Holt, Kathryn E; Baker, Stephen; Weill, François-Xavier; Holmes, Edward C; Kitchen, Andrew; Yu, Jun; Sangal, Vartul; Brown, Derek J; Coia, John E; Kim, Dong Wook; +8 more... Choi, Seon Young; Kim, Su Hee; da Silveira, Wanderley D; Pickard, Derek J; Farrar, Jeremy J; Parkhill, Julian; Dougan, Gordon; Thomson, Nicholas R; (2012) Shigella sonnei genome sequencing and phylogenetic analysis indicate recent global dissemination from Europe. Nature genetics, 44 (9). pp. 1056-1059. ISSN 1061-4036 DOI: https://doi.org/10.1038/ng.2369

Downloaded from: http://researchonline.lshtm.ac.uk/id/eprint/612562/

DOI: https://doi.org/10.1038/ng.2369

Usage Guidelines:

Please refer to usage guidelines at https://researchonline.lshtm.ac.uk/policies.html or alternatively contact researchonline@lshtm.ac.uk.

Available under license: http://creativecommons.org/licenses/by-nc-nd/2.5/
Supplementary Information

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

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).
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: Tn7/In2, Ss046 genes SSON_3891-SSON_3897, NC_007384; spA, Ss046 plasmid spA, NC_009345; blaTEM-1, GQ983346; Tn21/In1, plasmid R100, NC_002134; In38/catA1 C. freundii transposon, AY162283; catB3, DQ321671; TnJ0, 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.

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

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

<table>
<thead>
<tr>
<th>Identifier</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>TCTAAGTGATACCCATCATCGCATCCAGTGCGTC</td>
</tr>
<tr>
<td>27</td>
<td>ACCCTGACGCGCCGAGTATTTATATCTGCTCTGGC</td>
</tr>
<tr>
<td>7</td>
<td>ACGGGTGCGTGCTGGCTCCAGTCCAGTGCCGAGAACCGG</td>
</tr>
<tr>
<td>10</td>
<td>TCTTACTGCTTGGTATGCGGAATCACACCCTGAA</td>
</tr>
</tbody>
</table>
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 package41, 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 SNPs42. 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 IS911 (SSON_3904) present in lineages II and III but not lineage I, and IS640 (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* populations9,10, 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).