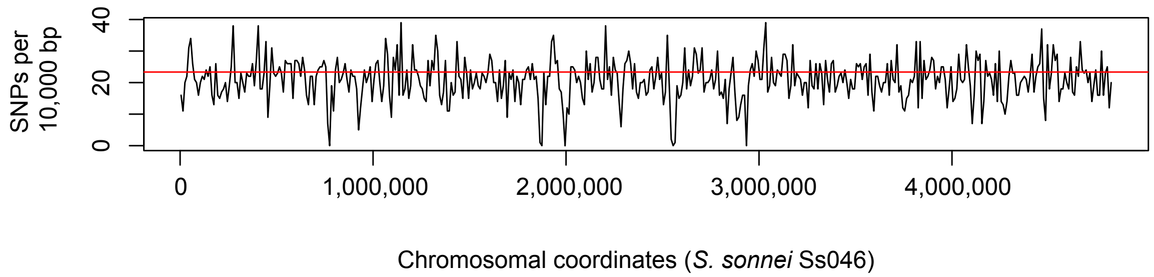
Supplementary Information

Shigella sonnei genome sequencing and phylogenetic analysis indicate recent global dissemination from Europe

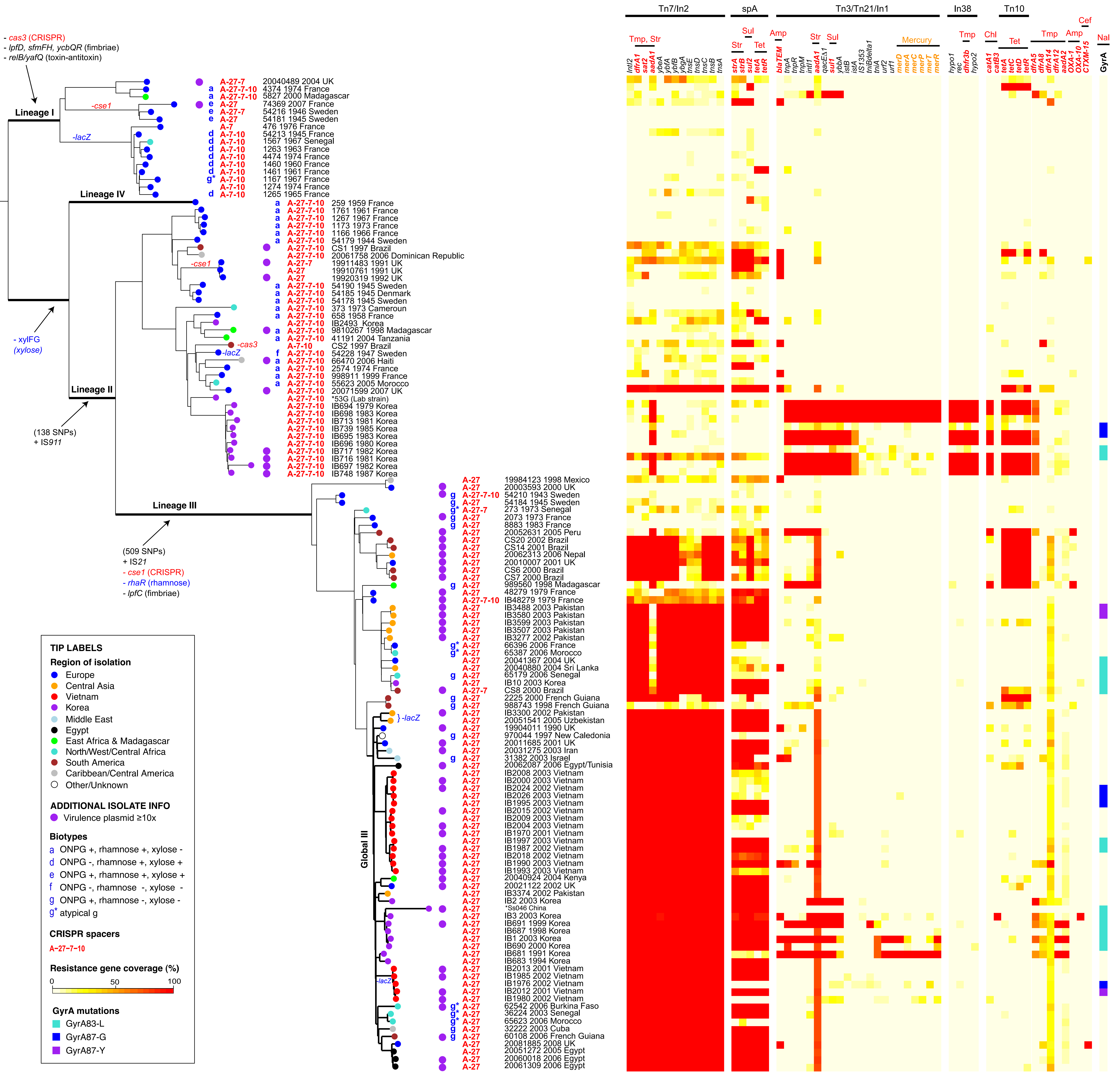
Kathryn E. Holt, Stephen Baker, François-Xavier Weill, Edward C. Holmes, Andrew Kitchen, Jun Yu, Vartul Sangal, Derek J. Brown, John E. Coia, Dong Wook Kim, Seon Young Choi, Su Hee Kim, Wanderley D. da Silveira, Derek J. Pickard, Jeremy J. Farrar, Julian Parkhill, Gordon Dougan, Nicholas R. Thomson



Supplementary Figure 1

Distribution of SNPs identified compared to the *S. sonnei* Ss046 reference genome.

Y-axis corresponds to SNP counts per 10,000 bp window; red line indicates the average rate of 23 SNPs per 10,000 bp (or 1 SNP per 430 bp).



Tn7/In2
spA
Tn3/Tn21/In1
In38
Tn10
Tmp, Str
Sul
Str
Tet
Amp
Str
Sul
Mercury
Tmp
Chl
Tet
Tmp
Amp
Cef
Nal
intI2
dfrA1
sat2
aadA1
ybeA
ybfA
ybfB
ybgA
tnsE
tnsD
tnsC
tnsB
tnsA
strA
strB
sul2
tetA
tetR
blaTEM
tnpA
tnpR
tnpM
intI1
aadA1
qacEΔ1
sul1
ybhA
istB
istA
IS1353
tniBdelta1
tniA
urf2
urf1
merD
merA
merC
merP
merT
merR
hypo1
rec
dhfr3b
hypo2
catA1
catB3
tetA
tetC
tetD
tetR
dfrA5
dfrA8
dfrA14
dfrA12
aadA2
OXA-1
OXA-10
CTXM-15
GyrA
- cas3 (CRISPR)
- lpfD, sfmFH, ycbQR (fimbriae)
- relB/yafQ (toxin-antitoxin)
Lineage I
-cse1
-lacZ
Lineage IV
- xylFG
(xylose)
Lineage II
(138 SNPs)
+ IS911
Lineage III
(509 SNPs)
+ IS21
- cse1 (CRISPR)
- rhaR (rhamnose)
- lpfC (fimbriae)
-cse1
-cas3
-lacZ
}-lacZ
-lacZ
Global III
A-27-7 20040489 2004 UK
a A-27-7-10 4374 1974 France
a A-27-7-10 5827 2000 Madagascar
e A-27 74369 2007 France
e A-27-7 54216 1946 Sweden
e A-27 54181 1945 Sweden
A-7 476 1976 France
d A-7-10 54213 1945 France
d A-7-10 1567 1967 Senegal
d A-7-10 1263 1963 France
d A-7-10 4474 1974 France
d A-7-10 1460 1960 France
d A-7-10 1461 1961 France
g* A-7-10 1167 1967 France
A-7-10 1274 1974 France
d A-7-10 1265 1965 France
a A-27-7-10 259 1959 France
a A-27-7-10 1761 1961 France
a A-27-7-10 1267 1967 France
a A-27-7-10 1173 1973 France
a A-27-7-10 1166 1966 France
a A-27-7-10 54179 1944 Sweden
A-27-7-10 CS1 1997 Brazil
A-27-7-10 20061758 2006 Dominican Republic
A-27-7 19911483 1991 UK
A-27 19910761 1991 UK
A-27 19920319 1992 UK
a A-27-7-10 54190 1945 Sweden
a A-27-7-10 54185 1945 Denmark
a A-27-7-10 54178 1945 Sweden
a A-27-7-10 373 1973 Cameroun
a A-27-7-10 658 1958 France
A-27-7-10 IB2493 Korea
a A-27-7-10 9810267 1998 Madagascar
a A-27-7-10 41191 2004 Tanzania
A-7-10 CS2 1997 Brazil
f A-27-7-10 54228 1947 Sweden
a A-27-7-10 66470 2006 Haiti
a A-27-7-10 2574 1974 France
a A-27-7-10 998911 1999 France
a A-27-7-10 55623 2005 Morocco
A-27-7-10 20071599 2007 UK
A-27-7-10 *53G (Lab strain)
A-27-7-10 IB694 1979 Korea
A-27-7-10 IB698 1983 Korea
A-27-7-10 IB713 1981 Korea
A-27-7-10 IB739 1985 Korea
A-27-7-10 IB695 1983 Korea
A-27-7-10 IB696 1980 Korea
A-27-7-10 IB717 1982 Korea
A-27-7-10 IB716 1981 Korea
A-27-7-10 IB697 1982 Korea
A-27-7-10 IB748 1987 Korea
A-27 19984123 1998 Mexico
A-27 20003593 2000 UK
g A-27-7-10 54210 1943 Sweden
g A-27 54184 1945 Sweden
g* A-27-7 273 1973 Senegal
g A-27 2073 1973 France
g A-27 8883 1983 France
A-27 20052631 2005 Peru
A-27 CS20 2002 Brazil
A-27 CS14 2001 Brazil
A-27 20062313 2006 Nepal
A-27 20010007 2001 UK
A-27 CS6 2000 Brazil
A-27 CS7 2000 Brazil
g A-27 989560 1998 Madagascar
A-27 48279 1979 France
A-27-7-10 IB48279 1979 France
A-27 IB3488 2003 Pakistan
A-27 IB3580 2003 Pakistan
A-27 IB3599 2003 Pakistan
A-27 IB3507 2003 Pakistan
A-27 IB3277 2002 Pakistan
g* A-27 66396 2006 France
g* A-27 65387 2006 Morocco
A-27 20041367 2004 UK
A-27 20040880 2004 Sri Lanka
g A-27 65179 2006 Senegal
A-27 IB10 2003 Korea
A-27-7 CS8 2000 Brazil
g A-27 2225 2000 French Guiana
g A-27 988743 1998 French Guiana
A-27 IB3300 2002 Pakistan
A-27 20051541 2005 Uzbekistan
A-27 19904011 1990 UK
g A-27 970044 1997 New Caledonia
A-27 20011685 2001 UK
A-27 20031275 2003 Iran
g A-27 31382 2003 Israel
A-27 20062087 2006 Egypt/Tunisia
A-27 IB2008 2003 Vietnam
A-27 IB2000 2003 Vietnam
A-27 IB2024 2002 Vietnam
A-27 IB2026 2003 Vietnam
A-27 IB1995 2003 Vietnam
A-27 IB2015 2002 Vietnam
A-27 IB2009 2003 Vietnam
A-27 IB2004 2003 Vietnam
A-27 IB1970 2001 Vietnam
A-27 IB1997 2003 Vietnam
A-27 IB1987 2002 Vietnam
A-27 IB2018 2002 Vietnam
A-27 IB1990 2003 Vietnam
A-27 IB1993 2003 Vietnam
A-27 20040924 2004 Kenya
A-27 20021122 2002 UK
A-27 IB3374 2002 Pakistan
A-27 IB2 2003 Korea
A-27 *Ss046 China
A-27 IB3 2003 Korea
A-27 IB691 1999 Korea
A-27 IB687 1998 Korea
A-27 IB1 2003 Korea
A-27 IB690 2000 Korea
A-27 IB681 1991 Korea
A-27 IB683 1994 Korea
A-27 IB2013 2001 Vietnam
A-27 IB1985 2002 Vietnam
A-27 IB1976 2002 Vietnam
A-27 IB2012 2001 Vietnam
A-27 IB1980 2002 Vietnam
g* A-27 62542 2006 Burkina Faso
g* A-27 36224 2003 Senegal
g* A-27 65623 2006 Morocco
g A-27 32222 2003 Cuba
g A-27 60108 2006 French Guiana
A-27 20081885 2008 UK
A-27 20051272 2005 Egypt
A-27 20060018 2006 Egypt
A-27 20061309 2006 Egypt
TIP LABELS
Region of isolation
● Europe
● Central Asia
● Vietnam
● Korea
● Middle East
● Egypt
● East Africa & Madagascar
● North/West/Central Africa
● South America
● Caribbean/Central America
○ Other/Unknown
ADDITIONAL ISOLATE INFO
● Virulence plasmid ≥10x
Biotypes
a ONPG +, rhamnose +, xylose -
d ONPG -, rhamnose +, xylose +
e ONPG +, rhamnose +, xylose +
f ONPG -, rhamnose -, xylose -
g ONPG +, rhamnose -, xylose -
g* atypical g
CRISPR spacers
A-27-7-10
Resistance gene coverage (%)
0
50
100
GyrA mutations
■ GyrA83-L
■ GyrA87-G
■ GyrA87-Y

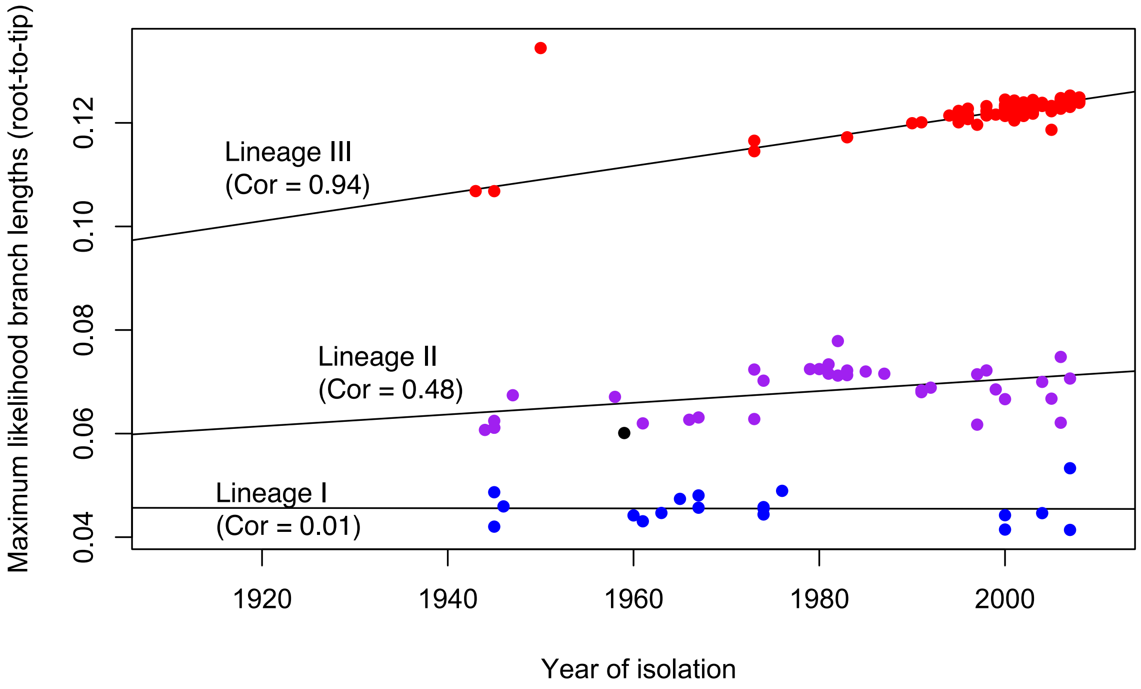
Supplementary Figure 2

Maximum likelihood tree of the 132 sequenced *S. sonnei* genomes, rooted using non-*S. sonnei* outgroups.

*Finished genome sequences (Ss046, NC_007384; 53G, accession pending). For each isolate is indicated the region of isolation, >10x coverage of pINV B, biotype, CRISPR spaces, isolate ID, year and country of isolation; according to inset legend. Gain (+) and loss (-) of genes on major branches are indicated with arrows; loss of gene function related to biotypes or CRISPR types are indicated in blue and red, respectively.

Distribution of antimicrobial resistance-associated genes and GyrA SNPs conferring resistance to Nalidixic acid are indicated via heatmap, according to inset legend which reflects percentage coverage of each gene sequence. Resistance-conferring genes are labelled in red and the antimicrobials they confer resistance to are indicated (Tmp, trimethoprim; Str, streptomycin; Sul, sulfonamides; Amp, ampicillin; Chl, chloramphenicol; Tet, tetracycline; Cef, third generation cephalosporins; Nal, Nalidixic acid).

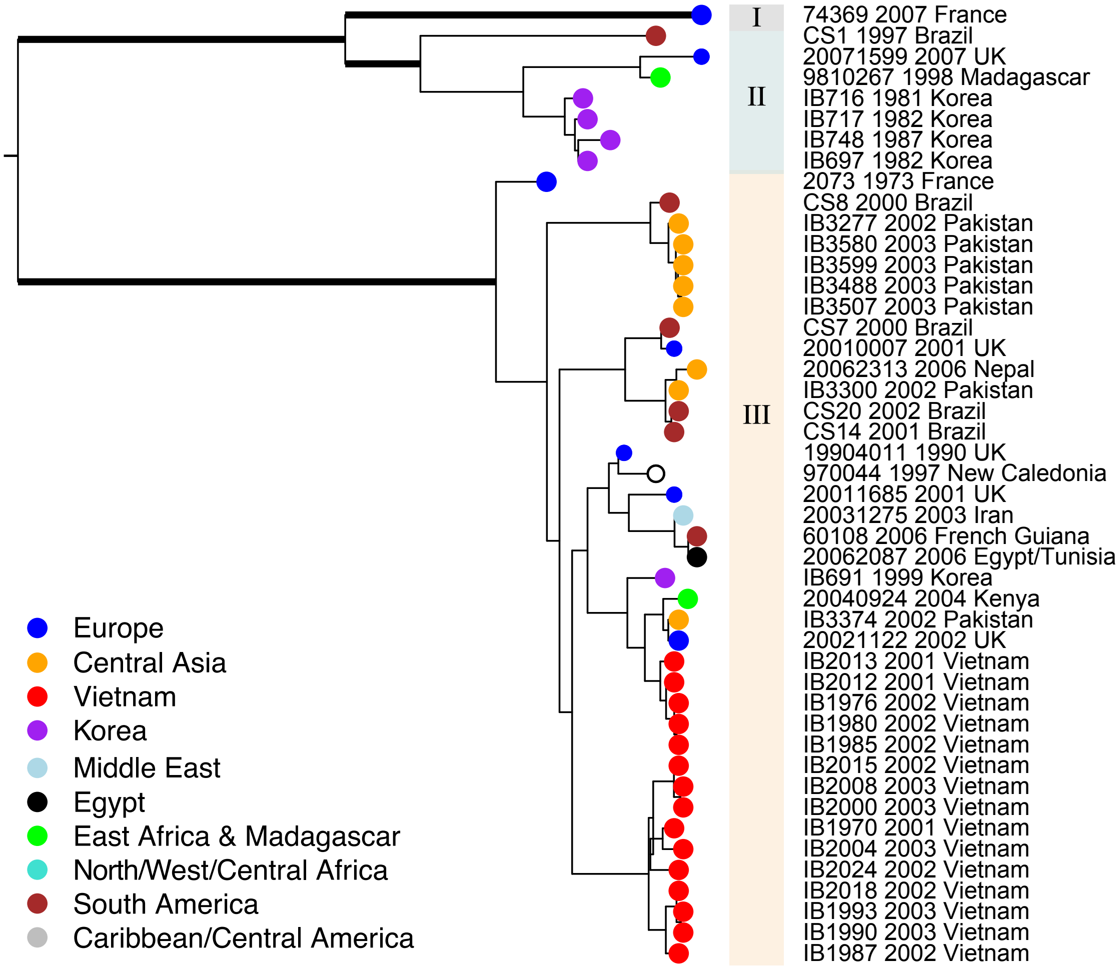
Reference sequences for resistance elements are: Tn*7*/In*2*, Ss046 genes SSON_3891-SSON_3897, NC_007384; spA, Ss046 plasmid spA, NC_009345; *blaTEM-1*, GQ983346; Tn*21*/In*1*, plasmid R100, NC_002134; In*38*/*catA1 C. freundii* transposon, AY162283; *catB3*, DQ321671; Tn*10*, J01830; *dfrA5*, X12868; *dfrA8*, U10186; *dfrA14*, Z50804; *dfrA12*, Z21672; *aadA2*, X68227; *blaOXA-1*, JN003856.1; *blaOXA-10*, DQ321671; *blaCTX-M-15*, DQ915953.



Supplementary Figure 3

Year of isolation vs. root-to-tip distances extracted from maximum likelihood phylogeny.

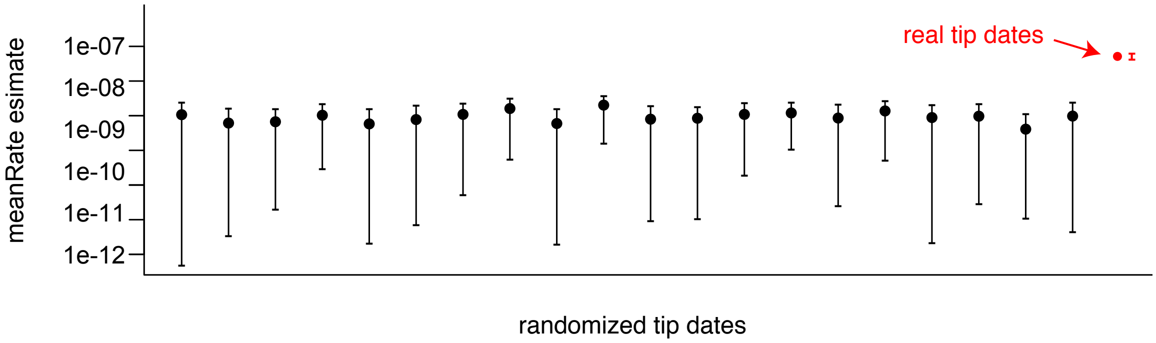
Linear regression lines and correlation coefficients are indicated separately for each lineage. Note the outlying red circle represents the reference genome *S. sonnei* Ss046 for which raw sequence reads were not available for SNP validation; hence it is possible that the long branch length reflects base call errors in the reference genome rather than true variation and this point was therefore excluded from linear regression.



Supplementary Figure 4

Bayesian (Maximum Clade Credibility) phylogenetic tree for *S. sonnei* virulence plasmid pINV B.

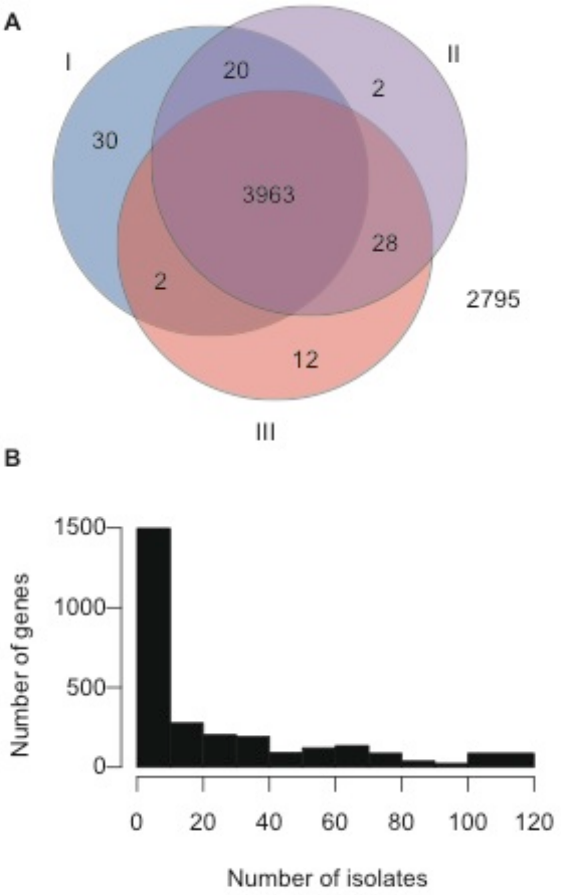
Note that the plasmid is commonly lost during laboratory culture, this tree includes only those 46 isolates with >10x mean read depth across the plasmid.



Supplementary Figure 5

Comparison of Bayesian estimates of nucleotide substitution rates for real and randomized tipdates.

Filled circles indicate mean estimates, while bars indicate values of the 95% highest probability density interval. The estimate from the real tipdate associations is shown in red, and estimates from randomized associations are shown in black. All randomized data sets were analyzed in BEAST using identical model settings as used in the analysis of the real tipdate data, with all Markov chains run for 200 million generations. Note the y-axis is on the log scale.



Supplementary Figure 6

Gene content variation in *S. sonnei*.

(A) Size of overlapping gene sets for *S. sonnei* lineages I, II and III. Conserved genes (N=3,963, central area) were present in ≥90% of isolates from each lineage. (B) Frequency distribution of 2,795 non-conserved genes that were not associated with any specific lineage.

Supplementary Tables

Supplementary Table 1 – Details of *Shigella sonnei* isolates used in this study.

Includes source; year and country of isolation; coverage information; lineage, biotype, CRISPR spacers; resistance phenotypes, gyrA SNPs and coverage of resistance-associated genes.

See separate Excel file

Supplementary Table 2 – *E. coli/Shigella* outgroups used to root the ML tree.

Serotype / species	Strain	GenBank Accession
Shigella boydii 4	Sb227	NC_007613
Shigella dysenteriae 1	Sd197	NC_007606
Shigella flexneri 2a	2457T	NC_004741
Shigella flexneri 2a	301	NC_004337
Shigella flexneri 5	8401	NC_008258
E. coli OR:H48:K- **(**K12)	MG1655	NC_000913
E. coli O157:H7	Sakai	NC_002695

Supplementary Table 3 – CRISPR spacer sequences analysed in this study.

Identifier	Sequence
A	TCTAAGTGATACCCATCATCGCATCCAGTGCGTC
27	ACCCTGACGCGCCGCAGTATTTATCTGCTCTGGC
7	ACGGGTGCGTGTGGCTGCCAGTGCCGGAGAACGG
10	TCTTACTGCTTGGTATGCGGAATCACACCCTGAA

Supplementary Note

SNP distribution and coding effects

The coding effect of each SNP was determined by mapping the SNP coordinate and allele to the reference genome *S. sonnei* strain Ss046 and its annotation (NC_007384). The available sites for synonymous (S) and nonsynonymous (N) substitutions in the 3,571 genes included in the analysis were obtained using the cusp (codon usage) tool in the EMBOSS package[41], resulting in a N:S sites ratio of 3.2. Mean d_N/d_S values were then approximated for each gene by dividing the ratio of observed nonsynonymous (n_g) and synonymous (s_g) SNPs by the genome-wide N:S ratio of 3.2, i.e. $d_N/d_S(g)=(n_g/s_g)/(N/S)$. This was possible only for the 1,455 genes with at least one synonymous and one nonsynonymous SNP.

The number of SNPs observed per gene was closely correlated with gene length, with a mean divergence of 0.23% within coding sequences (using linear regression model of SNPs on length; $R^2 = 0.67$, $p < 2 \times 10^{-16}$), same as the genome-wide mean divergence rate. Thus, SNPs were distributed randomly around the genome (Supplementary Fig. 1), randomly among genes and randomly across the length of genes. We therefore assumed a uniform distribution of SNPs along the chromosome with a mean nucleotide divergence of 0.23% and mean amino acid divergence of 0.16% (using linear regression model of nonsynonymous SNPs on gene length; $R^2 = 0.55$, $p < 2 \times 10^{-16}$). However, observed divergence values ranged up to 2% nucleotide divergence and 5.5% amino acid divergence. We calculated the probability of observing each gene *g*'s divergence levels using a Poisson distribution function with mean length(*g*)*0.0023 (nucleotide divergence) or length(*g*)*0.0018 (amino acid divergence). To correct for multiple testing, we adjusted p-values using Benjamini-Hochberg correction. The two genes discussed in the text are those with adjusted p<0.05.

As the *S. sonnei* genomes are closely related, sequences acquired by recombination would manifest as loci with a high density of SNPs[16,42]. Thus to screen for possible recombination events, we used the same statistical approach as above to detect regions with high SNP density, irrespective of coding consequence, within individual genomes (to reveal imports in each genome compared to the Ss046) and among SNPs defining each branch of the ML phylogenetic tree (to reveal imports into groups of strains that might include the reference Ss046, including recombination between *S. sonnei* strains or lineages). We detected clusters affecting five prophage sequences, six transposases or recombinases and one other locus. The latter involved a divergent copy of the *sitABCD* operon (SSON_1750-SSON_1753) in a single lineage III *S. sonnei* isolated in 2003 (strain 31382). The region assembled into a single contig and shared 99.38% nucleotide identity with *sitABCD* from *E. coli* 042 (FN554766.1), compared to 97.89% identity with 53G (lineage II) or Ss046 (lineage III).

Characteristics of the major S. sonnei lineages

The three main lineages of *S. sonnei* were defined by hundreds of SNPs (Supplementary Fig. 2). To determine whether they were also differentiated by the loss and acquisition of genes, we used a combined approach of read assembly and mapping to investigate gene content across the *S. sonnei* genomes (see Methods). We identified 6,852 genes in total, including 3,963 (58%) core genes that were present in ≥90% of isolates from each lineage (Supplementary Fig. 6A). Many accessory or non-core genes were present in less than half of isolates (N=2,459 genes, 85%) and very few genes were associated with specific lineages (Supplementary Fig. 6). Ongoing degradation of fimbrial genes complements was evident within lineages, including the loss of *ybcQR*, *sfmFH* and *lpfD* in lineage I and interruption of *lpfC* in lineage III. Lineages II and III shared the acquisition of a chromosomally encoded *relB*/*yafQ* toxin-antitoxin system and disruption of the xylose operon (deletion of *xylFG*) resulting in an inability to metabolise xylose. The acquisition of insertion sequences was also evident, with IS*911* (SSON_3904) present in lineages II and III but not lineage I, and IS*640* (SSON_1757) restricted to Lineage III.

Correlation of S. sonnei phylogenetic lineages with existing subtyping schemes

The genetically distinct lineages of *S. sonnei* were correlated with biotypes and CRISPR types used for subtyping *S. sonnei* populations[9,10], and revealed some of the genetic mechanisms driving the differentiation observed by subtyping (Supplementary Fig. 2, Supplementary Table 1).

Classical biotypes, based on breakdown of xylose, rhamnose and β-galactosidase (detected by cleavage of chromogenic substrate ortho-nitrophenyl-β-galactoside, ONPG)[9], were closely associated with the three lineages (Supplementary Fig. 2, Supplementary Table 1). Most biotyped lineage I isolates were able to metabolise rhamnose and xylose (biotypes d, e) but a subclade was ONPG-, which could be explained by a deletion in *lacZ* (Supplementary Fig. 2). Inactivating mutations affecting *lacZ* were also observed in lineage II (biotype f, ONPG-) and lineage III (biotypes unknown) (Supplementary Fig. 2). All biotyped lineage II and III isolates where xylose negative (biotypes a, f, g), due to a deletion of part of the *xyl* xylose operon that was conserved in all lineage II and III genomes. The biotyped Lineage III isolates were also rhamnose negative (biotype g), likely explained by a conserved nonsense mutation in codon 130 of *rhaR*, the transcriptional activator of the *rha* rhamnose operon.

The *S. sonnei* lineages also differed in their CRISPR types (Supplementary Fig. 2, Supplementary Table 1). We searched for four CRISPR spacer sequences (A, 27, 7, 10; Supplementary Table 3) by mapping to the CRISPR region of *S. sonnei* 53G. Loss of these spacer sequences was associated with inactivating mutations in CRISPR-associated genes *cas3* or *cse1*[10] (Supplementary Fig. 2).