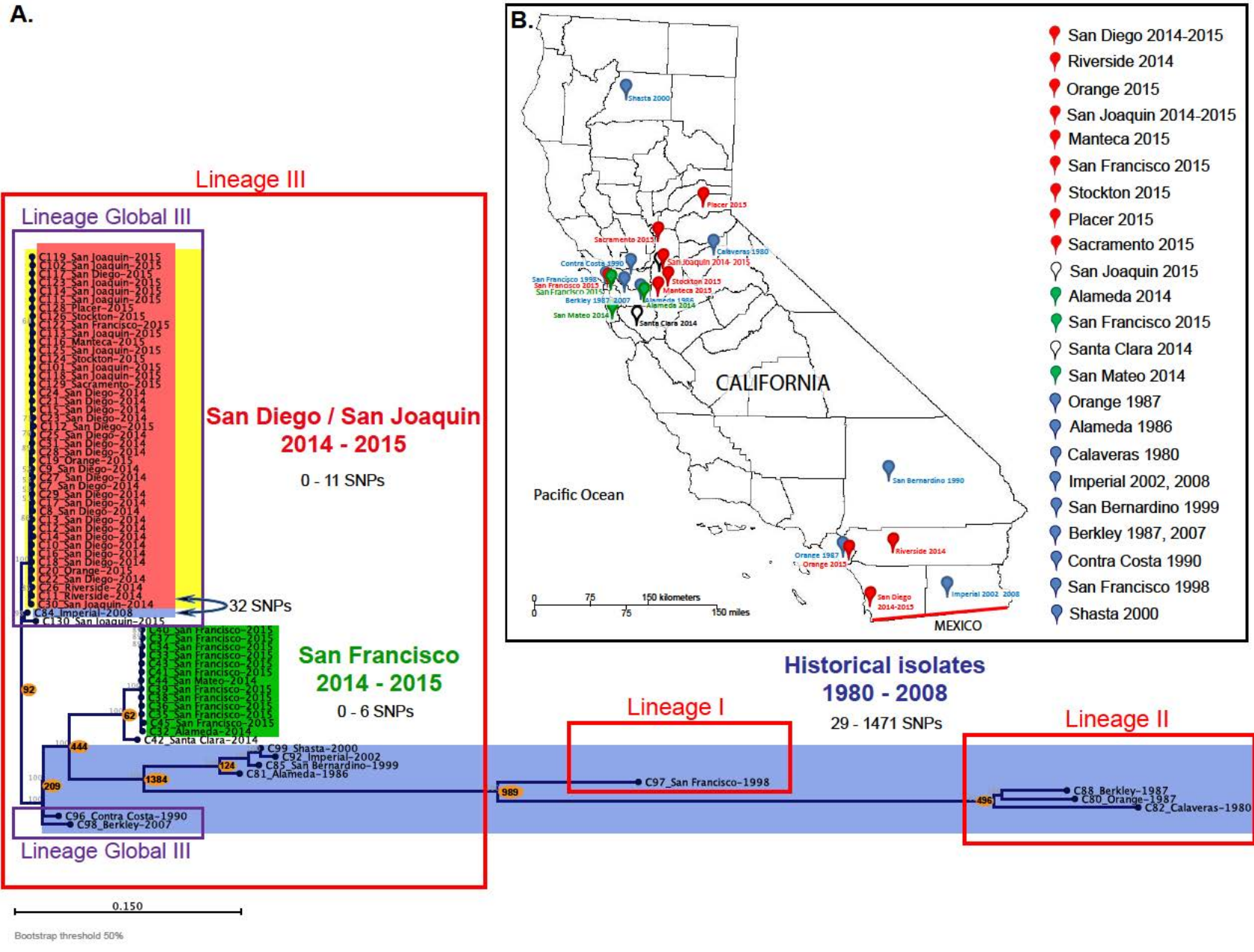

Lineage III
Lineage Global III
C119_San Joaquin-2015
C102_San Joaquin-2015
C117_San Diego-2015
C123_San Joaquin-2015
C114_San Joaquin-2015
C115_San Joaquin-2015
C128_Placer-2015
C126_Stockton-2015
C122_San Francisco-2015
C113_San Joaquin-2015
C116_Manteca-2015
C125_San Joaquin-2015
C124_Stockton-2015
C101_San Joaquin-2015
C118_San Joaquin-2015
C129_Sacramento-2015
C24_San Diego-2014
C21_San Diego-2014
C15_San Diego-2014
C23_San Diego-2014
C112_San Diego-2015
C25_San Diego-2014
C31_San Diego-2014
C28_San Diego-2014
C19_Orange-2015
C9_San Diego-2014
C27_San Diego-2014
C7_San Diego-2014
C29_San Diego-2014
C17_San Diego-2014
C8_San Diego-2014
C13_San Diego-2014
C12_San Diego-2014
C14_San Diego-2014
C10_San Diego-2014
C16_San Diego-2014
C18_San Diego-2014
C20_Orange-2015
C22_San Diego-2014
C26_Riverside-2014
C11_Riverside-2014
C30_San Joaquin-2014
C84_Imperial-2008
C130_San Joaquin-2015
San Diego / San Joaquin
2014 - 2015
0 - 11 SNPs
32 SNPs
C40_San Francisco-2015
C37_San Francisco-2015
C34_San Francisco-2015
C33_San Francisco-2015
C43_San Francisco-2015
C41_San Francisco-2015
C44_San Mateo-2014
C39_San Francisco-2015
C38_San Francisco-2015
C36_San Francisco-2015
C35_San Francisco-2015
C45_San Francisco-2015
C32_Alameda-2014
San Francisco
2014 - 2015
0 - 6 SNPs
C42_Santa Clara-2014
C99_Shasta-2000
C92_Imperial-2002
C85_San Bernardino-1999
C81_Alameda-1986
C96_Contra Costa-1990
C98_Berkley-2007
Lineage Global III
92
62
444
124
1384
209
989
C97_San Francisco-1998
Lineage I
Historical isolates
1980 - 2008
29 - 1471 SNPs
Lineage II
496
C88_Berkley-1987
C80_Orange-1987
C82_Calaveras-1980
0.150
Bootstrap threshold 50%
B.
Shasta 2000
Placer 2015
Sacramento 2015
Calaveras 1980
Contra Costa 1990
San Joaquin 2014-2015
San Francisco 1998
San Francisco 2015
Stockton 2015
Manteca 2015
Berkley 1987, 2007
Alameda 1986
Alameda 2014
San Mateo 2014
Santa Clara 2014
CALIFORNIA
Pacific Ocean
San Bernardino 1990
Orange 1987
Orange 2015
Riverside 2014
San Diego
2014-2015
Imperial 2002, 2008
0
75
150 kilometers
0
75
150 miles
MEXICO
San Diego 2014-2015
Riverside 2014
Orange 2015
San Joaquin 2014-2015
Manteca 2015
San Francisco 2015
Stockton 2015
Placer 2015
Sacramento 2015
San Joaquin 2015
Alameda 2014
San Francisco 2015
Santa Clara 2014
San Mateo 2014
Orange 1987
Alameda 1986
Calaveras 1980
Imperial 2002, 2008
San Bernardino 1999
Berkley 1987, 2007
Contra Costa 1990
San Francisco 1998
Shasta 2000

Figure 2

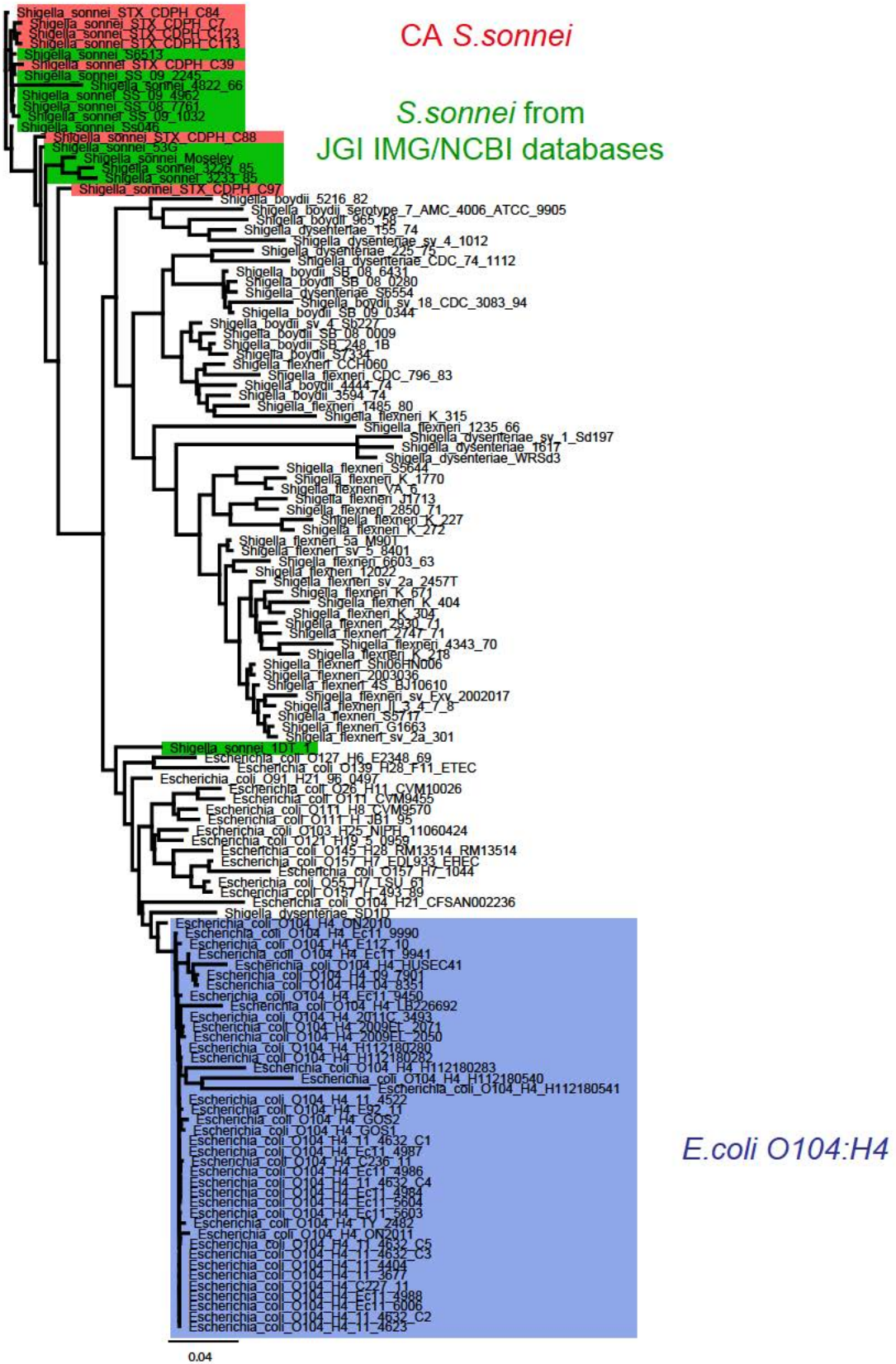
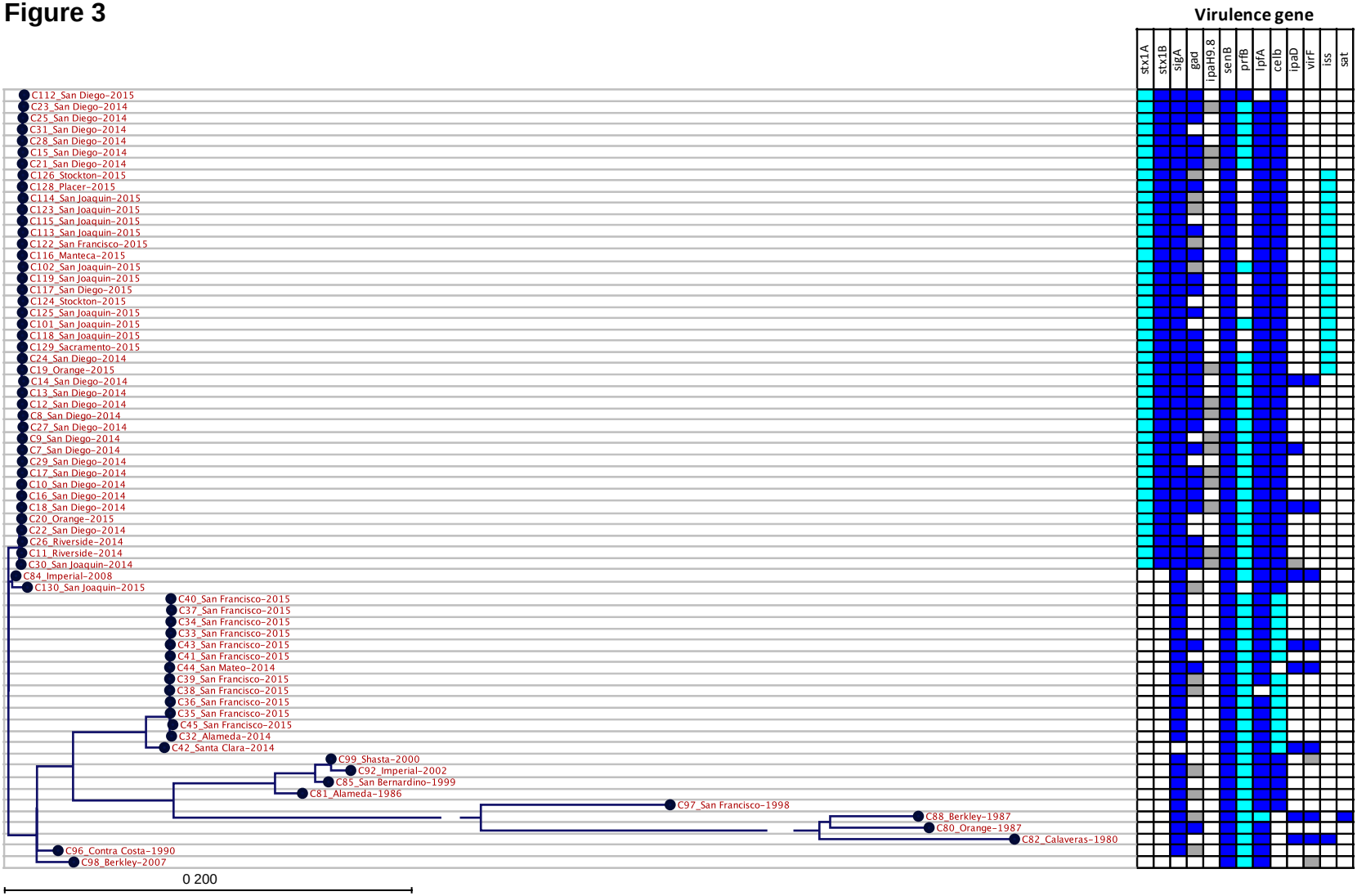

CA S.sonnei
S.sonnei from JGI IMG/NCBI databases
E.coli O104:H4
0.04

Figure 3



- Virulence gene present;
- Virulence gene with 99.77-99.88% ID to VirulenceFinder database sequence present;
- Incomplete sequence of the virulence gene present

Figure 4


A.
CA Shigella sonnei stx-phage
Escherichia coli O104:H4 str 2011C-3493 phage
Escherichia coli O104:H4 phage P13374
Escherichia coli O104:H4 str 2009EL-2071 phage
Escherichia coli O111:H2 phage phi191
Escherichia coli O118 str.FHI29 phage
Escherichia coli O157:H7 str.Sakai phage
Escherichia coli O157:H7 str.EDL933 phage
Shigella flexneri phage POCJ13
Shigella sonnei phage HUN/2013
stx gene

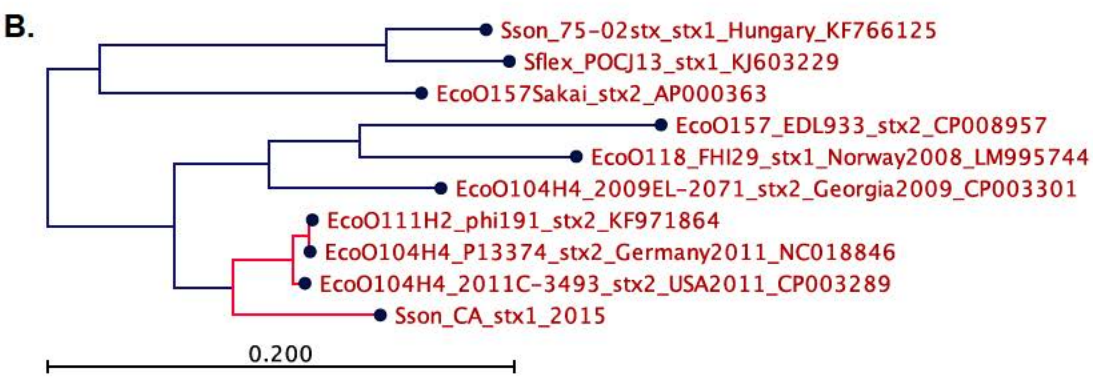
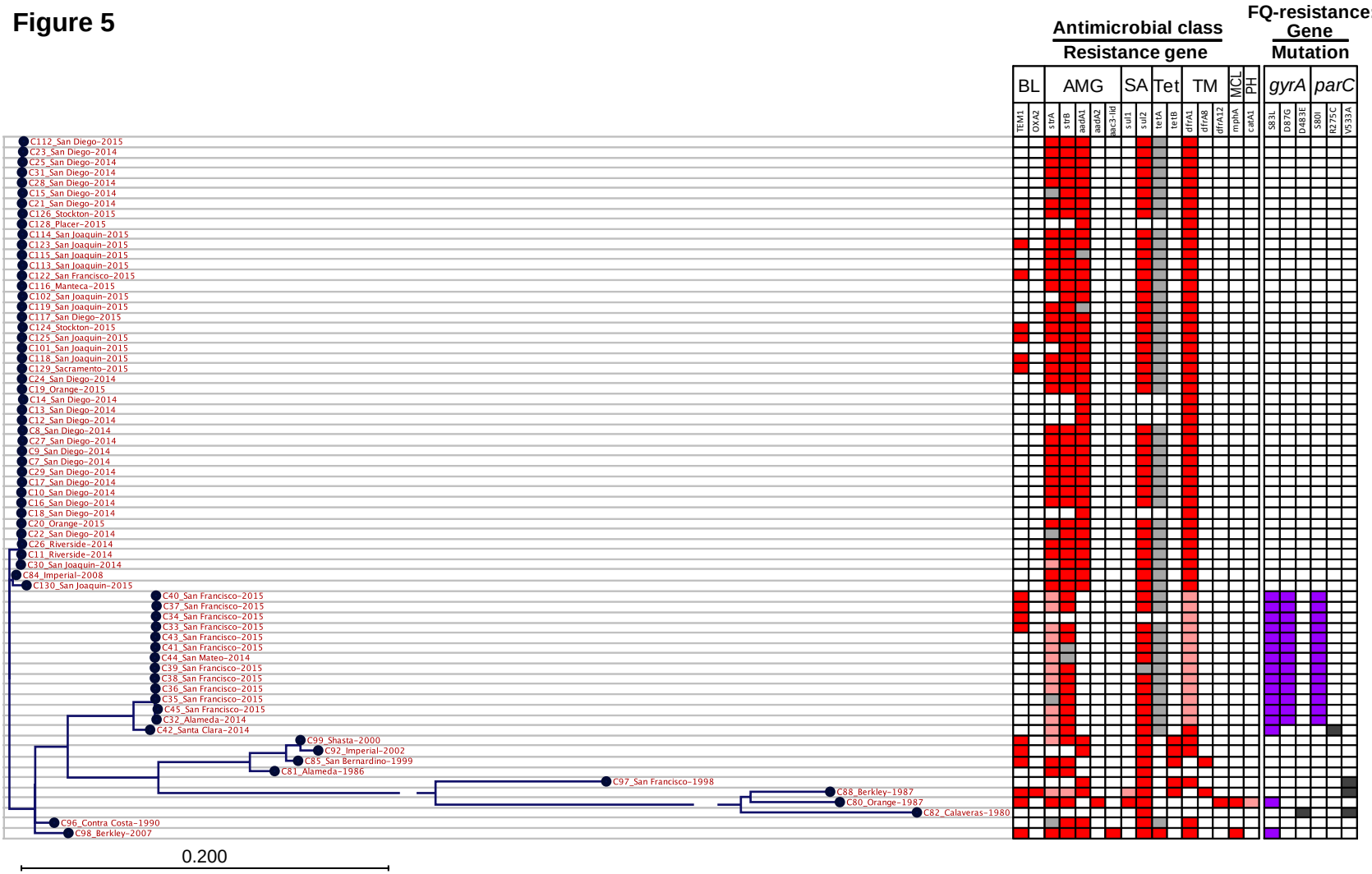
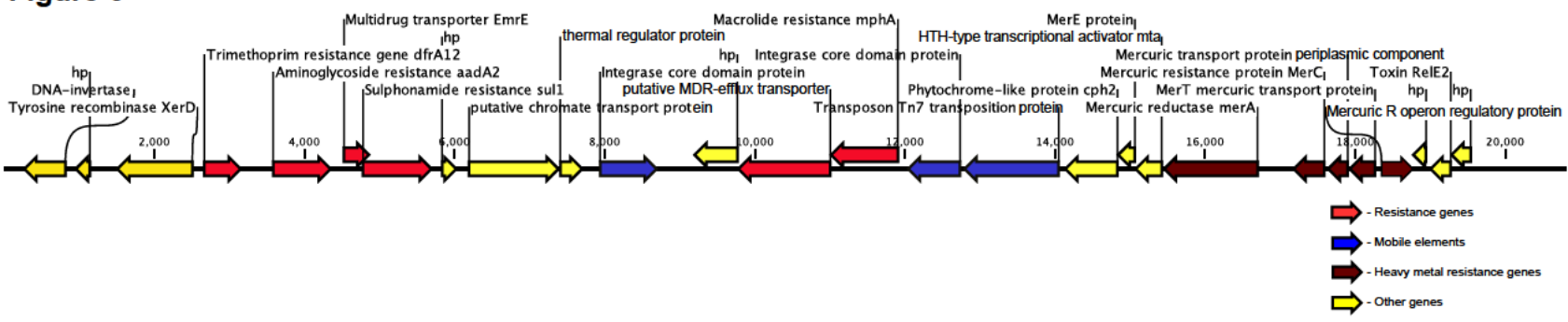

B.
Sson_75-02stx_stx1_Hungary_KF766125
Sflex_POCJ13_stx1_KJ603229
EcoO157Sakai_stx2_AP000363
EcoO157_EDL933_stx2_CP008957
EcoO118_FHI29_stx1_Norway2008_LM995744
EcoO104H4_2009EL-2071_stx2_Georgia2009_CP003301
EcoO111H2_phi191_stx2_KF971864
EcoO104H4_P13374_stx2_Germany2011_NC018846
EcoO104H4_2011C-3493_stx2_USA2011_CP003289
Sson_CA_stx1_2015
0.200

Figure 5



- Resistance gene present;
- Resistance gene with 99.77 99.88 % ID to ResFinder database sequence present;
- Incomplete sequence of the resistance gene present;
- FQ-resistance conferring mutation;
- No data available about effect of the mutation on the FQ-resistance.

Figure 6


Multidrug transporter EmrE
hp
Trimethoprim resistance gene dfrA12
Aminoglycoside resistance aadA2
Sulphonamide resistance sul1
putative chromate transport protein
thermal regulator protein
Macrolide resistance mphA
hp
Integrase core domain protein
putative MDR-efflux transporter
Transposon Tn7 transposition protein
MerE protein
HTH-type transcriptional activator mta
Integrase core domain protein
Phytochrome-like protein cph2
Mercuric reductase merA
Mercuric transport protein periplasmic component
Mercuric resistance protein MerC
MerT mercuric transport protein
Toxin RelE2
hp
hp
Mercuric R operon regulatory protein
hp
DNA-invertase
Tyrosine recombinase XerD
2,000
4,000
6,000
8,000
10,000
12,000
14,000
16,000
18,000
20,000
- Resistance genes
- Mobile elements
- Heavy metal resistance genes
- Other genes

SUPPLEMENTAL MATERIAL

Figure S1. Clustering of CA *S. sonnei* isolates with *S. sonnei* strains of global lineages as per Holt et al. based on a maximum-likelihood phylogenetic tree built using genome-wide hqSNPs. Color of node labels: Red- *S.sonnei* isolates from CA; Blue- global isolates from Holt et al. publication. Branches highlight color: Yellow- Lineage I; Blue- Lineage II; Green- Lineage III; Purple- Lineage IV. Tree is rooted to *E. coli*.

Figure S2. Hierarchical Clustering of CA representative *S. sonnei* isolates with *E.coli* and other *Shigella* species from JGI IMG database. A. Clustering based on COG profiles (presence/absence & abundance). Color of the brackets: Red- representative *S.sonnei* from California; Green- other *S. sonnei* from JGI IMG database; Blue- *E.coli* O104:H4 strains from JGI IMG database. **B.** Clustering based on pfam profiles (presence/absence & abundance). Color of the brackets: Red- representative *S.sonnei* from California; Green- other *S. sonnei* from JGI IMG database; Blue- *E.coli* O104:H4 strains from JGI IMG database.

Figure S3. Comparison of CA *S. sonnei* with *E. coli* strains and other publicly-available genomes of *S. sonnei* based on nucleotide sequence. A. Maximum likelihood clustering from PhyloSift nucleotide-based phylogeny. Bootstrap threshold 20%. **B.** Maximum likelihood phylogeny of CA *S. sonnei*, *E.coli*, and publicly-available *S. sonnei* genomes

based on genome-wide hqSNPs. Bootstrap threshold 50%. Background color: Red- *S.sonnei* from California; Green- other S. sonnei from JGI IMG and NCBI databases; Blue- *E.coli* O104:H4 strains from JGI IMG database.

Figure S4. Comparison of COG abundance profiles of representative CA *S. sonnei* with *E. coli* strains and other *Shigella* species from JGI IMG database. The heat map represents gene count for different COGs.

Figure S5. Revised pfam phylogeny. Maximum Likelihood clustering of CA representative *S.sonnei* isolates with *E.coli* and other *Shigella* species from IMG JGI database based on pfam profiles (presence/absence). Background color: Red- representative *S.sonnei* from California; Green- other *S. sonnei* from JGI IMG database; Blue- *E.coli* O104:H4 strains from JGI IMG database.

Figure S6. Neighbor-Joining phylogeny of the CA STX1-phage based on amino acid sequence of Integrase protein. Subtree containing CA STX1-phage is highlighted with red branch color.

Figure S7. Distribution of CA STX1-bacteriophage genes in different *E.coli* and *Shigella* serovars from IMG database. Gene IDs from IMG database are listed on the left side of the graph. Color of the blocks: red - gene is present, blue - gene is absent. Min 10% ID was used as a similarity cutoff.

Figure S8. Cryptic STX-converting prophages found in CA *S.sonnei* genomes.

Figure S9. Plasmids and other mobile elements encoding antibiotic resistance in CA *S. sonnei*. A. Organization of *bla*$_{TEM-1}$- encoding IncB/O/K/Z conjugative plasmid from modern SF population isolates. Different integration sites of Tn3::*bla*$_{TEM-1}$ on IncB/O/K/Z plasmids found in recent SF isolates and in historical *S. sonnei*. **B.** A representative genetic surrounding of *bla*$_{TEM-1}$ gene on a putative IncI1 conjugative plasmid from a modern SJ *S. sonnei* isolate. **C.** Representative organization of *sul2-strA-strB-tetA* resistance genes cluster in SD/SJ and historical CA *S. sonnei* isolates.

56

Supplementary Table 1. List of CA *S. sonnei* isolates

Isolate ID	Species	Shiga-toxin gene	Date of isolation	Geographical location (CA County)	MLST type	Lineage (Holt et al.)*	Outbreak code (PFGE)	NCBI accession number (assembly / raw reads)
C7	Shigella sonnei	*stx1*	2014	San Diego	ST-152	III G	1504CAJ16-1	LXVV00000000 / SRR3441855
C8	Shigella sonnei	*stx1*	2014	San Diego	ST-152	III G	1504CAJ16-1	LRRZ00000000 / SRR3441856
C9	Shigella sonnei	*stx1*	2014	San Diego	ST-152	III G	1504CAJ16-1	LRSA00000000 / SRR3441868
C10	Shigella sonnei	*stx1*	2014	San Diego	ST-152	III G	1504CAJ16-1	LRSB00000000 / SRR3441880
C11	Shigella sonnei	*stx1*	2014	Riverside	ST-152	III G	1504CAJ16-1	LRSC00000000 / SRR3447562
C12	Shigella sonnei	*stx1*	2014	San Diego	ST-152	III G	1504CAJ16-1	LRSD00000000 / SRR3452059
C13	Shigella sonnei	*stx1*	2014	San Diego	ST-152	III G	1504CAJ16-1	LXVU00000000 / SRR3452073
C14	Shigella sonnei	*stx1*	2014	San Diego	ST-152	III G	1504CAJ16-1	LXVT00000000 / SRR3452083
C15	Shigella sonnei	*stx1*	2014	San Diego	ST-152	III G	1504CAJ16-1	LYEW00000000 / SRR3452085
C16	Shigella sonnei	*stx1*	2014	San Diego	ST-152	III G	1504CAJ16-1	LYEV00000000 / SRR3452086
C17	Shigella sonnei	*stx1*	2014	San Diego	ST-152	III G	1504CAJ16-1	LXVS00000000 / SRR3441857
C18	Shigella sonnei	*stx1*	2014	San Diego	ST-152	III G	1504CAJ16-1	LXVR00000000 / SRR3441858
C19	Shigella sonnei	*stx1*	2015	Orange	ST-152	III G	1504CAJ16-1	LYEU00000000 / SRR3441859
C20	Shigella sonnei	*stx1*	2015	Orange	ST-152	III G	1504CAJ16-1	LXVQ00000000 / SRR3441860
C21	Shigella sonnei	*stx1*	2014	San Diego	ST-152	III G	1504CAJ16-1	LXVP00000000 / SRR3441861
C22	Shigella sonnei	*stx1*	2014	San Diego	ST-152	III G	1504CAJ16-1	LXVO00000000 / SRR3441862
C23	Shigella sonnei	*stx1*	2014	San Diego	ST-152	III G	1504CAJ16-1	LXVN00000000 / SRR3441863
C24	Shigella sonnei	*stx1*	2014	San Diego	ST-152	III G	1504CAJ16-1	LXVM00000000 / SRR3441865
C25	Shigella sonnei	*stx1*	2014	San Diego	ST-152	III G	1504CAJ16-1	LXVL00000000 / SRR3441866
C26	Shigella sonnei	*stx1*	2014	Riverside	ST-152	III G	1504CAJ16-1	LXVK00000000 / SRR3441867
C27	Shigella sonnei	*stx1*	2014	San Diego	ST-152	III G	1504CAJ16-1	LXVJ00000000 / SRR3441869
C28	Shigella sonnei	*stx1*	2014	San Diego	ST-152	III G	1504CAJ16-1	LXVI00000000 / SRR3441870
C29	Shigella sonnei	*stx1*	2014	San Diego	ST-152	III G	1504CAJ16-1	LXVH00000000 / SRR3441871
C30	Shigella sonnei	*stx1*	2014	San Joaquin	ST-152	III G	1504CAJ16-1	LYET00000000 / SRR3452090
C31	Shigella sonnei	*stx1*	2014	San Diego	ST-152	III G	1504CAJ16-1	LYES00000000 / SRR3441873
C32	Shigella sonnei	negative	2014	Alameda	ST-152	III	1407MLJ16-2	LYER00000000 / SRR3441874
C33	Shigella sonnei	negative	2015	San Francisco	ST-152	III	1407MLJ16-2	LYEQ00000000 / SRR3441875
C34	Shigella sonnei	negative	2015	San Francisco	ST-152	III	1407MLJ16-2	LYEP00000000 / SRR3441876

57

58 Supplementary Table 1 continued.

C35	Shigella sonnei	negative	2015	San Francisco	ST-152	III	1407MLJ16-2	LYEO00000000 / SRR3441878
C36	Shigella sonnei	negative	2015	San Francisco	ST-152	III	1407MLJ16-2	LYEN00000000 / SRR3441879
C37	Shigella sonnei	negative	2015	San Francisco	ST-152	III	1407MLJ16-2	LYEM00000000 / SRR3441881
C38	Shigella sonnei	negative	2015	San Francisco	ST-152	III	1407MLJ16-2	LYEL00000000 / SRR3441882
C39	Shigella sonnei	negative	2015	San Francisco	ST-152	III	1407MLJ16-2	LYEK00000000 / SRR3441883
C40	Shigella sonnei	negative	2015	San Francisco	ST-152	III	1407MLJ16-2	LYEJ00000000 / SRR3441884
C41	Shigella sonnei	negative	2015	San Francisco	ST-152	III	1407MLJ16-2	LYEI00000000 / SRR3441886
C42	Shigella sonnei	negative	2014	Santa Clara	ST-152	III	1407MLJ16-2	LYEH00000000 / SRR3441887
C43	Shigella sonnei	negative	2015	San Francisco	ST-152	III	1407MLJ16-2	LYEG00000000 / SRR3442202
C44	Shigella sonnei	negative	2014	San Mateo	ST-152	III	1407MLJ16-2	LYEF00000000 / SRR3443504
C45	Shigella sonnei	negative	2015	San Francisco	ST-152	III	1407MLJ16-2	LYEE00000000 / SRR3445394
C80	Shigella sonnei	negative	1987	Orange	ST-152	II	N/A (historical)	LYED00000000 / SRR3446557
C81	Shigella sonnei	negative	1986	Alameda	ST-152	III	N/A (historical)	LYEC00000000 / SRR3448695
C82	Shigella sonnei	negative	1980	Calaveras	ST-152	II	N/A (historical)	LXVG00000000 / SRR3449672
C84	Shigella sonnei	negative	2008	Imperial	ST-152	III G	N/A (historical)	LYEB00000000 / SRR3451341
C85	Shigella sonnei	negative	1999	San Bernardino	ST-152	III	N/A (historical)	LXVF00000000 / SRR3452052
C88	Shigella sonnei	negative	1987	Berkley	ST-152	II	N/A (historical)	LXVE00000000 / SRR3452053
C92	Shigella sonnei	negative	2002	Imperial	ST-152	III	N/A (historical)	LXVD00000000 / SRR3452054
C96	Shigella sonnei	negative	1990	Contra Costa	ST-152	III G	N/A (historical)	LYEA00000000 / SRR3452055
C97	Shigella sonnei	negative	1998	San Francisco	ST-1502	I	N/A (historical)	LYDZ00000000 / SRR3452056
C98	Shigella sonnei	negative	2007	Berkley	ST-152	III G	N/A (historical)	LYDY00000000 / SRR3452057
C99	Shigella sonnei	negative	2000	Shasta	ST-152	III	N/A (historical)	LYDX00000000 / SRR3452058
C101	Shigella sonnei	*stx1*	2015	San Joaquin	ST-152	III G	1504CAJ16-1	LXVC00000000 / SRR3452061
C102	Shigella sonnei	*stx1*	2015	San Joaquin	ST-152	III G	1504CAJ16-1	LXVB00000000 / SRR3452062
C112	Shigella sonnei	*stx1*	2015	San Diego	ST-152	III G	1504CAJ16-1	LXVA00000000 / SRR3452063
C113	Shigella sonnei	*stx1*	2015	San Joaquin	ST-152	III G	1504CAJ16-1	LXUZ00000000 / SRR3452064
C114	Shigella sonnei	*stx1*	2015	San Joaquin	ST-152	III G	1504CAJ16-1	LXUY00000000 / SRR3452065
C115	Shigella sonnei	*stx1*	2015	San Joaquin	ST-152	III G	1504CAJ16-1	LXUX00000000 / SRR3452066
C116	Shigella sonnei	*stx1*	2015	Manteca	ST-152	III G	1504CAJ16-1	LXUW00000000 / SRR3452067

59

Supplementary Table 1 continued.

C117	Shigella sonnei	*stx1*	2015	San Diego	ST-152	III G	1504CAJ16-1	LXUV00000000 / SRR3452069
C118	Shigella sonnei	*stx1*	2015	San Joaquin	ST-152	III G	1504CAJ16-1	LXUU00000000 / SRR3452070
C119	Shigella sonnei	*stx1*	2015	San Joaquin	ST-152	III G	1504CAJ16-1	LXUT00000000 / SRR3452071
C122	Shigella sonnei	*stx1*	2015	San Francisco	ST-152	III G	1504CAJ16-1	LXUS00000000 / SRR3452074
C123	Shigella sonnei	*stx1*	2015	San Joaquin	ST-152	III G	1504CAJ16-1	LXUR00000000 / SRR3452075
C124	Shigella sonnei	*stx1*	2015	Stockton	ST-152	III G	1504CAJ16-1	LXUQ00000000 / SRR3452076
C125	Shigella sonnei	*stx1*	2015	San Joaquin	ST-152	III G	1504CAJ16-1	LXUP00000000 / SRR3452077
C126	Shigella sonnei	*stx1*	2015	Stockton	ST-152	III G	1504CAJ16-1	LXUO00000000 / SRR3452078
C128	Shigella sonnei	*stx1*	2015	Placer	ST-152	III G	1504CAJ16-1	LXUN00000000 / SRR3452080
C129	Shigella sonnei	*stx1*	2015	Sacramento	ST-152	III G	1504CAJ16-1	LXUM00000000 / SRR3452081
C130	Shigella sonnei	negative	2015	San Joaquin	ST-152	III G	Not assigned to an outbreak	LXUL00000000 / SRR3452082

*Lineage designation I-IV according to Holt et al. publication; III G-Global III Lineage.

Supplementary Table 2. List of isolates from Holt et al. publication used for phylogenetic comparison.

ID	Isolate	Country	Region	Year	Lineage	Global III	ENA Accession Number
54210	54210	Sweden	Europe	1943	III	no	ERR025732
54179	54179	Sweden	Europe	1944	II	no	ERR025727
5827	#00 5827	Madagascar	East Africa & Madagascar	2000	I	no	ERR025765
54185	54185	Denmark	Europe	1945	II	no	ERR025730
259	2-59	France	Europe	1959	IV	no	ERR025737
4474	44-74	France	Europe	1974	I	no	ERR025755
IB694	IB694	Korea	Korea	1979	II	no	ERR024619
19984123	19984123	Mexico	Caribbean/Central America	1998	III	no	ERR028691
CS20	CS20	Brazil	South America	2002	III	no	ERR028685
74369	#07 4369	France	Europe	2007	I	no	ERR024606
20062087	20062087	Egypt_Tunisia	Egypt	2006	III	yes	ERR028699
20071599	20071599	UK	Europe	2007	II	no	ERR024092
32222	#03 2222	Cuba	Caribbean/Central America	2003	III	yes	ERR025768
20081885	20081885	UK	Europe	2008	III	yes	ERR024070

Supplementary Table 3. Genes of representative SD/SJ *S.sonnei* isolates with homologs in *E.coli* O104:H4, but not in other *S. sonnei* (except *S.sonnei* strain 1DT-1-).

Green highlight- STX-bacteriophage genes; yellow- bla_{TEM-1} plasmid.

Homologies in SD/SJ *S.sonnei* isolate C7:

Result	Gene Object ID	Locus Tag	Gene Name	Length	COG	Pfam	% ID
1	2634646959	Ga0082014_10014	hypothetical protein	2793	-	-	97.9
2	2634646960	Ga0082014_10015	hypothetical protein	421	-	-	93.6
3	2634646961	Ga0082014_10016	Bacteriocin class II with double-glycine leader peptide	83	-	pfam10439	100
4	2634646962	Ga0082014_10017	hypothetical protein	148	-	-	99.3
5	2634646964	Ga0082014_10019	hypothetical protein	133	-	-	100
6	2634646967	Ga0082014_100112	hypothetical protein	205	-	-	100
7	2634646969	Ga0082014_100114	protein of unknown function (DUF1983)	422	-	pfam09327	99.5
8	2634646970	Ga0082014_100115	hypothetical protein	563	-	-	99
9	2634646971	Ga0082014_100116	Phage tail fibre adhesin Gp38	170	-	pfam05268	100
10	2634646973	Ga0082014_100118	hypothetical protein	216	-	-	99.5
11	2634646974	Ga0082014_100119	hypothetical protein	187	-	-	100
12	2634646975	Ga0082014_100120	hypothetical protein	153	-	-	100
13	2634646978	Ga0082014_100123	hypothetical protein	335	-	-	99.4
14	2634646979	Ga0082014_100124	hypothetical protein	714	-	-	100
15	2634646981	Ga0082014_100126	hypothetical protein	268	-	-	100
16	2634646985	Ga0082014_100130	hypothetical protein	44	-	-	100
17	2634646987	Ga0082014_100132	lysozyme	177	COG3772	pfam00959	96.6
18	2634646989	Ga0082014_100134	Protein of unknown function (DUF826)	81	-	pfam05696	96.3
19	2634646992	Ga0082014_100137	shiga toxin subunit B	89	-	pfam02258	58
20	2634646993	Ga0082014_100138	shiga toxin subunit A	315	-	pfam00161	54.8
21	2634646995	Ga0082014_100140	Phage NinH protein	64	-	pfam06322	100
22	2634646996	Ga0082014_100141	Bacteriophage Lambda NinG protein	187	-	pfam05766	100
23	2634646997	Ga0082014_100142	Protein of unknown function (DUF1367)	149	-	pfam07105	100
24	2634646999	Ga0082014_100144	ATP-dependent helicase IRC3	506	COG1061	pfam00271 pfam04851	98.2
25	2634647000	Ga0082014_100145	hypothetical protein	125	-	-	100
26	2634647001	Ga0082014_100146	Cro protein	67	-	pfam09048	100
27	2634647002	Ga0082014_100147	SOS-response transcriptional repressor LexA (RecA-mediated autopeptidase)	237	COG1974	pfam00717 pfam01381	100
28	2634647004	Ga0082014_100149	BsuBI/PstI restriction endonuclease C-terminus	317	-	pfam06616	100
29	2634647008	Ga0082014_100154	hypothetical protein	70	-	-	100
30	2634647009	Ga0082014_100155	hypothetical protein	73	-	-	100
31	2634647010	Ga0082014_100156	hypothetical protein	93	-	-	100
32	2634647011	Ga0082014_100157	recombination protein RecT	316	COG3723	pfam03837	99.1
33	2634647012	Ga0082014_100158	YqaJ-like recombinase domain-containing protein	229	-	pfam09588	100
34	2634647013	Ga0082014_100159	hypothetical protein	195	-	-	99
35	2634647016	Ga0082014_100162	hypothetical protein	147	-	-	90.1
36	2634647020	Ga0082014_100166	Ead/Ea22-like protein	233	-	pfam13935	95.5
37	2634647021	Ga0082014_100167	hypothetical protein	171	-	-	100
38	2634647025	Ga0082014_100171	hypothetical protein	143	-	-	100
39	2634647026	Ga0082014_100172	N6-adenosine-specific RNA methylase IME4	208	COG4725	pfam05063	100
40	2634647028	Ga0082014_100174	Protein of unknown function (DUF1627)	129	-	pfam07789	96.9
41	2634647030	Ga0082014_100176	protein of unknown function (DUF4222)	83	-	pfam13973	98.8

Supplementary Table 3 continued.

Homologies in SD / SJ *S.sonnei* isolate C123:

Result	Gene Object ID	Locus Tag	Gene Name	Length	COG	Pfam	% ID
1	2640155234	Ga0099679_10193	protein of unknown function (DUF4222)	83	-	pfam13973	98.8
2	2640155236	Ga0099679_10195	Protein of unknown function (DUF1627)	129	-	pfam07789	89.9
3	2640155238	Ga0099679_10197	N6-adenosine-specific RNA methylase IME4	208	COG4725	pfam05063	99
4	2640155239	Ga0099679_10198	hypothetical protein	143	-	-	100
5	2640155243	Ga0099679_101912	hypothetical protein	171	-	-	100
6	2640155245	Ga0099679_101914	hypothetical protein	79	-	-	88.6
7	2640155246	Ga0099679_101915	hypothetical protein	78	-	-	91
8	2640155251	Ga0099679_101920	hypothetical protein	195	-	-	99
9	2640155254	Ga0099679_101923	hypothetical protein	93	-	-	100
10	2640155255	Ga0099679_101924	hypothetical protein	73	-	-	100
11	2640155256	Ga0099679_101925	hypothetical protein	70	-	-	100
12	2640155260	Ga0099679_101930	BsuBI/PstI restriction endonuclease C-terminus	317	-	pfam06616	100
13	2640155263	Ga0099679_101933	Cro protein	67	-	pfam09048	100
14	2640155264	Ga0099679_101934	hypothetical protein	125	-	-	100
15	2640155265	Ga0099679_101935	ATP-dependent helicase IRC3	506	COG1061	pfam04851 pfam00271	98.2
16	2640155267	Ga0099679_101937	Protein of unknown function (DUF1367)	149	-	pfam07105	100
17	2640155268	Ga0099679_101938	Bacteriophage Lambda NinG protein	187	-	pfam05766	100
18	2640155271	Ga0099679_101941	shiga toxin subunit A	315	-	pfam00161	54.8
19	2640155272	Ga0099679_101942	shiga toxin subunit B	89	-	pfam02258	58
20	2640155275	Ga0099679_101945	Protein of unknown function (DUF826)	81	-	pfam05696	96.3
21	2640155283	Ga0099679_101953	hypothetical protein	268	-	-	100
22	2640155285	Ga0099679_101955	hypothetical protein	714	-	-	98
23	2640155286	Ga0099679_101956	hypothetical protein	335	-	-	98.5
24	2640155288	Ga0099679_101958	hypothetical protein	129	-	-	98.4
25	2640155289	Ga0099679_101959	hypothetical protein	153	-	-	99.3
26	2640155290	Ga0099679_101960	hypothetical protein	187	-	-	98.9
27	2640155291	Ga0099679_101961	hypothetical protein	216	-	-	98.1
28	2640156889	Ga0099679_109011	Site-specific recombinase XerD	387	COG4974	pfam00589	73.4
29	2640158194	Ga0099679_11592	hypothetical protein	2793	-	-	97.9
30	2640158195	Ga0099679_11593	hypothetical protein	421	-	-	93.6
31	2640158196	Ga0099679_11594	Bacteriocin class II with double-glycine leader peptide	83	-	pfam10439	100
32	2640158197	Ga0099679_11595	hypothetical protein	148	-	-	99.3
33	2640158198	Ga0099679_11596	hypothetical protein	218	-	-	99.1
34	2640158199	Ga0099679_11597	hypothetical protein	133	-	-	98.5
35	2640158202	Ga0099679_115910	hypothetical protein	205	-	-	100
36	2640158204	Ga0099679_115912	protein of unknown function (DUF1983)	422	-	pfam09327	97.9
37	2640158205	Ga0099679_115913	hypothetical protein	563	-	-	98
38	2640158206	Ga0099679_115914	Phage tail fibre adhesin Gp38	170	-	pfam05268	100
39	2640158457	Ga0099679_11721	shufflon protein, N-terminal constant region	361	-	pfam04917	100
40	2640158458	Ga0099679_11722	Type IV leader peptidase family protein	218	-	pfam01478	100
41	2640158460	Ga0099679_11724	PilS N terminal	204	-	pfam08805	100
42	2640158461	Ga0099679_11725	Type II secretory pathway, component PulF	361	COG1459	pfam00482	100
43	2640158463	Ga0099679_11727	type IV pilus biogenesis protein PilP	150	-	pfam11356	98.7
44	2640158466	Ga0099679_117210	PilM protein	145	-	pfam07419	97.9
45	2640158467	Ga0099679_117211	Toxin co-regulated pilus biosynthesis protein Q	355	-	pfam10671	100

79 Supplementary Table 3 continued.

46	2640158472	Ga0099679_117216	Protein of unknown function (DUF2913)	227	-	pfam11140	99.1
47	2640158473	Ga0099679_117217	Transcription termination factor nusG	177	-	pfam02357	99.4
48	2640158474	Ga0099679_117218	hypothetical protein	95	-	-	100
49	2640158478	Ga0099679_117223	ProQ/FINO family protein	200	-	pfam04352	97
50	2640158479	Ga0099679_117224	hypothetical protein	448	-	-	97.8
51	2640158480	Ga0099679_117225	Transposase and inactivated derivatives, TnpA family	1001	COG4644	pfam01526 pfam13700	100
52	2640158482	Ga0099679_117227	beta-lactamase class A TEM	286	COG2367	pfam13354	100
53	2640158490	Ga0099679_117235	chromosome partitioning protein	214	COG1192	pfam01656	99.5
54	2640158496	Ga0099679_117241	DNA-damage-inducible protein I	82	-	pfam06183	100
55	2640158497	Ga0099679_117242	Protein of unknown function (DUF1281)	308	-	pfam06924	92.9
56	2640158499	Ga0099679_117244	hypothetical protein	73	-	-	98.6
57	2640158500	Ga0099679_117245	Protein of unknown function (DUF1380)	144	-	pfam07128	97.9
58	2640158501	Ga0099679_117246	hypothetical protein	258	-	-	98.8
59	2640158504	Ga0099679_117249	Protein of unknown function (DUF1380)	140	-	pfam07128	98.6
60	2640158505	Ga0099679_117250	hypothetical protein	63	-	-	100
61	2640158508	Ga0099679_117253	chromosome partitioning protein, ParB family	652	COG1475	pfam02195	94.3
62	2640158509	Ga0099679_117254	SOS inhibition protein (PsiB)	144	-	pfam06290	96.5
63	2640158512	Ga0099679_117257	Antirestriction protein	166	COG4734	pfam07275	99.4
64	2640158513	Ga0099679_117258	hypothetical protein	144	-	-	100
65	2640158514	Ga0099679_117259	hypothetical protein	88	-	-	97.7
66	2640158516	Ga0099679_117261	hypothetical protein	60	-	-	79.4
67	2640158517	Ga0099679_117262	hypothetical protein	83	-	-	100
68	2640158519	Ga0099679_117264	hypothetical protein	111	-	-	100
69	2640158521	Ga0099679_117266	Relaxase/Mobilisation nuclease domain-containing protein	898	-	pfam03432	99.3
70	2640158522	Ga0099679_117267	intracellular multiplication protein IcmO	763	-	-	99.9
71	2640158523	Ga0099679_117268	TrbB protein	356	-	pfam13098	97.2
72	2640158524	Ga0099679_117269	intracellular multiplication protein IcmP	402	-	-	99.5
73	2640158527	Ga0099679_117272	hypothetical protein	83	-	-	100
74	2640158528	Ga0099679_117273	hypothetical protein	98	-	-	98
75	2640158529	Ga0099679_117274	hypothetical protein	58	-	-	96.6
76	2640158531	Ga0099679_117276	hypothetical protein	216	-	-	53.5
77	2640158533	Ga0099679_117278	hypothetical protein	194	-	-	99.3
78	2640158534	Ga0099679_117279	hypothetical protein	400	-	-	99.5
79	2640158535	Ga0099679_117280	hypothetical protein	204	-	-	98.6
80	2640158536	Ga0099679_117281	intracellular multiplication protein IcmB	1014	-	pfam12846	99.9
81	2640158537	Ga0099679_117282	hypothetical protein	266	-	-	97
82	2640158538	Ga0099679_117283	hypothetical protein	62	-	-	98.4
83	2640158539	Ga0099679_117284	hypothetical protein	134	-	-	95.5
84	2640158540	Ga0099679_117285	hypothetical protein	175	-	-	100
85	2640158541	Ga0099679_117286	intracellular multiplication protein IcmG	234	-	-	99.1
86	2640158542	Ga0099679_117287	intracellular multiplication protein IcmE	429	-	-	99.1
87	2640158543	Ga0099679_117288	intracellular multiplication protein IcmK	327	-	pfam12293	100
88	2640158544	Ga0099679_117289	intracellular multiplication protein IcmL	230	-	pfam11393	99.1
89	2640158545	Ga0099679_117290	TraL protein	115	-	-	100
90	2640158547	Ga0099679_117292	PLD-like domain-containing protein	183	-	pfam13091	98.4
91	2640158550	Ga0099679_117295	defect in organelle trafficking protein DotC	272	-	pfam16932	100

80

81 Supplementary Table 3 continued.

92	2640158551	Ga0099679_117296	defect in organelle trafficking protein DotD	152	-	pfam16816	100
93	2640158554	Ga0099679_117299	hypothetical protein	274	-	-	98.9
94	2640158555	Ga0099679_1172100	Site-specific recombinase XerC	384	COG4973	pfam00589	99.7
95	2640159987	Ga0099679_14131	Tn*3* transposase DDE domain-containing protein	74	-	pfam01526	63.2

82

83

84

85

86 **Supplementary Table 4. Antibiotic resistance genotype and**
87 **phenotype of CA S. sonnei isolates.**

							Susceptibility categories (% of S/I/R with given resistance determinant)		
Antimicrobial Class	**Resistance determinant**	**No of isolates with the determinant**	**Antibiotic**	**MIC Range**	**MIC50***	**MIC90***	**S, # of isolates (%)**	**I, # of isolates (%)**	**R, # of isolates (%)**
Aminoglycoside resistance	strA (both 100% and 99.8% ID; both FL and Δ) + strB (both 100% and 99.8% ID) + aadA1(both FL and Δ)	40	Streptomycin	N/A	N/A	N/A	0	0	40 (100%)
			Amikacin	≤4-8	≤4	≤4	40 (100%)	0	0
			Gentamicin	1-2	2	2	40 (100%)	0	0
			Tobramycin	1-2	2	2	40 (100%)	0	0
	strA + strB + aadA2	1	Streptomycin	N/A	N/A	N/A	0	0	1 (100%)
			Amikacin	8	N/A	N/A	1 (100%)	0	0
			Gentamicin	4	N/A	N/A	1 (100%)	0	0
			Tobramycin	2	N/A	N/A	1 (100%)	0	0
	strA (both 100% and 99.8% ID; both FL and Δ) + strB (both FL and Δ)	15	Streptomycin	N/A	N/A	N/A	0	0	14 (100%)
			Amikacin	4	4	4	15 (100%)	0	0
			Gentamicin	1-2	2	2	15 (100%)	0	0
			Tobramycin	1-2	2	2	15 (100%)	0	0
	strA + strB + aadA1 + aac(3)-IId (99.8%ID)	1	Streptomycin	N/A	N/A	N/A	0	0	1 (100%)
			Amikacin	4	N/A	N/A	1 (100%)	0	0
			Gentamicin	>8	N/A	N/A	0	0	1 (100%)
			Tobramycin	8	N/A	N/A	0	1 (100%)	0
	strB + aadA1	2	Streptomycin	N/A	N/A	N/A	0	0	2 (100%)
			Amikacin	4	4	4	2 (100%)	0	0
			Gentamicin	1-2	1	2	2 (100%)	0	0
			Tobramycin	1-2	1	2	2 (100%)	0	0
	aadA1	7	Streptomycin	N/A	N/A	N/A	0	0	7 (100%)
			Amikacin	4	4	4	7 (100%)	0	0
			Gentamicin	1-2	2	2	7 (100%)	0	0
			Tobramycin	1-2	1	2	7 (100%)	0	0
	no aminoglycoside resistance genes	2	Streptomycin	N/A	N/A	N/A	0	2 (100%)	0
			Amikacin	4	4	4	2 (100%)	0	0
			Gentamicin	2	2	2	2 (100%)	0	0
			Tobramycin	1-2	1	2	2 (100%)	0	0
Beta-lactam resistance	blaTEM-1	15	Ampicillin	>16	>16	>16	0	0	15 (100%)
			Ampicillin/Sulbactam	8/4- >16/8	16/8	>16/8	1 (6.67%)	6 (40%)	8 (53.33%)
			Aztreonam	4	4	4	15 (100%)	0	0
			Cefazolin	2-8	2	4	15 (100%)	0	0
			Cefepime	2	2	2	15 (100%)	0	0
			Cefotaxime	2	2	2	15 (100%)	0	0
			Cefotaxime/CA	0.5	0.5	0.5	15 (100%)	0	0
			Cefoxitin	8	8	8	15 (100%)	0	0
			Ceftazidime	1	1	1	15 (100%)	0	0
			Ceftazidime/CA	0.25	0.25	0.25	15 (100%)	0	0
			Cephalothin	8-16	8	16	10(66.67%)	5 (33.33%)	0
			Ertapenem	1	1	1	15 (100%)	0	0
			Imipenem	1	1	1	15 (100%)	0	0
			Meropenem	1	1	1	15 (100%)	0	0
			Piperacillin	>64	>64	>64	0	0	15 (100%)
			Piperacillin/ Tazobactam	8	8	8	15 (100%)	0	0
			Ticarcillin/K Clavulanate	8-32	8	32	12 (80%)	3 (20%)	0

88

89 Supplementary Table 4 continued.

	blaTEM-1 + blaOXA-2	1	Ampicillin	>16	N/A	N/A	0	0	1 (100%)
			Ampicillin/Sulbactam	16/8	N/A	N/A	0	1 (100%)	0
			Aztreonam	4	N/A	N/A	1 (100%)	0	0
			Cefazolin	2	N/A	N/A	1 (100%)	0	0
			Cefepime	2	N/A	N/A	1 (100%)	0	0
			Cefotaxime	2	N/A	N/A	1 (100%)	0	0
			Cefotaxime/CA	0.5	N/A	N/A	1 (100%)	0	0
			Cefoxitin	8	N/A	N/A	1 (100%)	0	0
			Ceftazidime	1	N/A	N/A	1 (100%)	0	0
			Ceftazidime/CA	0.25	N/A	N/A	1 (100%)	0	0
			Cephalothin	8	N/A	N/A	1 (100%)	0	0
			Ertapenem	1	N/A	N/A	1 (100%)	0	0
			Imipenem	1	N/A	N/A	1 (100%)	0	0
			Meropenem	1	N/A	N/A	1 (100%)	0	0
			Piperacillin	>64	N/A	N/A	0	0	1 (100%)
			Piperacillin/ Tazobactam	8	N/A	N/A	1 (100%)	0	0
			Ticarcillin/K Clavulanate	8	N/A	N/A	1 (100%)	0	0
	No beta-lactam resistance genes	52	Ampicillin	2-4	2	4	52 (100%)	0	0
			Ampicillin/ \Sulbactam	8/4	8/4	8/4	52 (100%)	0	0
			Aztreonam	4	4	4	52 (100%)	0	0
			Cefazolin	2	2	2	52 (100%)	0	0
			Cefepime	2	2	2	52 (100%)	0	0
			Cefotaxime	2	2	2	52 (100%)	0	0
			Cefotaxime/CA	0.5	0.5	0.5	52 (100%)	0	0
			Cefoxitin	8	8	8	52 (100%)	0	0
			Ceftazidime	1	1	1	52 (100%)	0	0
			Ceftazidime/CA	0.25	0.25	0.25	52 (100%)	0	0
			Cephalothin	8	8	8	52 (100%)	0	0
			Ertapenem	1	1	1	52 (100%)	0	0
			Imipenem	1	1	1	52 (100%)	0	0
			Meropenem	1	1	1	52 (100%)	0	0
			Piperacillin	16	16	16	52 (100%)	0	0
			Piperacillin/ Tazobactam	8	8	8	52 (100%)	0	0
			Ticarcillin/K Clavulanate	8	8	8	52 (100%)	0	0
Macrolide resistance	mph(A)	1	Azithromycin	N/A	N/A	N/A	0	0	1 (100%)
	No macrolide resistance genes	67	Azithromycin	N/A	N/A	N/A	67 (100%)	0	0
Phenicol resistance	catA1 (99.85% ID)	1	Chloramphenicol	>16	N/A	N/A	0	0	1 (100%)
	No phenicol resistance genes	67	Chloramphenicol	8	N/A	N/A	67 (100%)	0	0
Tetracycline resistance	tet(A) (both FL and Δ)	54	Tetracycline	>8	>8	>8	0	0	54 (100%)
	tet(B)	5	Tetracycline	>8	>8	>8	0	0	5 (100%)
	No tetracycline resistance genes	9	Tetracycline	4->8	4	>8	7 (77.77%)	0	2 (22.22%)
Trimethoprim/ Sulphonamide resistance	sul2 + dfrA8	1	Trimethoprim/ Sulfamethoxazole	>2/38	N/A	N/A	0	0	1 (100%)
	sul2 (both FL and Δ) + dfrA1	56	Trimethoprim/ Sulfamethoxazole	>2/38	>2/38	>2/38	0	0	56 (100%)
	sul2 + dfrA1 + dfrA12	1	Trimethoprim/ Sulfamethoxazole	>2/38	N/A	N/A	0	0	1 (100%)
	sul2 + sul1 (99.77% ID) + dfrA8	1	Trimethoprim/ Sulfamethoxazole	>2/38	N/A	N/A	0	0	1 (100%)

90

Supplementary Table 4 continued.

	sul2 + sul1 + dfrA12	1	Trimethoprim/ Sulfamethoxazole	>2/38	N/A	N/A	0	0	1 (100%)
	dfrA1 (both 100 and 99.79% ID)	6	Trimethoprim/ Sulfamethoxazole	>2/38	>2/38	>2/38	0	0	6 (100%)
	No trimethoprim or sulphonamide resistance genes	2	Trimethoprim/ Sulfamethoxazole	≤2/38	N/A	N/A	2(100%)	0	0
Fluoroquinolone resistance	gyrA S83L + parC [R275C]▲	1	Ciprofloxacin	0.5	N/A	N/A	1 (100%)	0	0
			Levofloxacin	1	N/A	N/A	1 (100%)	0	0
	gyrA S83L	2	Ciprofloxacin	0.5	0.5	0.5	2(100%)	0	0
			Levofloxacin	1	1	1	2 (100%)	0	0
	gyrA S83L + gyrA D87G + parC S80I	13	Ciprofloxacin	>2	>2	>2	0	0	13 (100%)
			Levofloxacin	1-4	4	4	13 (100%)	0	0
	gyrA [D483E] + parC [V533A]	1	Ciprofloxacin	0.5	N/A	N/A	1 (100%)	0	0
			Levofloxacin	1	N/A	N/A	1 (100%)	0	0
	parC [V533A]	2	Ciprofloxacin	0.5	N/A	N/A	2 (100%)	0	0
			Levofloxacin	1	N/A	N/A	2 (100%)	0	0
	No mutations in target genes	49	Ciprofloxacin	0.5	0.5	0.5	49 (100%)	0	0
			Levofloxacin	1	1	1	49 (100%)	0	0

Footnotes and abbreviations:

*MIC50 and MIC90 values calculated as the lowest concentration of the antibiotic at which the growth of respectively 50% and 90% of the isolates is inhibited.

▲- mutations which weren't previously described as associated with a quinolone resistance are marked with square brackets

FL- Full-length

Δ- truncated

Figure S1. Clustering of CA *S. sonnei* isolates with *S. sonnei* strains of global lineages as per Holt et al. based on a maximum-likelihood phylogenetic tree built using genome-wide hqSNPs.

Color of node labels: Red- S.sonnei isolates from CA; Blue- global isolates from Holt et al. publication. **Branches highlight color:** Yellow- Lineage I; Blue- Lineage II; Green- Lineage III; Purple- Lineage IV. Tree is rooted to E. coli.

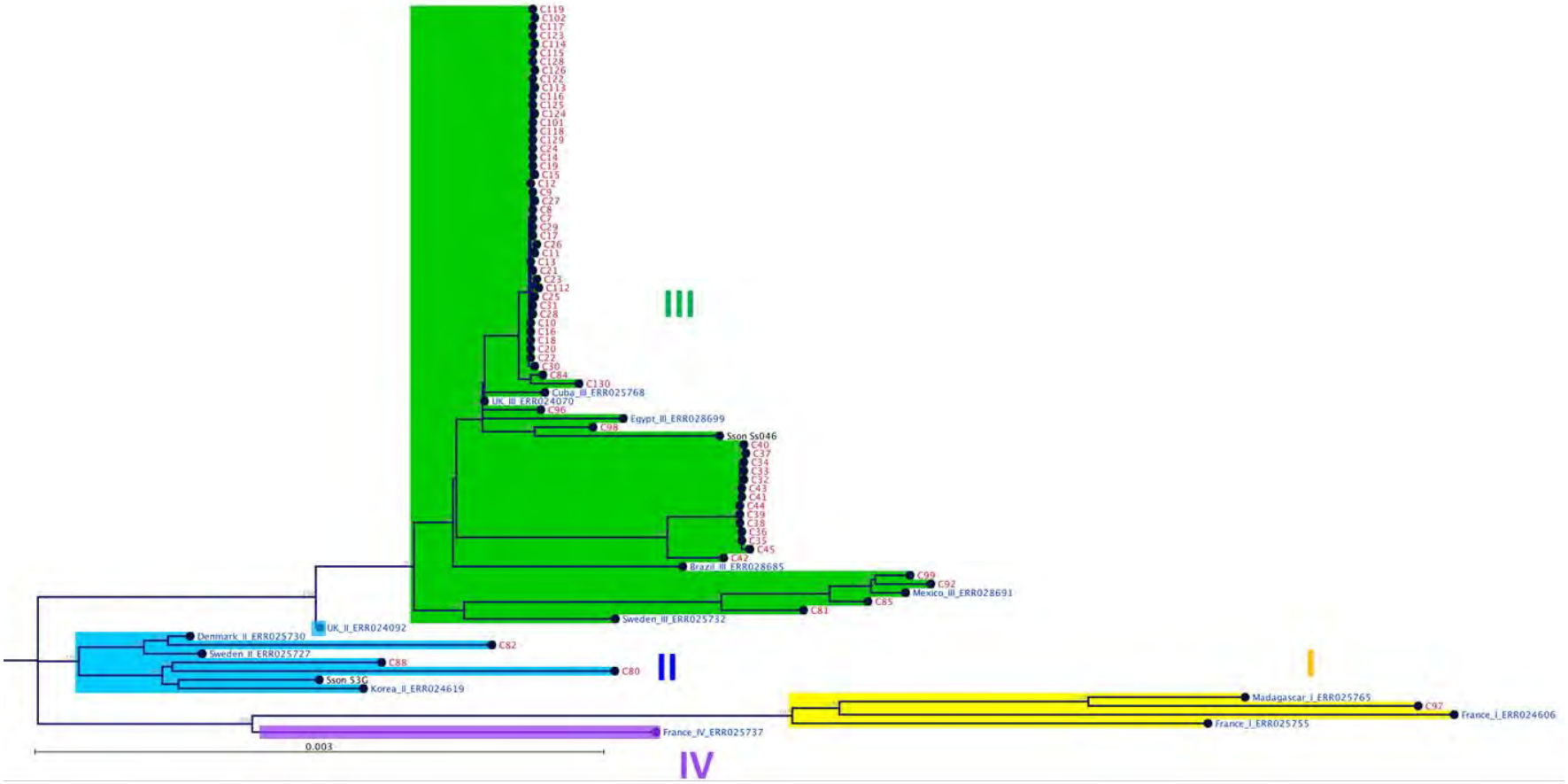
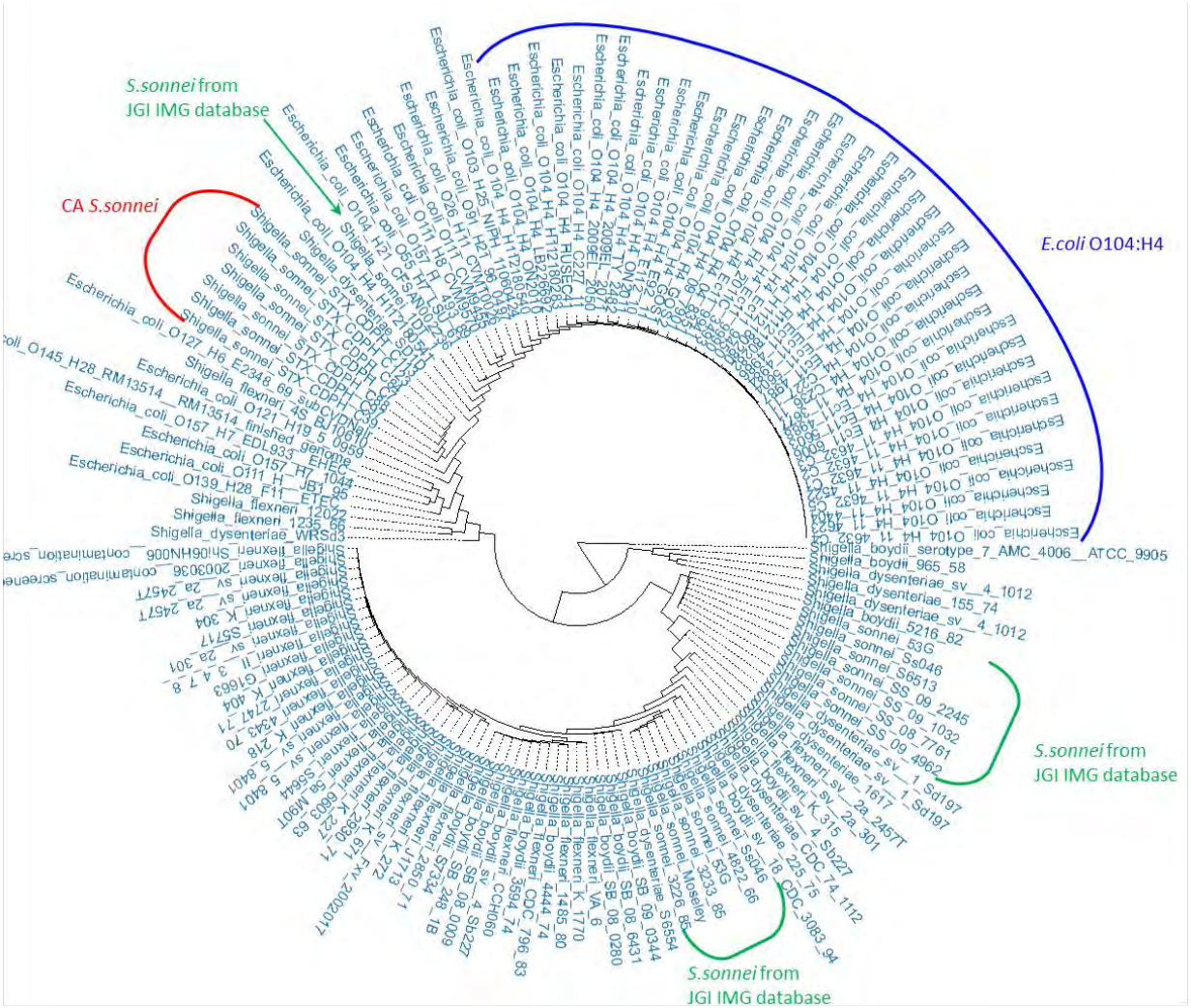


Figure S2. Hierarchical Clustering of CA representative *S. sonnei* isolates with *E.coli* and other *Shigella* species from JGI IMG database.

A. Clustering based on COG profiles (presence / absence & abundance).

Color of the brackets: Red- representative *S.sonnei* from California; Green- other *S. sonnei* from JGI IMG database; Blue- *E.coli* O104:H4 strains from JGI IMG database



B. Clustering based on pfam profiles (presence / absence & abundance).

Color of the brackets: Red- representative *S.sonnei* from California; Green- other *S. sonnei* from JGI IMG database; Blue- *E.coli* O104:H4 strains from JGI IMG database.

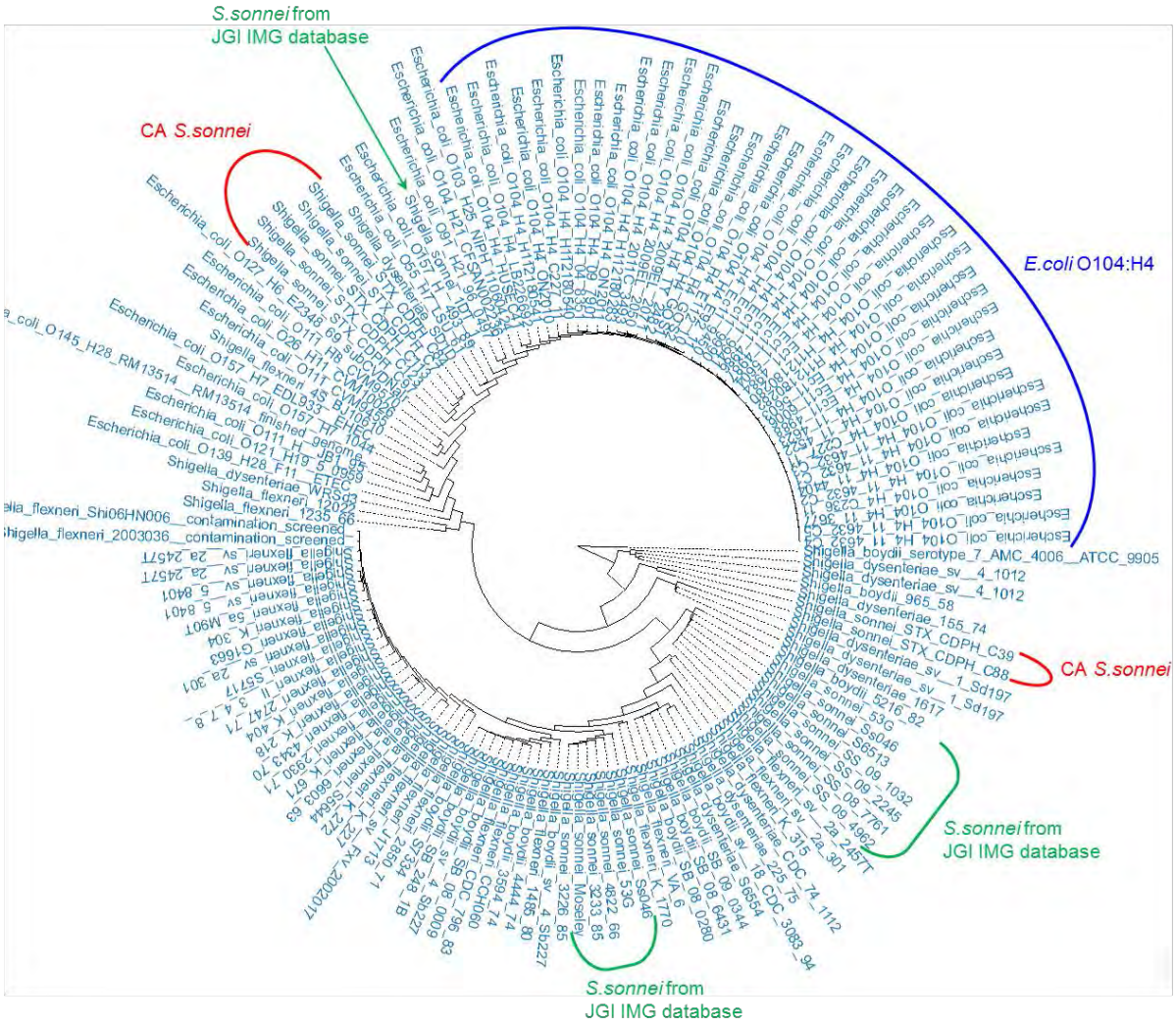
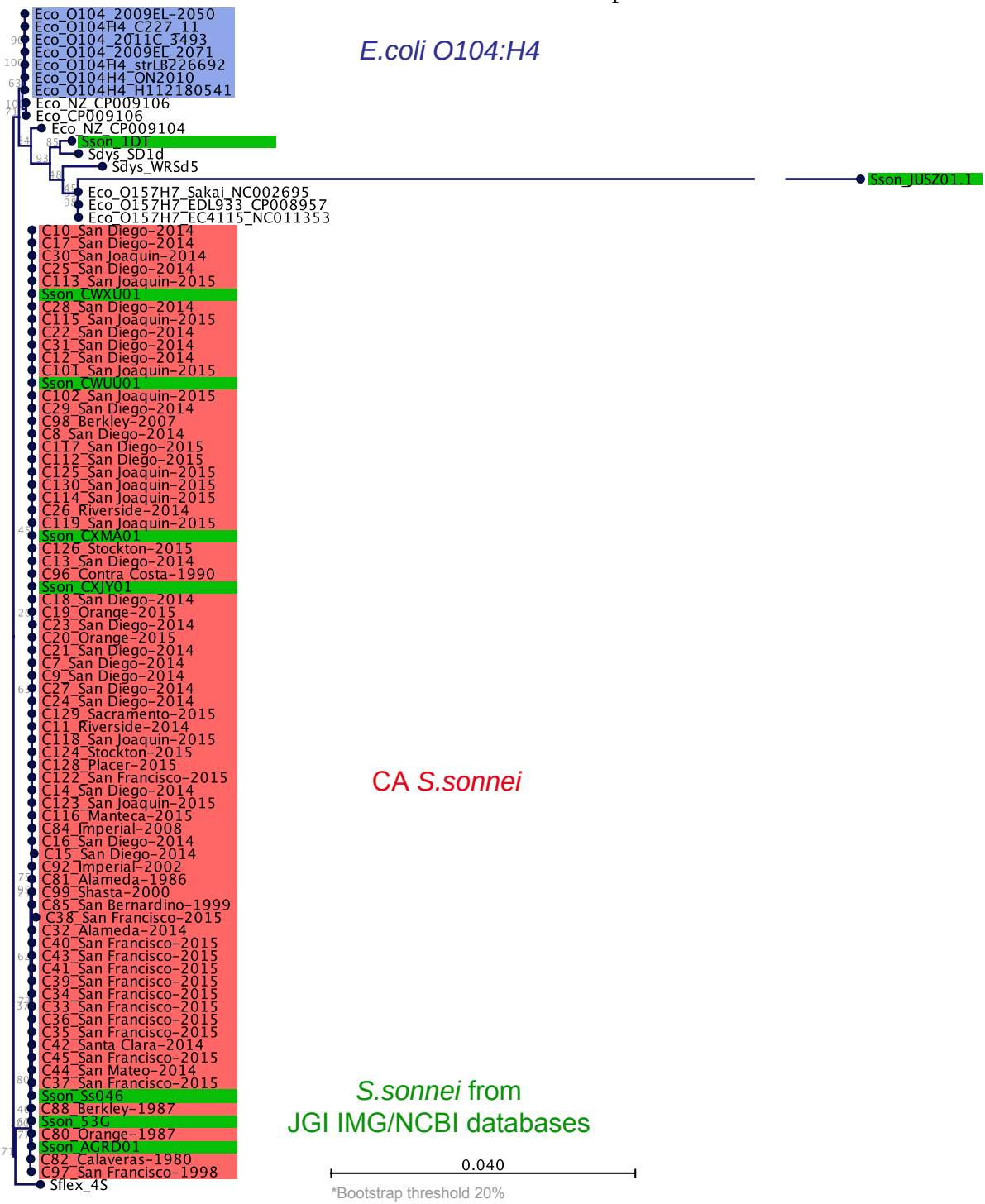


Figure S3. Comparison of CA *S. sonnei* with *E. coli* strains and other publicly-available genomes of *S. sonnei* based on nucleotide sequence.

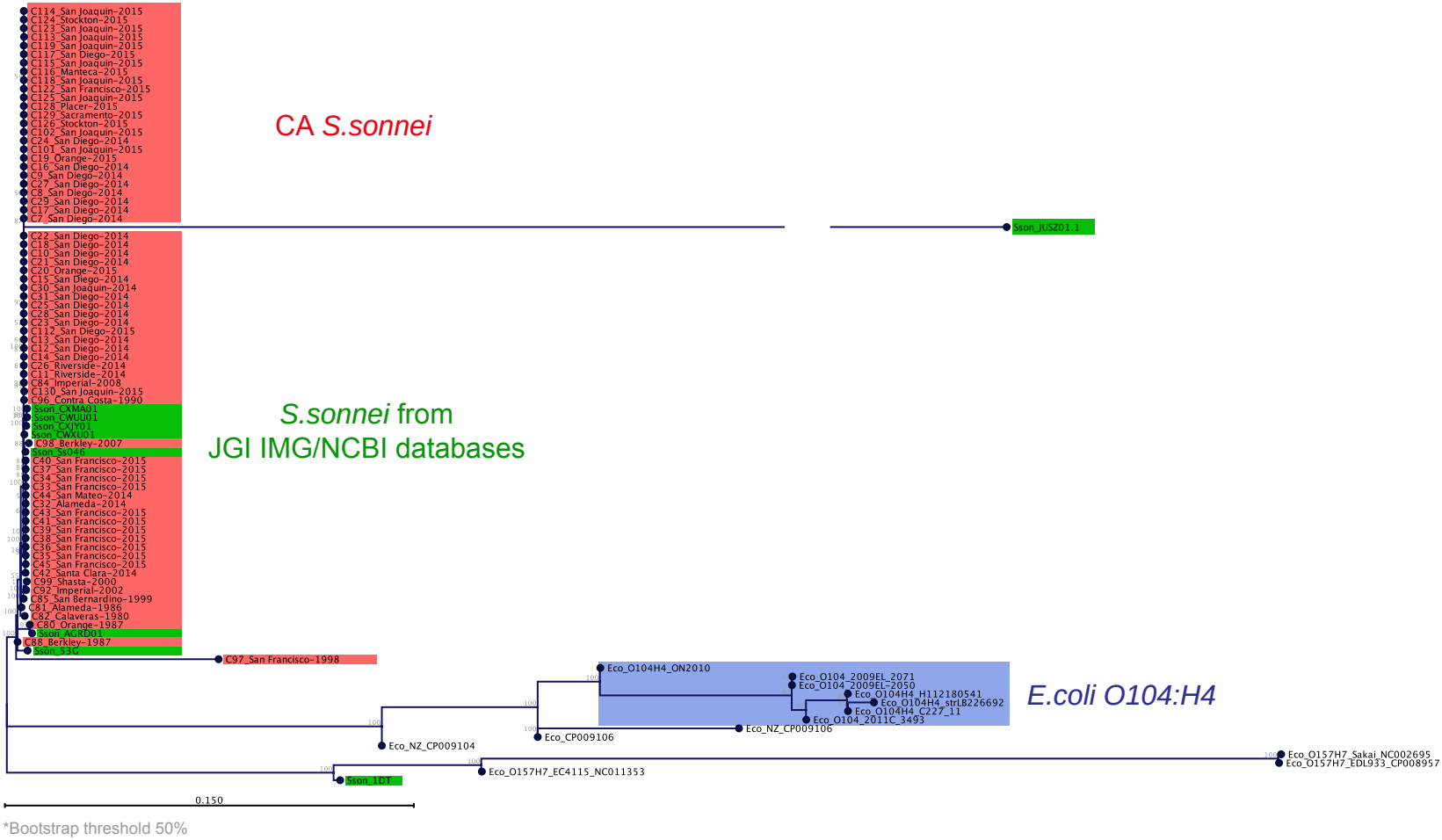
A. Maximum likelihood clustering from PhyloSift nucleotide-based phylogeny.

Background color: Red- *S.sonnei* from California; Green- other *S.sonnei* from JGI IMG and NCBI databases; Blue- *E.coli* O104:H4 strains from JGI IMG database.Bootstrap threshold 20%.



B. Maximum likelihood phylogeny of CA *S. sonnei*, *E.coli*, and publicly-available *S. sonnei* genomes based on genome-wide hqSNPs.

Background color: Red- *S.sonnei* from California; Green- other *S. sonnei* from JGI IMG and NCBI databases; Blue- *E.coli* O104:H4 strains from JGI IMG database. Bootstrap threshold 50%.



*Bootstrap threshold 50%

Figure S4. Comparison of COG abundance profiles of representative CA *S. sonnei* with *E. coli* strains and other *Shigella* species from JGI IMG database

The heat map represents gene count for different COGs.

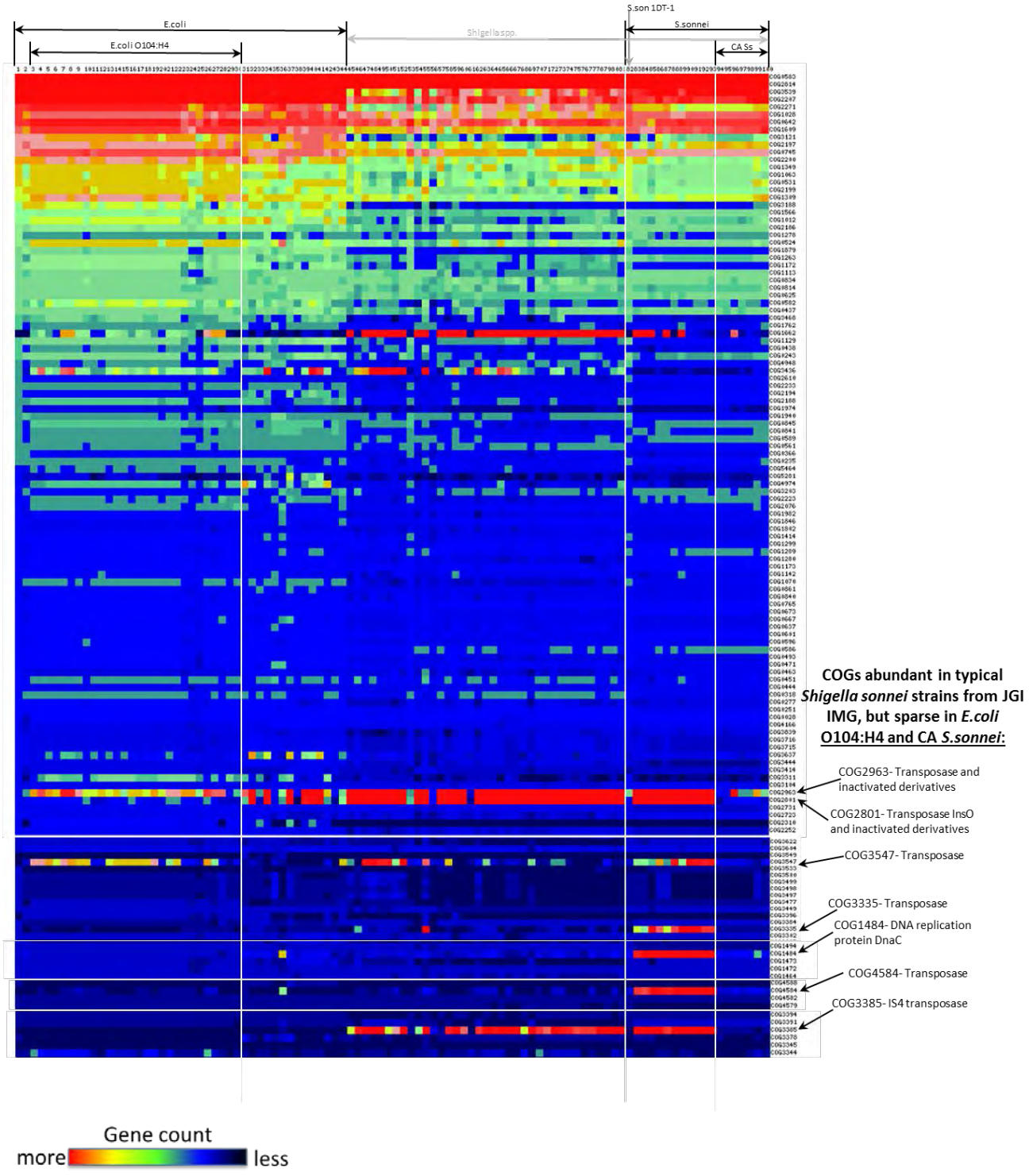


Figure S4 continued.

Genomes included into the COG abundance analysis:

1 - Escherichia coli O103:H25 NIPH-11060424
2 - Escherichia coli O104:H21 CFSAN002236
3 - Escherichia coli O104:H4 04-8351
4 - Escherichia coli O104:H4 09-7901
5 - Escherichia coli O104:H4 11-4404
6 - Escherichia coli O104:H4 11-4632 C3
7 - Escherichia coli O104:H4 2009EL-2050
8 - Escherichia coli O104:H4 2009EL-2071
9 - Escherichia coli O104:H4 2011C-3493
10 - Escherichia coli O104:H4 C227-11
11 - Escherichia coli O104:H4 E112/10
12 - Escherichia coli O104:H4 E92/11
13 - Escherichia coli O104:H4 Ec11-4986
14 - Escherichia coli O104:H4 Ec11-4988
15 - Escherichia coli O104:H4 Ec11-5603
16 - Escherichia coli O104:H4 Ec11-5604
17 - Escherichia coli O104:H4 Ec11-9450
18 - Escherichia coli O104:H4 Ec11-9941
19 - Escherichia coli O104:H4 Ec11-9990
20 - Escherichia coli O104:H4 GOS1
21 - Escherichia coli O104:H4 H112180280
22 - Escherichia coli O104:H4 H112180282
23 - Escherichia coli O104:H4 H112180283
24 - Escherichia coli O104:H4 H112180540
25 - Escherichia coli O104:H4 H112180541
26 - Escherichia coli O104:H4 HUSEC41
27 - Escherichia coli O104:H4 LB226692
28 - Escherichia coli O104:H4 ON2010
29 - Escherichia coli O104:H4 ON2011
30 - Escherichia coli O104:H4 TY-2482
31 - Escherichia coli O111 CVM9455
32 - Escherichia coli O111:H- JB1-95
33 - Escherichia coli O111:H8 CVM9570
34 - Escherichia coli O121:H19 5.0959
35 - Escherichia coli O127:H6 E2348/69 subCVDNalr
36 - Escherichia coli O139:H28 F11 (ETEC)
37 - Escherichia coli O145:H28 RM13514 (RM13514 finished genome)
38 - Escherichia coli O157:H- 493-89
39 - Escherichia coli O157:H7 1044
40 - Escherichia coli O157:H7 EDL933 (EHEC)
41 - Escherichia coli O157:H7 Sakai (EHEC)
42 - Escherichia coli O26:H11 CVM10026
43 - Escherichia coli O55:H7 LSU-61
44 - Escherichia coli O91:H21 96.0497
45 - Shigella boydii 3594-74
46 - Shigella boydii 5216-82
47 - Shigella boydii SB_08-0009
48 - Shigella boydii SB_08-6431
49 - Shigella boydii SB_248-1B
50 - Shigella boydii sv. 4 Sb227
51 - Shigella dysenteriae CDC 74-1112
52 - Shigella dysenteriae S6554
53 - Shigella dysenteriae SD1D
54 - Shigella dysenteriae sv. 1 Sd197
55 - Shigella dysenteriae sv. 4 1012
56 - Shigella dysenteriae WRSd3
57 - Shigella flexneri 12022
58 - Shigella flexneri 1485-80
59 - Shigella flexneri 2850-71

Figure S4 continued.

60 - Shigella flexneri 2930-71
61 - Shigella flexneri 4S BJ10610
62 - Shigella flexneri CDC 796-83
63 - Shigella flexneri G1663
64 - Shigella flexneri II:(3)4,7(8)
65 - Shigella flexneri K-1770
66 - Shigella flexneri K-227
67 - Shigella flexneri K-272
68 - Shigella flexneri K-304
69 - Shigella flexneri K-315
70 - Shigella flexneri K-404
71 - Shigella flexneri K-671
72 - Shigella flexneri S5644
73 - Shigella flexneri S5717
74 - Shigella flexneri Shi06HN006
75 - Shigella flexneri sv. 2a 2457T
76 - Shigella flexneri sv. 2a 2457T
77 - Shigella flexneri sv. 2a 301
78 - Shigella flexneri sv. 5 8401
79 - Shigella flexneri sv. 5 8401
80 - Shigella flexneri sv. Fxv 2002017
81 - Shigella flexneri VA-6
82 - Shigella sonnei 1DT-1-
83 - Shigella sonnei 3226-85
84 - Shigella sonnei 3233-85
85 - Shigella sonnei 4822-66
86 - Shigella sonnei 53G
87 - Shigella sonnei Moseley
88 - Shigella sonnei S6513
89 - Shigella sonnei Ss046
90 - Shigella sonnei Ss046
91 - Shigella sonnei SS_08-7761
92 - Shigella sonnei SS_09-1032
93 - Shigella sonnei SS_09-2245
94 - Shigella sonnei STX_CDPH_C113
95 - Shigella sonnei STX_CDPH_C123
96 - Shigella sonnei STX_CDPH_C39
97 - Shigella sonnei STX_CDPH_C7
98 - Shigella sonnei STX_CDPH_C84
99 - Shigella sonnei STX_CDPH_C88
100 - Shigella sonnei STX_CDPH_C97

Figure S5. Revised pfam phylogeny. Maximum Likelihood clustering of CA representative *S.sonnei* isolates with *E.coli* and other *Shigella* species from IMG JGI database based on pfam profiles (presence/absence).

Background color: Red- representative *S.sonnei* from California; Green- other *S. sonnei* from JGI IMG database; Blue- *E.coli* O104:H4 strains from JGI IMG database.

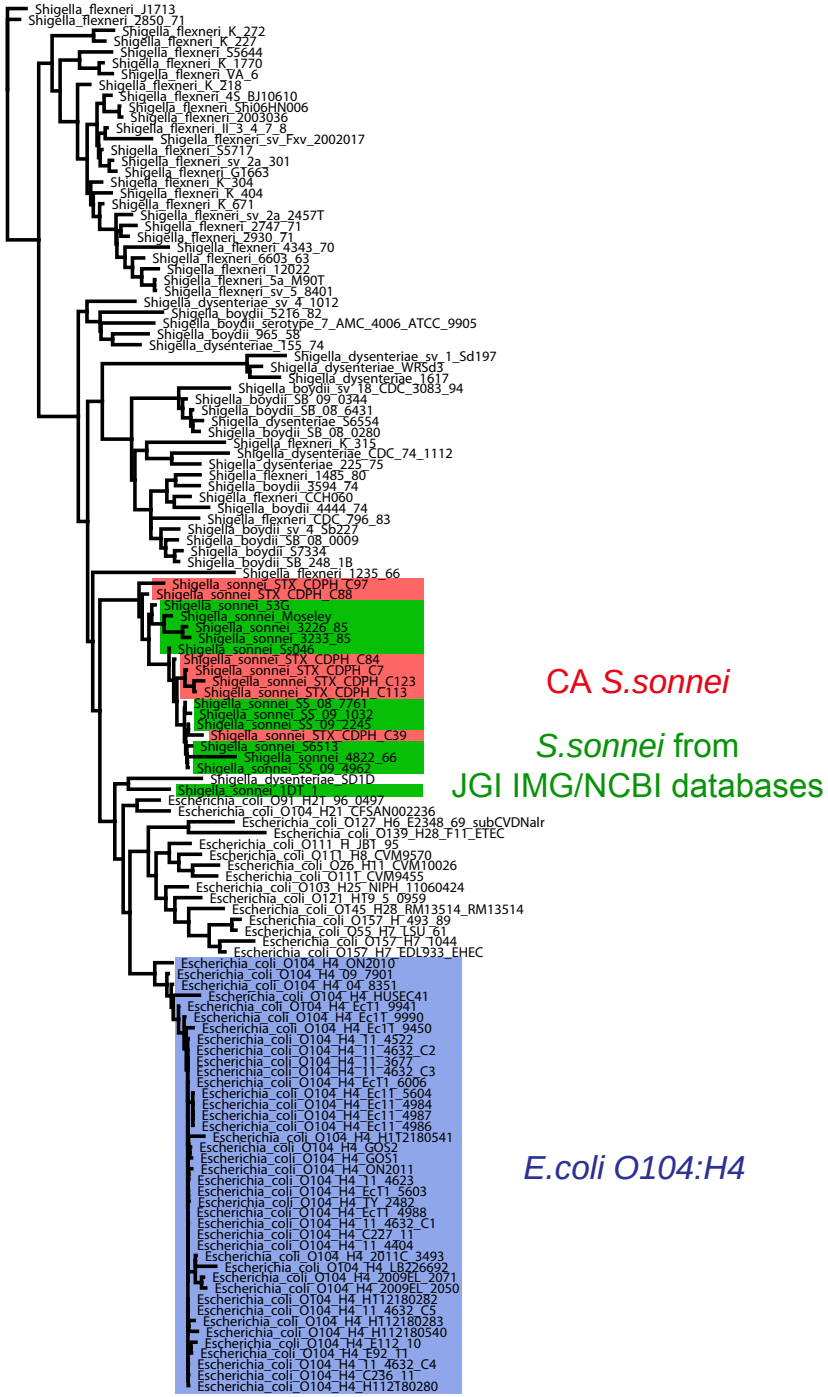


Figure S6. Neighbor-Joining phylogeny of the CA STX1-phage based on amino acid sequence of Integrase protein.

Subtree containing CA STX1-phage is highlighted with red branch color.

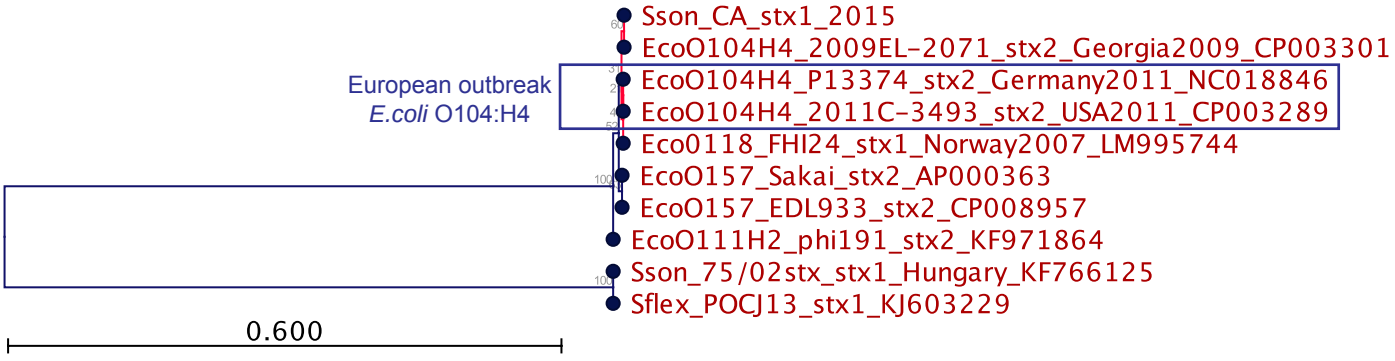


Figure S7. Distribution of CA STX1-bacteriophage genes in different *E.coli* and *Shigella* serovars from IMG database

Gene IDs from IMG database are listed on the left side of the graph. Color of the blocks: red - gene is present, blue - gene is absent. Min 10% ID was used as a similarity cutoff.

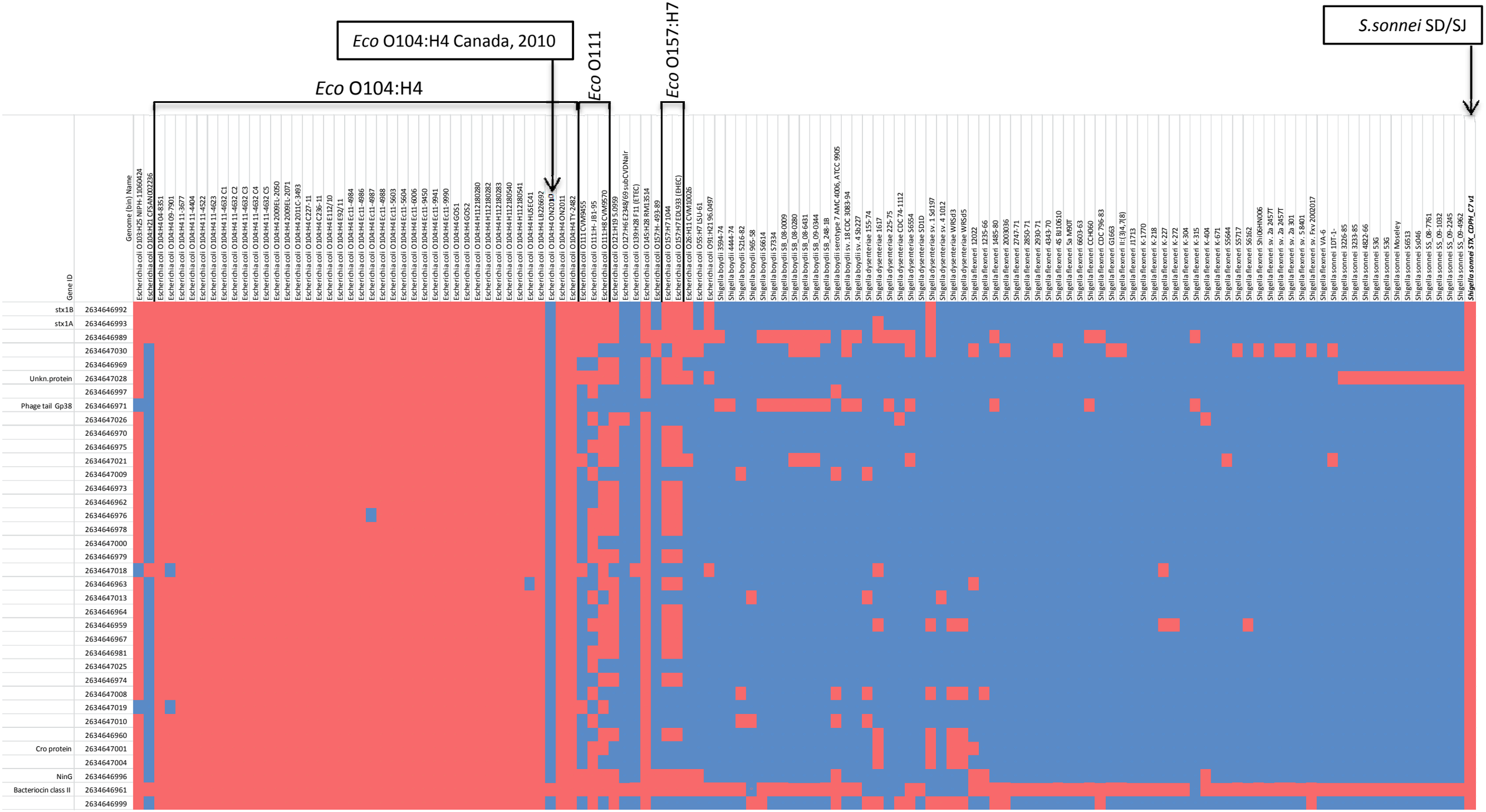


Figure S8. Cryptic STX-converting prophages found in CA *S.sonnei* genomes

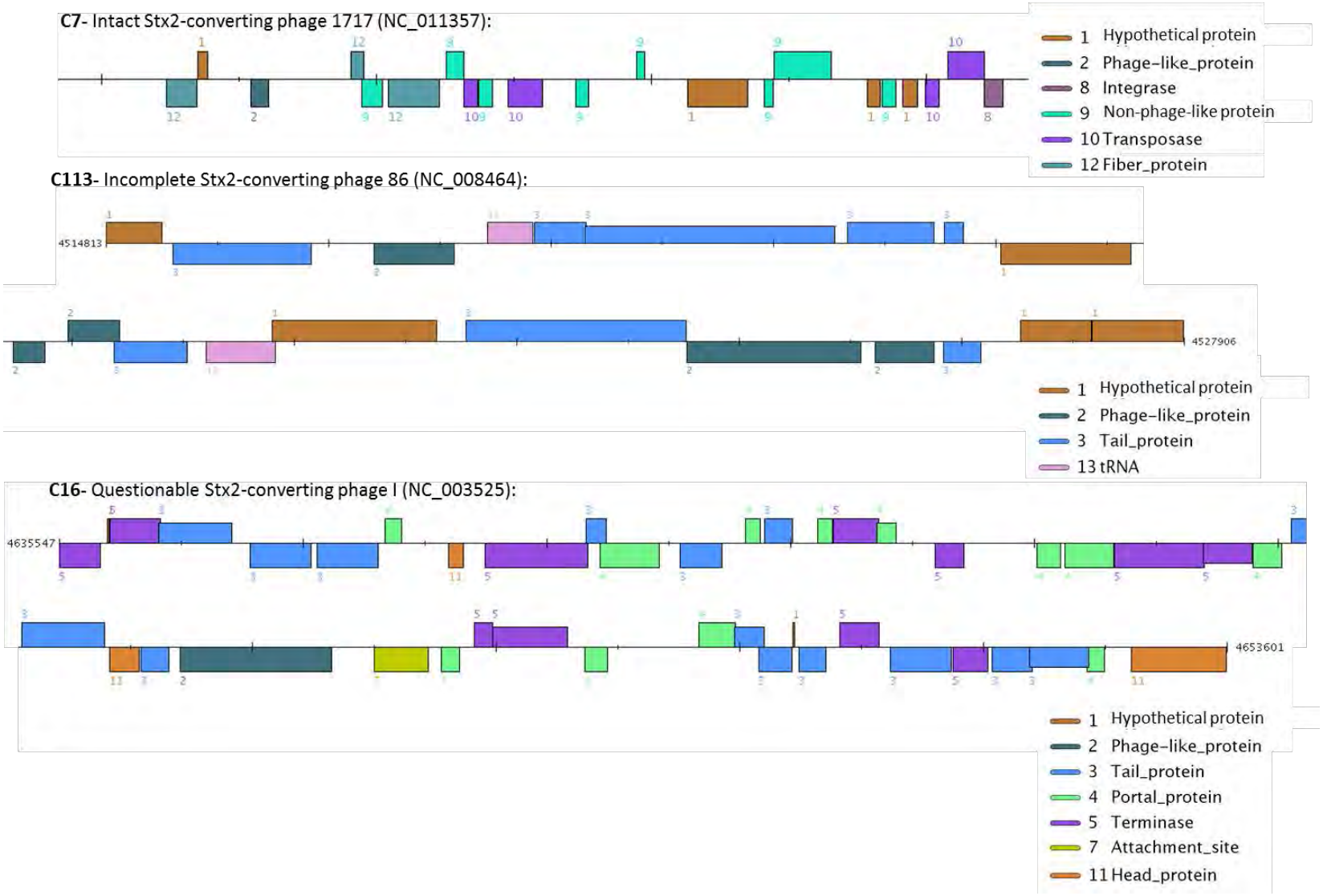


Figure S9. Plasmids and other mobile elements encoding antibiotic resistance in CA *S. sonnei*

A. Organization of bla_{TEM-1}- encoding IncB/O/K/Z conjugative plasmid from modern SF population isolates

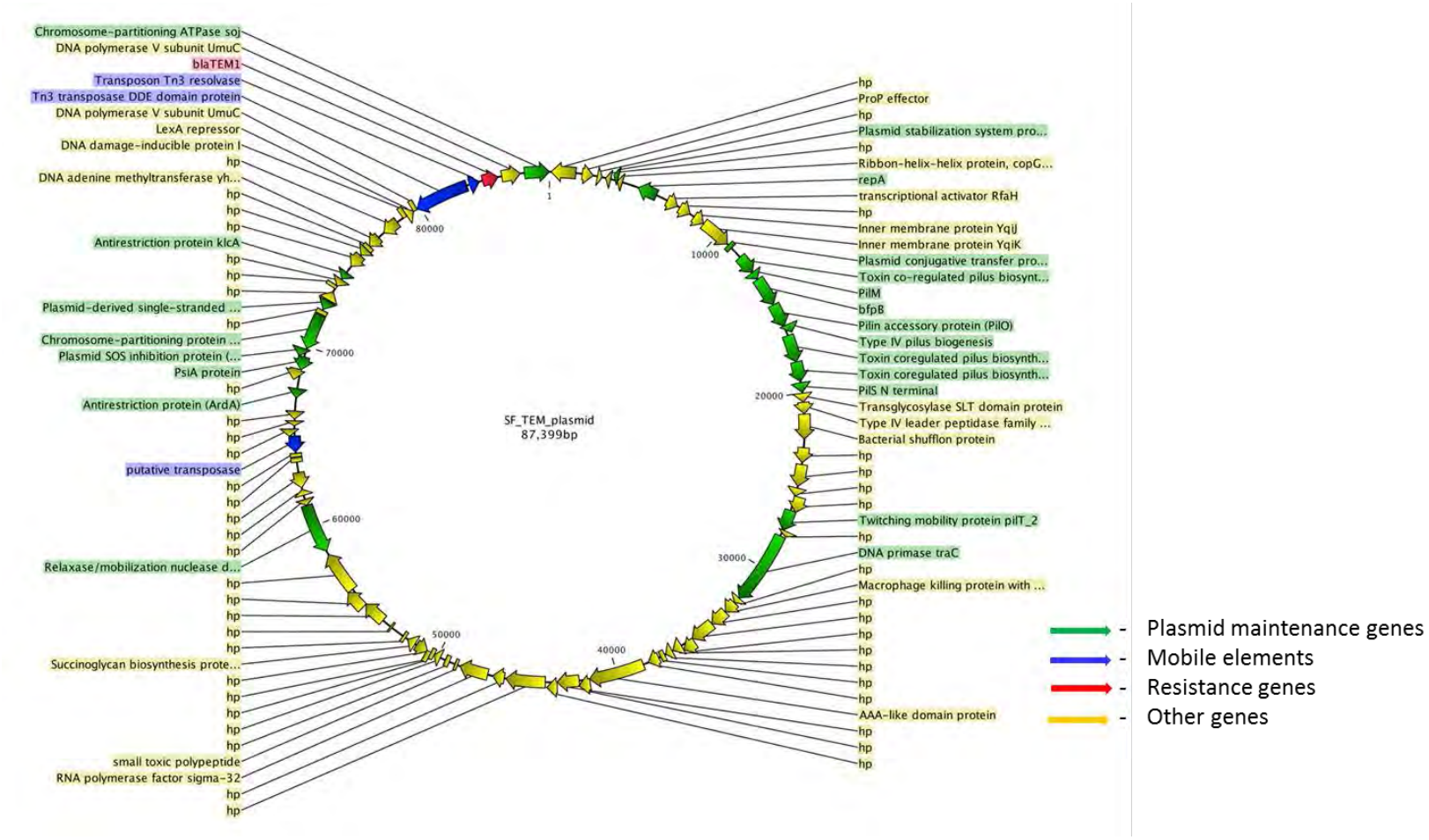


Figure S9 continued

Different integration sites of Tn3::*bla*TEM-1 on IncB/O/K/Z plasmids found in recent SF isolates and in historical *S. sonnei*:

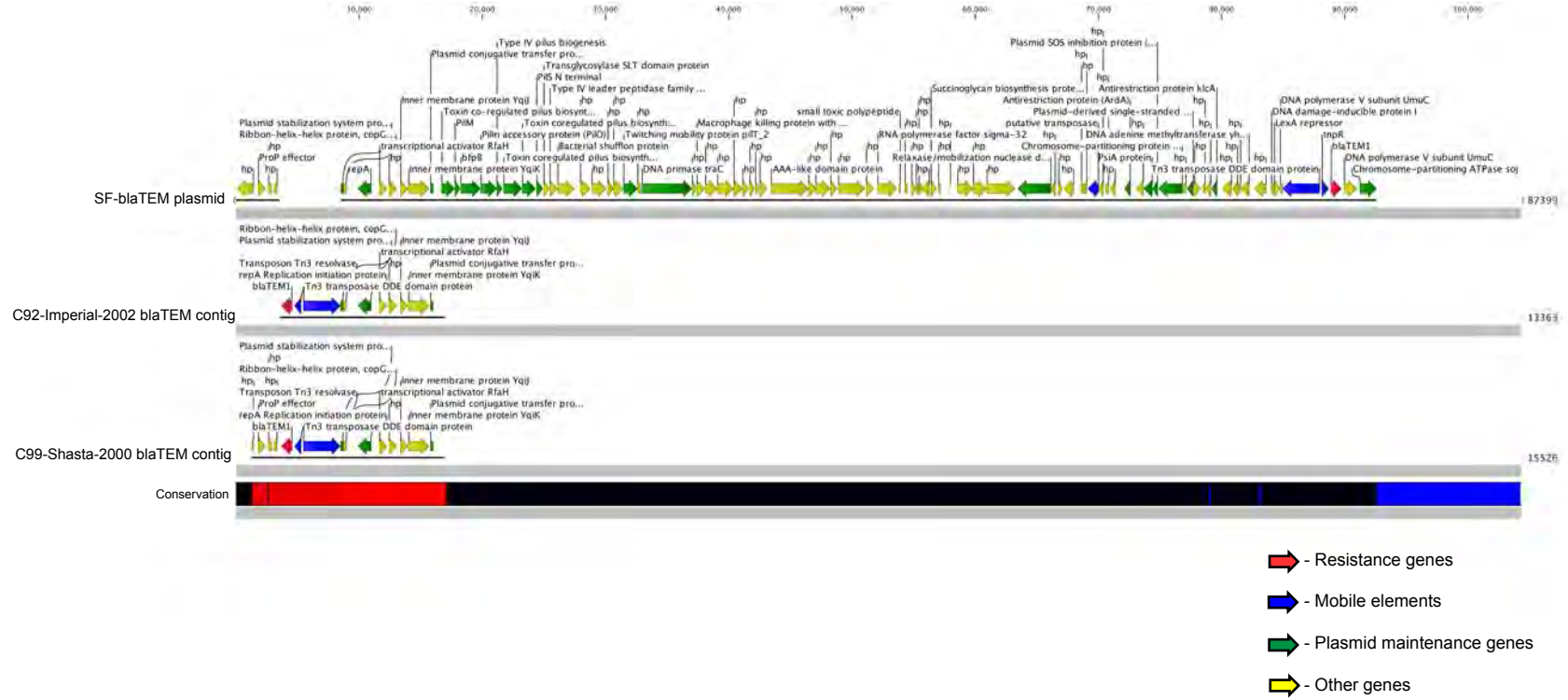


Figure S9 continued

B. A representative genetic surrounding of *bla*$_{TEM-1}$ gene on a putative IncI1 conjugative plasmid from a modern SJ *S. sonnei* isolate

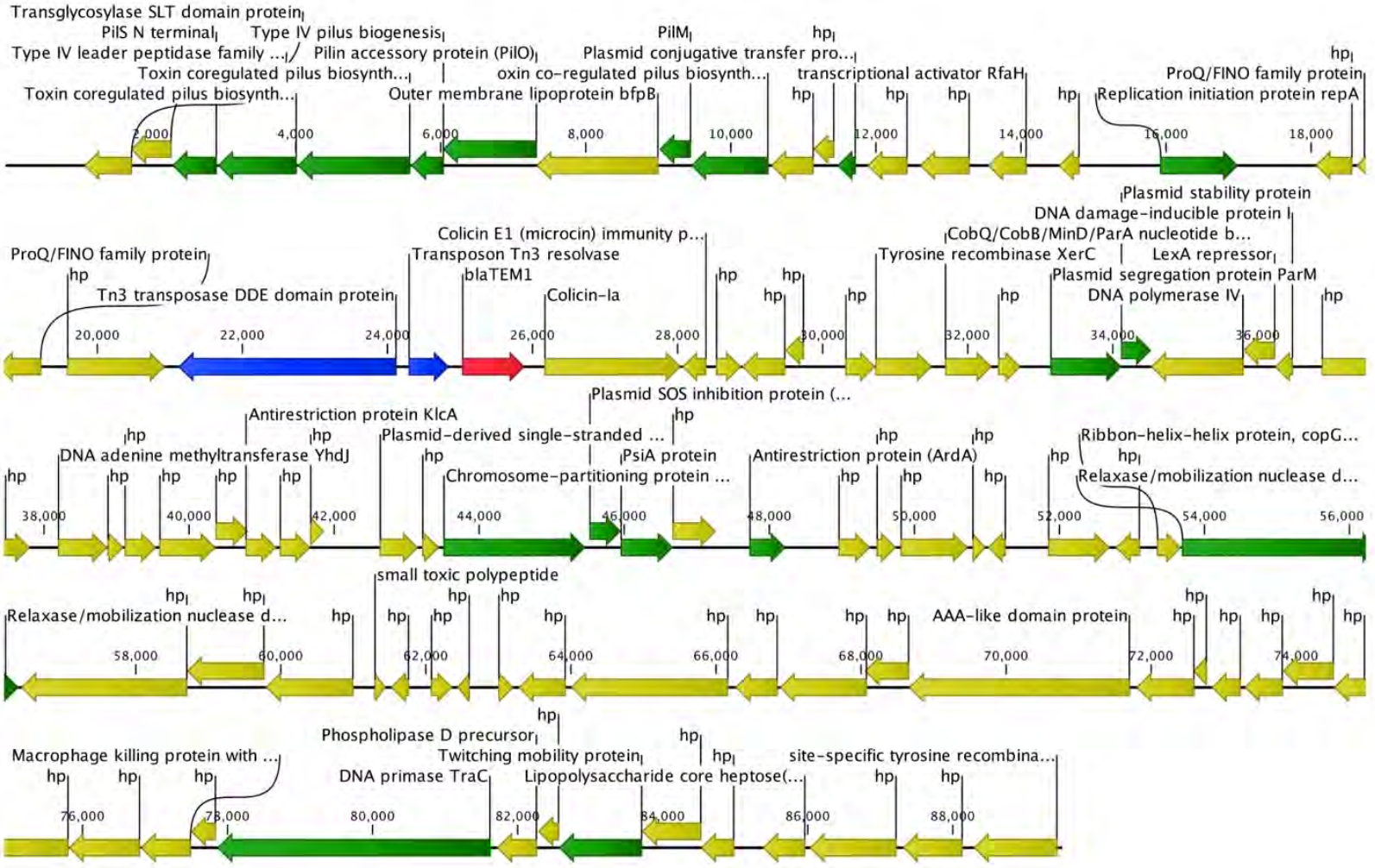


Figure S9 continued

C. Representative organization of *sul2-strA-strB-tetA* resistance genes cluster in SD / SJ and historical CA *S. sonnei* isolates