Manuscript Information

Journal name: Nature genetics
Manuscript #: 49152
Manuscript Title: Shigella sonnei genome sequencing and phylogenetic analysis indicate recent global dissemination from Europe
Principal Investigator: J Yu (jun.yu@strath.ac.uk)
Submitter: Nature Publishing Group (repositorynotifs@nature.com)

Grant/Project/Contract/Support Information

Name	Support ID#	Title
J Yu	MRC_G0800173 (86345)	Genotyping epidemic Shigella sonnei

Manuscript Files

Type	Fig/Table #	Filename	Size	Uploaded
manuscript	1	article_1.doc	276480	2012-07-06 08:00:36
figure	1	figure_1.ai	3313183	2012-07-06 08:00:53
supplement	1	supp_info_1.pdf	7713830	2012-07-06 08:01:30
supplement	table 1	supp table 1.xls	167936	2012-07-09 13:31:40

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

Kathryn E. Holt[1], Stephen Baker[2], François-Xavier Weill[3], Edward C. Holmes[4,5], Andrew Kitchen[4], Jun Yu[6], Vartul Sangal[6], Derek J. Brown[7], John E. Coia[7], Dong Wook Kim[8,9], Seon Young Choi[8], Su Hee Kim[8], Wanderley D. da Silveira[10], Derek J. Pickard[11], Jeremy J. Farrar[2], Julian Parkhill[11], Gordon Dougan[11], Nicholas R. Thomson[11]

1. University of Melbourne, Department of Microbiology and Immunology, Royal Parade, Melbourne, Victoria, 3010, Australia
2. The Hospital for Tropical Diseases, Wellcome Trust Major Overseas Programme, Oxford University Clinical Research Unit, Ho Chi Minh City, Vietnam
3. Institut Pasteur, Unité des Bactéries Pathogènes Entériques, Paris, France
4. Center for Infectious Disease Dynamics, Department of Biology, The Pennsylvania State University, University Park, PA 16802, USA
5. Fogarty International Center, National Institutes of Health, Bethesda, MD 20892, USA
6. Strathclyde Institute of Pharmacy and Biomedical Sciences, University of Strathclyde, Glasgow, G4 0RE, UK
7. Scottish *Salmonella*, *Shigella* and *Clostridium difficile* Reference Laboratory, Stobhill Hospital, 133 Balornock Road, Glasgow, UK
8. Molecular Biology Laboratory, International Vaccine Institute (IVI), Seoul, Republic of Korea
9. Department of Pharmacy, College of Pharmacy, Hanyang University Ansan, Kyeonggi- do, 426-791, Korea
10. Department of Genetics, Evolution and Bioagents, Biology Institute, Campinas State University – UNICAMP, Brazil
11. Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK CB10 1SA

Abstract/First paragraph

Shigella are human-adapted *Escherichia coli* that have gained the ability to invade the human gut mucosa and cause dysentery[1,2], spreading efficiently via low-dose fecal-oral transmission[3,4]. Historically, *S. sonnei* has been predominantly responsible for dysentery in developed countries, but is now emerging as a problem in the developing world, apparently replacing the more diverse *S. flexneri* in areas undergoing economic development and improvements in water quality[4-6]. Classical approaches have shown *S. sonnei* is genetically conserved and clonal[7]. We report here whole-genome sequencing of 132 globally-distributed isolates. Our phylogenetic analysis shows that the current *S. sonnei* population descends from a common ancestor that existed less than 500 years ago and has diversified into several distinct lineages with unique characteristics. Our analysis suggests the majority of this diversification occurred in Europe, followed by more recent establishment of local pathogen populations in other continents predominantly due to the pandemic spread of a single, rapidly-evolving, multidrug resistant lineage.

To establish an accurate population framework we sequenced the whole genomes of 132 *S. sonnei* isolated between 1943 and 2008, spanning four continents (Supplementary Table 1). We detected 10,111 chromosomal single nucleotide polymorphisms (SNPs) randomly distributed around the *S. sonnei* chromosome, approximately one per 430 bp (0.23% nucleotide divergence) (Supplementary Fig. 1). To investigate the population structure of *S. sonnei*, we analysed these chromosomal SNPs using multiple phylogenetic methods. Maximum likelihood (ML) phylogenetic analysis (Supplementary Fig. 2) revealed a strong correlation between root-to-tip

branch lengths and the known dates of isolation for the sequenced *S. sonnei*, indicative of rapid, clock-like evolution (Supplementary Fig. 3). There appears to be some rate variation between lineages, possibly associated with differences in effective population size or in the mean number of generations per year (replication rate), which may in turn be associated with different lifestyles or niches. We used a Bayesian approach (BEAST[8]) to infer the evolutionary dynamics of the global *S. sonnei* population as a whole. Importantly, this yielded the same tree topology as the ML analysis, while also providing estimates of nucleotide substitution rates and divergence times for key *S. sonnei* lineages (Fig. 1). Interestingly, the phylogenies identified four distinct *S. sonnei* lineages, three encompassing isolates spanning the 1940s through the 2000s and another comprising a single isolate from France (Fig. 1). These lineages each had 100% ML bootstrap support, 100% Bayesian posterior support (BEAST) and were also recovered using a Bayesian clustering analysis (see Online Methods). Whilst these lineages are uniquely characterized by hundreds of SNPs they display only minor differences in gene content and were correlated with traditional typing methods used to subdivide *S. sonnei* (biotypes a-g[9] and CRISPR types[10]) (Supplementary Note, Supplementary Fig. 2, Supplementary Table 3). We estimated a mean substitution rate of 2.0×10^{-4} site^{-1} year^{-1} among the 10,111 chromosomal SNP loci [95% Highest Posterior Density (HPD) $1.6 \times 10^{-4} – 2.3 \times 10^{-4}$], corresponding to the accumulation of approximately 2.2 SNPs chromosome^{-1} year^{-1} ([95% HPD 1.8 – 2.6], excluding repeated and phage regions). This scales to a genome-wide substitution rate of 6.0×10^{-7} substitutions site^{-1} year^{-1} [95% HPD = 5.2×10^{-7} - 6.7×10^{-7}], which likely represents the upper bound of the true genome-wide substitution rate and is similar to that calculated for the enteric pathogen *Vibrio cholerae* (8×10^{-7} site^{-1} year^{-1})[11] but lies between the rates estimated for *Yersinia*

pestis (2×10^{-8})[12] and *Staphylococcus aureus* (3×10^{-6})[13]. From BEAST analysis, we estimated the most recent common ancestor (MRCA) of all contemporary *S. sonnei* existed less than 500 years ago [median calendar year for divergence date, 1669; 95% HPD, 1554 - 1763] (Fig. 1). Similarly, we estimate the MRCA for each of Lineages I and II existed in the early 19th century and that all Lineage III isolates descend from a hypothetical ancestor that existed around the turn of the 20th century (Fig. 1). Critically, these data indicate that though the extant *S. sonnei* population descends from a single ancestor existing in the 17th century, by the late 19th century *S. sonnei* had become segregated into at least four distinct lineages that still persist today.

There was strong evidence for regional clustering of *S. sonnei* within the phylogenetic tree (Fig. 1), indicating significant geographic structure in the global bacterial population ($p < 1 \times 10^{-5}$ for association between phylogeny and geographic region[14]). Interestingly, the European population shows the richest diversity, with isolates distributed across all four lineages (31% lineage I, 35% lineage II, 31% lineage III, sole lineage IV isolate) and occupying basal branches in each lineage (Fig. 1). In contrast, *S. sonnei* isolates from Asia, Africa and America were mainly from lineage III (67-77%) with fewer lineage II representatives (22-26%) and just two from Lineage I. Furthermore, ancestral state reconstruction analysis indicated a >50% likelihood of a European common ancestor for each of the lineages I, II and III (Fig. 1). The data also indicate Lineage III has been more successful at global dispersal than other lineages, with only low numbers of Lineage I or II detected outside Europe (Fig. 1). In particular, a recently derived clade within Lineage III (Global III, MRCA = 1972 [95% HPD = 1964-1979 C.E.]) has been particularly successful at global dissemination, comprising 49% of all isolates sampled since 1995 and detected in all

104 regions represented in our collection (Fig. 1). Unlike the European isolates, isolates
105 from non-European countries form tight shallow-rooted phylogenetic clusters,
106 consistent with and suggestive of contemporary dispersal (Fig. 1). In many cases,
107 these clusters contain multiple isolates from the same country, indicating localized
108 clonal expansions (Fig. 1). For example, isolates from Korea formed two subclades
109 within lineages II and III that likely represent separate introductions of *S. sonnei* into
110 Korea during the 1960s and 1970s, each followed by local clonal expansions (Fig. 1).
111 Similarly, isolates originating in Vietnam form two subclades, indicating the local
112 establishment of Lineage III clones in Vietnam in the 1990s (Fig. 1). At a regional
113 level, there appears to have been an establishment of a Lineage III subclade in South
114 America during the 1950s to which isolates from Brazil and Peru could be traced,
115 followed by dissemination of the Global III clade into Africa and America in the early
116 1980s (Fig. 1).
117
118 Critically, the phylogeographic analysis indicates that all contemporary *S. sonnei*
119 infections are caused by a small number of clones that have recently become globally
120 dispersed (Fig. 1). The distribution of antimicrobial resistance genes and mutations
121 within the *S. sonnei* phylogeny suggest that selection for multiple drug resistance
122 (MDR) played a pivotal role in driving this global dissemination (Fig. 1,
123 Supplementary Fig. 2, Supplementary Table 1). In particular, the establishment of
124 local *S. sonnei* Lineage III populations outside Europe is intimately associated with
125 the carriage of transposon Tn*7* and class II integrons (In*2*) encoding resistance to
126 multiple antimicrobials (Fig. 1). All three major Lineage III subgroups carry a distinct
127 In*2* variant, which is either plasmid-encoded (South America III) or integrated into
128 the chromosome adjacent to *glmS* (Central Asia IIIa, Global III), suggesting

129 independent acquisitions of the integron in each group during the 1960s-1970s
130 followed by clonal expansion and subsequent international spread (Fig. 1). Studies
131 from Europe, Asia, Africa, South America and Australia have reported a high
132 prevalence of In*2*-bearing, MDR, biotype g *S. sonnei*, often associated with local
133 epidemics[15]. Our data demonstrate biotype g is a marker for Lineage III due to a
134 conserved nonsense mutation in rhamnose regulatory gene *rhaR* (Supplementary Fig.
135 2) and indicate that the global distribution of MDR biotype g/In*2 S. sonnei* is the
136 result of global dissemination of multiple In*2*-bearing subclades of Lineage III *S.*
137 *sonnei*. Half of the In*2*-bearing Lineage III isolates also harboured the small MDR
138 plasmid spA[2] containing *tetAR*, *strAB* and *sul2* genes, which confer additional
139 resistance to tetracycline, streptomycin and sulfonamides (Fig. 1). All quinolone
140 resistant isolates harboured one of three point mutations in the chromosomal DNA
141 gyrase gene, *gyrA,* known to confer quinolone resistance (Fig. 1, Supplementary
142 Table 1; we detected no plasmid-mediated quinolone resistance genes). The
143 distribution of *gyrA* mutations within the phylogeny shows these resistance mutations
144 have arisen independently on at least nine occasions among our *S. sonnei* collection,
145 including two separate mutations within the clonal group Korea II, indicative of
146 surprisingly strong selection for quinolone resistance even among MDR isolates (Fig.
147 1). To investigate other signals of selection, we examined the clustering of SNPs
148 within genes and chromosomal regions (Supplementary Note). We found evidence of
149 phage and transposase insertions and a single case of homologous recombination
150 affecting the *sitABCD* operon in isolate 31382, but identified only two genes
151 displaying amino acid variation significantly higher than expected under a random
152 distribution of SNPs. Neither of these genes (*rpoS* and *mreB*) encodes an extracellular
153 protein, suggesting a lack of immune selection, in common with another human

154 restricted pathogen *Salmonella* Typhi (typhoid fever)[16]. However, we detected a large
155 number of nonsynonymous SNPs (nsSNPs) and a high rate of nonsyonymous to
156 synonymous substitutions per site (d_N/d_S) in the drug efflux pump component genes
157 *acrD* (8 nsSNPs, d_N/d_S =2.5) and *acrB* (12 nsSNPs, d_N/d_S =1.8). Currently,
158 antimicrobial treatment is recommended for the management of dysentery[17], but may
159 not significantly impact the resolution of *S. sonnei* or *S. flexneri* infections[18,19].
160 However, there is evidence such treatment can prevent shedding of *S. sonnei* after the
161 resolution of symptoms[20]. Thus, while antimicrobial resistance may have only minor
162 implications for dysentery treatment, this phenotype may be important in sustaining *S.*
163 *sonnei* transmission within human populations and our data indicates there is a strong
164 selective pressure for its maintenance. It has been hypothesized that free-living
165 amoebae may represent an environmental reservoir for *Shigella*, which are able to
166 survive intracellularly within *Acanthamoeba*[21,22]. This could potentially provide
167 another niche in which selective pressure for antibiotic resistance may be exerted,
168 although intracellular *Shigella* are likely to be protected from most antibiotics by their
169 amoebae hosts[23,24].
170
171 Previous studies have proposed that the acquisition of virulence plasmid pINV B,
172 encoding the *Plesiomonas shigelloides* related O antigen, was the defining event in
173 the emergence of *S. sonnei*[25]. Unfortunately, the *S. sonnei* virulence plasmid is highly
174 unstable on laboratory media and is commonly lost on sub-culturing[26] and, as a
175 consequence, less than half of our isolates yielded sufficient virulence plasmid
176 sequence data for analysis (46 isolates with >10x read depth). Phylogenetic analysis
177 of the available virulence plasmid sequences (which contained 84 SNPs) identified

three distinct lineages (Supplementary Fig. 4). There was a parallel relationship between chromosomal and plasmid lineages, consistent with co-evolution of the plasmid and host chromosome, stable maintenance of the plasmid in the natural environment and no transfer of plasmid variants among host bacteria. It has also been proposed that exposure to *P. shigelloides* via contaminated water protects humans from *S. sonnei* infection[5] as the O antigens are indistinguishable and cross-react[27,28]. This may explain increases in *S. sonnei* incidence following economic development and water quality improvements, as the result of a decline in passive cross-protection by environmental immunization with *P. shigelloides*. If this cross-protection acts as a barrier to the establishment of *S. sonnei* in human populations, one would predict that *S. sonnei* infections would gradually increase following improvements in water quality, and that the geographical expansion of *S. sonnei* will be characterized by the introduction and expansion of novel clones moving into human populations with falling natural immunity previously obtained from exposure to *P. shigelloides*. Our model of recent dissemination out of Europe is remarkably consistent with these hypotheses. Transmission of *S. sonnei* into other continents has likely occurred sporadically over centuries through human migration, trade and travel; however the establishment of local *S. sonnei* populations – which we would observe as geographically clustered clonal groups outside Europe – is not evident until the last few decades.

Our findings have major implications for global public health and diarrheal infections. Improvement of drinking water, one of the Millenium Development Goals, is an undeniably important aim and is expected to reduce morbidity and mortality due to a diverse array of waterborne diseases. However, we predict that fulfilling this aim will

produce a concurrent increase in *S. sonnei* dysentery incidence in transitional countries. The combination of increased incidence and excessive antimicrobial resistance among globally disseminated *S. sonnei* indicates an anti-*S. sonnei* vaccine will be increasingly important for the control and long-term prevention of dysentery and associated morbidity and mortality. A suitable vaccine is an achievable goal, since all *S. sonnei* share a single O antigen that has proven to be a successful vaccine target[29]. Interestingly, the success of *S. sonnei* in the face of diminishing *S. flexneri* incidence suggests important epidemiological distinctions in transmission of the two pathogens. *S. sonnei* outbreaks have been associated with schools, care facilities, contaminated food and insects moving between fecal waste and food preparation areas[30-32]. These modes of transmission are considerably more direct than waterborne transmission and may explain the persistence of *S. sonnei* even when water infrastructure is improved, implying that vaccination and improved hygiene standards will be pivotal in eliminating *S. sonnei* infections in industrializing countries.

URLs

Illumina sequence data provided at http://www.ebi.ac.uk/ena/data/view/ERP000182

TreeStat: http://tree.bio.ed.ac.uk/software/treestat/

Velvet Optimiser: http://www.ebi.ac.uk/~zerbino/velvet/

Acknowledgements

This work was supported by the Wellcome Trust (#0689); and a Victorian Life Sciences Computation Initiative (VLSCI) grant (#VR0082) on its Peak Computing Facility at the University of Melbourne, an initiative of the Victorian Government, Australia. KEH was supported by a Fellowship from the NHMRC of Australia (#628930); SB is supported by an OAK Foundation Fellowship through Oxford University (#OAKF9) and the Li Ka Shing foundation (#LG13); FXW was partially funded by the Institut de Veille Sanitaire; JY was supported by a MRC grant (#G0800173); DWK was partially supported by grant RTI05-01-01 from the Ministry of Knowledge and Economy (MKE). We thank Myron Levine (University of Maryland School of Medicine, Center for Vaccine Development, USA) and Christoph Tang (University of Oxford, UK) for their kind gift of *Shigella sonnei* strain 53G.

Author Contributions

KEH, NRT, ECH and AK analysed the data and performed phylogenetic analysis. NRT, GD, JY, SB, JJF, KEH and JP were involved in the study design. FXW, DJB, JEC, JY, VS, DWK, SYC, SHK, WDS and DJP were involved in isolate collection, DNA analysis and resistance phenotyping. KEH, SB, NRT, GD, AK, ECH and FXW contributed to the manuscript writing.

Accession Numbers

The finished genome of *S. sonnei* 53G is available under EMBL accessions HE616528 (chromosome) and HE616529, HE616530, HE616531 and HE616532 (plasmids). Sequence reads for the 132 Illumina-sequenced *S. sonnei* are deposited in the European Nucleotide Archive under accession ERP000182.

247

248 The authors declare no competing financial interests.

Figure Legends

Figure 1. Bayesian maximum clade credibility phylogeny for *S. sonnei*. Branches defining major lineages in bold (each had 100% posterior support); pie charts indicate ML estimates for geographic origin of major nodes, according to inset legend (lower left). Time (x-axis) is relative to the Common Era; divergence dates (median estimate and 95% HPD) are given in blue for major nodes. Distribution of antimicrobial resistance determinants is indicated in the heatmap according to the legends provided, which reflect percentage of bases in each gene sequence that are covered by reads from each isolate (top right). Geographically localised clonal expansions are highlighted on the right, labeled with their median estimated divergence date.

Online Methods

Bacterial isolates and sequencing

Bacterial isolates analysed in this study are detailed in Supplementary Table 1. DNA was prepared using the Wizard Genomic DNA Kit (Promega, Madison, WI) or phenol extraction. Index-tagged paired end Illumina sequencing libraries were prepared using one of 12 unique indexing tags as previously described[13]. These were combined into pools each containing 11-12 uniquely tagged libraries and sequenced on the Illumina Genome Analyzer GAII according to manufacturer's protocols to generate tagged 54 bp paired-end reads.

Read alignment and SNP detection

Reads from each isolate were mapped to the *S. sonnei* reference genome (strain Ss046 chromosome, NC_007384; strain Ss046 plasmids, NC_007385, NC_009347, NC_009346, NC_009345; plasmid pEG356, NC_013727) using BWA[33] with default parameters. Average read depths are given in Supplementary Table 1. SNPs were identified using SamTools[34]. SNPs in the previously sequenced *S. sonnei* strain 53G were identified using the same mapping procedure to analyse reads simulated from the finished genome (chromosome: HE616528; plasmids: HE616529, HE616530, HE616531 and HE616532) using SamTools' wgsim algorithm. SNPs called in phage regions or repetitive sequences (10.2% of bases and 15.5% of genes in the Ss046 reference chromosome) were excluded[16], resulting in a final set of 10,111 chromosomal SNP loci. The allele at each locus in each isolate was determined by reference to the consensus base in that genome (using SamTools pileup and removing

low confidence alleles with consensus base quality ≤20, read depth ≤5 or a heterozygous base call).

The SNP calling procedure was repeated using *S. sonnei* 53G (Lineage II) as the reference for mapping. This resulted in an identical tree topology with near-identical branch lengths (Pearson correlation coefficient = 0.995, $p<1x10^{-15}$), demonstrating the robustness of the method and its independence from the choice of reference genome. The Ss046-mapped data was used for all analyses reported, since the Ss046 genome has been widely used in previous comparative studies while the 53G genome is reported here for the first time.

The same procedures were followed to identify SNPs in the invasion plasmid. The analysis was restricted to strains with a mean plasmid read depth of ≥10x and the 137 kbp of non-repetitive plasmid sequence (63% of the *S. sonnei* pSs046 reference plasmid sequence).

Alleles in outgroup genomes were determined using the same approach to analyse reads simulated from other *Shigella* and *E. coli* reference genomes (Supplementary Table 2) using wgsim (distributed with SamTools).

Phylogenetic and temporal analyses

Chromosomal SNP alleles were concatenated for each strain to generate a multiple alignment of all SNPs (where high confidence base calls could not be determined, the allele was recorded as a gap character). Clusters of SNPs introduced via horizontal transfer (see *SNP distribution* section below) were removed from the alignment. The

312 resulting alignment was further filtered to remove loci at which alleles were unknown
313 for >40% of isolates (indicating the site is not conserved) and an ML phylogeny was
314 estimated using RAxML[35]. The BEAST package[8] was utilized for the Bayesian
315 inference of phylogeny and divergence dates. Additionally, we used the *BAPS*
316 program (Bayesian Analysis of Population Structure)[36] to examine clustering of
317 isolates based on SNP data.
318
319 For ML analysis, RAxML was run ten times using the generalized time-reversible
320 model with a Γ distribution to model site-specific rate variation (i.e., the GTR+Γ
321 substitution model; GTRGAMMA in RAxML). 1000 bootstrap pseudo-replicate
322 analyses were performed to assess support for the ML phylogeny. The final result
323 (Supplementary Fig. 2) is the tree with the highest likelihood across all ten runs, with
324 ML estimates of branch length and confidence in major bipartitions calculated using
325 the bootstrap values across all runs. This phylogeny was rooted using *E. coli* and
326 *Shigella* outgroups (Supplementary Table 2).
327
328 Root-to-tip branches were extracted from the ML tree using the program TreeStat (see
329 URLs). The relationship between root-to-tip distances, year of isolation and lineage
330 were analysed using linear regression. Plots and regression lines are shown in
331 Supplementary Figure 3, along with Pearson correlation coefficients.
332
333 For BEAST analysis, we also used the GTR+Γ substitution model and defined tip
334 dates as the year of isolation (restricting the analysis to those sequences with recorded
335 dates). We performed multiple analyses using both constant size and Bayesian skyline
336 demographic models, in combination with either a strict molecular clock or a relaxed

clock (uncorrelated lognormal distribution). BEAST (v1.6) uses a Markov chain Monte Carlo (MCMC) method for sampling the posterior probability distributions. Analyses of all model combinations (demographic and clock) were performed using ten chains of 100 million generations each to ensure convergence, with samples taken every 1,000 MCMC generations. Parameters were estimated after combining all replicate analyses, totaling 900 million MCMC generations post-burnin, with all reported parameter estimates (i.e., medians and 95% Highest Probability Densities – HPDs) calculated using the program Tracer v1.5. The relaxed clock models provided much better fit to the data (Bayes Factor > 100; using the harmonic mean estimator of the marginal likelihood) and the standard deviation of inferred substitution rates across branches was 0.45 [95% HPD = 0.38 - 0.52], providing additional strong support for a relaxed molecular clock. Bayesian skyline plots indicated a constant population size through time and estimates under a constant population model yielded very similar results to that under a Bayesian skyline model. Therefore, all parameter estimates quoted are from analyses using relaxed clock and Bayesian skyline demographic models. To test the validity of the temporal signal in the data, we performed 20 additional BEAST runs (of 200 million MCMC generations each) with identical substitution (GTR+Γ), clock (relaxed), and demographic (Bayesian skyline) models, but with randomized tip dates (Supplementary Fig. 5). This randomization procedure produces a null set of tipdate and sequence correlations that may be analysed to produce null substitution rate distributions, which can then be compared with empirical rate estimates.

Phylogeographic analysis

361 The geographic region of isolation of each *S. sonnei* was analysed as a discrete
362 character trait using two complementary methods. Phylogeographic analyses were
363 performed using the 126 isolates which had complete information on both year and
364 geographic region of isolation (see Supplementary Table 1). First, the association
365 between the phylogenetic relationships of *S. sonnei* isolates (inferred by BEAST) and
366 their geographic region of isolation was tested using the Bayesian Tip-Significance
367 software (BaTS[14]). A random selection of 50,000 trees sampled during the Bayesian
368 phylogenetic analysis described above were used as input, and 1,000 randomizations
369 were used to generate a null distribution for significance testing. Second, ancestral
370 state reconstruction of the geographic origin of hypothetical common ancestors (i.e.,
371 internal nodes in the phylogeny) was performed using the 'ace' function implemented
372 in the 'ape' package for R[37]. The percent probability estimates quoted, and illustrated
373 by pie charts in Figure 1, are scaled likelihoods for the discrete character trait (i.e.,
374 region of isolation) at each node.
375
376 ***Gene content analysis***
377 Each read set was assembled using the *de novo* short read assembler Velvet[38] and
378 Velvet Optimiser (see URLs). Contigs less than 100 bp in size were excluded from
379 further analysis. The *S. sonnei* 53G genome (chromosome: HE616528; plasmids:
380 HE616529, HE616530, HE616531 and HE616532) and *de novo* assembled contig
381 sets were mapped iteratively to the pan-genome reference set (initialized as the
382 concatenation of *S. sonnei* Ss046 chromosome, NC_007384; Ss046 plasmids,
383 NC_007385, NC_009347, NC_009346, NC_009345; plasmid pEG356, NC_013727)
384 using MUMmer (nucmer algorithm)[39]. At each iteration *i*, sequences not aligning to
385 the current pan-genome P_{i-1} set were incorporated into an extended pan-genome, P_i.

386 The final pan-genome, *P*, was annotated using a combination of annotation transfer
387 (for *S. sonnei* reference sequences) and *de novo* annotation using the RAST
388 annotation server[40] for novel sequences assembled from reads. The latter included
389 1.67 Mbp of sequence in 862 contigs, in which 2,422 genes were annotated
390 (incorporating 80.5% of bases), resulting in a total of 6,852 genes.
391
392 *S. sonnei* read sets were then aligned to the pan-genome using BWA[27] with default
393 mapping parameters. A pileup was generated for each aligned read set using
394 SamTools[28] and used to summarize, for each annotated gene in the pan-genome *P*, the
395 coverage (% of bases covered) and presence of inactivating mutations (nonsense
396 SNPs or non-triplet indels resulting in frameshifts) in each genome. The results were
397 used to identify genes whose presence or inactivation was associated with specific
398 lineages (Supplementary Note, Supplementary Fig. 6).
399
400 ***Resistance gene analysis***
401 The presence of resistance genes was initially determined from mapping data
402 described above. The genetic context of resistance genes was examined by blastn
403 search of each contig set with known resistance, transposase or integrase genes as
404 query sequences. The resulting contigs were compared to the NCBI non-redundant
405 nucleotide database to annotate the resistance genes and mobile elements. Mapping
406 was then repeated using annotated mobile elements to generate the gene coverage
407 maps shown in Figure 1 and Supplementary Figure 2, which indicate the proportion
408 of bases in each gene sequence that are covered by reads from each isolate (reference
409 sequences are provided in Supplementary Fig. 2).

References

1. Pupo, G.M., Lan, R. & Reeves, P.R. Multiple independent origins of Shigella clones of Escherichia coli and convergent evolution of many of their characteristics. *Proc Natl Acad Sci U S A* **97**, 10567-72 (2000).

2. Yang, F. *et al.* Genome dynamics and diversity of Shigella species, the etiologic agents of bacillary dysentery. *Nucleic Acids Res* **33**, 6445-58 (2005).

3. DuPont, H.L., Levine, M.M., Hornick, R.B. & Formal, S.B. Inoculum size in shigellosis and implications for expected mode of transmission. *J Infect Dis* **159**, 1126-8 (1989).

4. Kotloff, K.L. *et al.* Global burden of Shigella infections: implications for vaccine development and implementation of control strategies. *Bull World Health Organ* **77**, 651-66 (1999).

5. Sack, D.A., Hoque, A.T., Huq, A. & Etheridge, M. Is protection against shigellosis induced by natural infection with Plesiomonas shigelloides? *Lancet* **343**, 1413-5 (1994).

6. Vinh, H. *et al.* A changing picture of shigellosis in southern Vietnam: shifting species dominance, antimicrobial susceptibility and clinical presentation. *BMC Infect Dis* **9**, 204 (2009).

7. Karaolis, D.K., Lan, R. & Reeves, P.R. Sequence variation in Shigella sonnei (Sonnei), a pathogenic clone of Escherichia coli, over four continents and 41 years. *J Clin Microbiol* **32**, 796-802 (1994).

8. Drummond, A.J. & Rambaut, A. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol Biol* **7**, 214 (2007).

9. Nastasi, A., Pignato, S., Mammina, C. & Giammanco, G. rRNA gene restriction patterns and biotypes of Shigella sonnei. *Epidemiol Infect* **110**, 23-30 (1993).

10. Touchon, M. *et al.* CRISPR distribution within the Escherichia coli species is not suggestive of immunity-associated diversifying selection. *J Bacteriol* **193**, 2460-7 (2011).

11. Mutreja, A. *et al.* Evidence for several waves of global transmission in the seventh cholera pandemic. *Nature* **477**, 462-5 (2011).

12. Morelli, G. *et al.* Yersinia pestis genome sequencing identifies patterns of global phylogenetic diversity. *Nat Genet* **42**, 1140-3 (2010).

13. Harris, S.R. *et al.* Evolution of MRSA during hospital transmission and intercontinental spread. *Science* **327**, 469-74 (2010).

14. Parker, J., Rambaut, A. & Pybus, O.G. Correlating viral phenotypes with phylogeny: accounting for phylogenetic uncertainty. *Infect Genet Evol* **8**, 239-46 (2008).

15. Ranjbar, R. *et al.* Genetic relatedness among isolates of Shigella sonnei carrying class 2 integrons in Tehran, Iran, 2002-2003. *BMC Infect Dis* **7**, 62 (2007).

16. Holt, K.E. *et al.* High-throughput sequencing provides insights into genome variation and evolution in Salmonella Typhi. *Nat Genet* **40**, 987-93 (2008).

17. World Health Organization Guidelines for the control of shigellosis, including epidemics due to Shigella dysenteriae 1. WHO Document Production Services, Geneva, Switzerland (2005).

18. Christopher, P.R., David, K.V., John, S.M. & Sankarapandian, V. Antibiotic therapy for Shigella dysentery. *Cochrane Database of Systematic Reviews*, CD006784 (2010).

458 19. Vinh, H. *et al.* A multi-center randomized trial to assess the efficacy of
459 gatifloxacin versus ciprofloxacin for the treatment of shigellosis in Vietnamese
460 children. *PLoS Negl Trop Dis* **5**, e1264 (2011).
461 20. Vinh, H. *et al.* Treatment of bacillary dysentery in Vietnamese children: two
462 doses of ofloxacin versus 5-days nalidixic acid. *Trans Royal Soc Trop Med Hyg* **94**,
463 323-6 (2000).
464 21. Jeong, H.J. *et al.* Acanthamoeba: could it be an environmental host of
465 Shigella? *Exp Parasitol* **115**, 181-6 (2007).
466 22. Saeed, A., Abd, H., Edvinsson, B. & Sandstrom, G. Acanthamoeba castellanii
467 an environmental host for Shigella dysenteriae and Shigella sonnei. *Arch Microbiol*
468 **191**, 83-8 (2009).
469 23. Winiecka-Krusnell, J. & Linder, E. Free-living amoebae protecting Legionella
470 in water: the tip of an iceberg? *Scand J Infect Dis* **31**, 383-5 (1999).
471 24. Greub, G. & Raoult, D. Microorganisms resistant to free-living amoebae. *Clin*
472 *Microbiol Rev* **17**, 413-33 (2004).
473 25. Shepherd, J.G., Wang, L. & Reeves, P.R. Comparison of O-antigen gene
474 clusters of Escherichia coli (Shigella) sonnei and Plesiomonas shigelloides O17:
475 sonnei gained its current plasmid-borne O-antigen genes from P. shigelloides in a
476 recent event. *Infect Immunity* **68**, 6056-61 (2000).
477 26. Sansonetti, P.J., Kopecko, D.J. & Formal, S.B. Shigella sonnei plasmids:
478 evidence that a large plasmid is necessary for virulence. *Infect Immunity* **34**, 75-83
479 (1981).
480 27. Van de Verg, L.L., Herrington, D.A., Boslego, J., Lindberg, A.A. & Levine,
481 M.M. Age-specific prevalence of serum antibodies to the invasion plasmid and

482 lipopolysaccharide antigens of Shigella species in Chilean and North American
483 populations. *J Infect Dis* **166**, 158-61 (1992).
484 28. Shimada, T. & Sakazaki, R. On the serology of Plesiomonas shigelloides. *Jap*
485 *J Med Science Biol* **31**, 135-42 (1978).
486 29. Kaminski, R.W. & Oaks, E.V. Inactivated and subunit vaccines to prevent
487 shigellosis. *Exp Review Vaccines* **8**, 1693-704 (2009).
488 30. Genobile, D. *et al.* An outbreak of shigellosis in a child care centre.
489 *Communicable Dis Intell* **28**, 225-9 (2004).
490 31. Lewis, H.C. *et al.* Outbreaks of Shigella sonnei infections in Denmark and
491 Australia linked to consumption of imported raw baby corn. *Epidemiol Infect* **137**,
492 326-34 (2009).
493 32. Cohen, D. *et al.* Reduction of transmission of shigellosis by control of
494 houseflies (Musca domestica). *Lancet* **337**, 993-7 (1991).
495 33. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-
496 Wheeler transform. *Bioinformatics* **25**, 1754-60 (2009).
497 34. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools.
498 *Bioinformatics* **25**, 2078-9 (2009).
499 35. Stamatakis, A. RAxML-VI-HPC: maximum likelihood-based phylogenetic
500 analyses with thousands of taxa and mixed models. *Bioinformatics* **22**, 2688-90
501 (2006).
502 36. Tang, J., Hanage, W.P., Fraser, C. & Corander, J. Identifying currents in the
503 gene pool for bacterial populations using an integrative approach. *PLoS Comp Biol* **5**,
504 e1000455 (2009).
505 37. Paradis, E., Claude, J. & Strimmer, K. APE: Analyses of Phylogenetics and
506 Evolution in R language. *Bioinformatics* **20**, 289-90 (2004).

38. Zerbino, D.R. & Birney, E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* **18**, 821-9 (2008).

39. Kurtz, S. *et al.* Versatile and open software for comparing large genomes. *Genome Biol* **5**, R12 (2004).

40. Aziz, R.K. *et al.* The RAST Server: rapid annotations using subsystems technology. *BMC Genomics* **9**, 75 (2008).

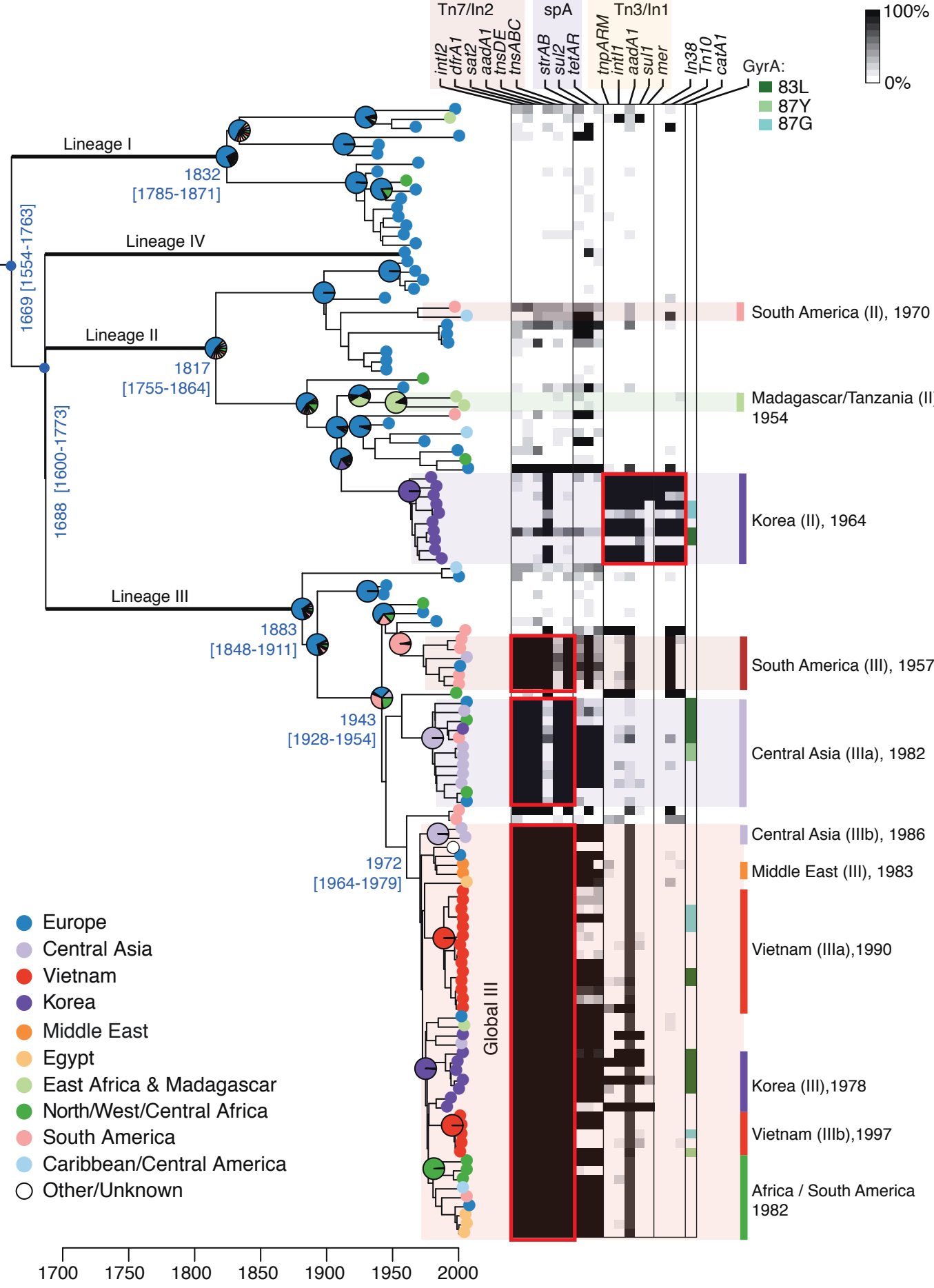
41. Rice, P., Longden, I. & Bleasby, A. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet* **16**, 276-7 (2000).

42. Croucher, N.J. *et al.* Rapid pneumococcal evolution in response to clinical interventions. *Science* **331**, 430-4 (2011).

Editorial Summary (AOP and Month, same):

Nicholas Thomson and colleagues report whole-genome sequencing of 132 globally distributed isolates of *Shigella sonnei*, a cause of human dysentery. Their phylogeographic analyses suggest that the current *S. sonnei* population is under 500 years old, and the authors are able to trace several distinct lineages that have spread out of Europe to other continents over the last few decades.

Type of file: figure
Label: 1
Filename: figure_1.ai


Tn7/In2
spA
Tn3/In1
intI2
dfrA1
sat2
aadA1
tnsDE
tnsABC
strAB
sul2
tetAR
tnpARM
intI1
aadA1
sul1
mer
In38
Tn10
catA1
100%
0%
GyrA:
83L
87Y
87G
Lineage I
1832
[1785-1871]
1669 [1554-1763]
Lineage IV
Lineage II
1817
[1755-1864]
1688 [1600-1773]
Lineage III
1883
[1848-1911]
1943
[1928-1954]
1972
[1964-1979]
Global III
South America (II), 1970
Madagascar/Tanzania (II
1954
Korea (II), 1964
South America (III), 1957
Central Asia (IIIa), 1982
Central Asia (IIIb), 1986
Middle East (III), 1983
Vietnam (IIIa),1990
Korea (III),1978
Vietnam (IIIb),1997
Africa / South America
1982
Europe
Central Asia
Vietnam
Korea
Middle East
Egypt
East Africa & Madagascar
North/West/Central Africa
South America
Caribbean/Central America
Other/Unknown
1700
1750
1800
1850
1900
1950
2000

UKPMC+ has received the file 'supp_info_1.pdf' as supplementary data. The file will not appear in this PDF Receipt, but it will be linked to the web version of your manuscript.

UKPMC+ has received the file 'supp table 1.xls' as supplementary data. The file will not appear in this PDF Receipt, but it will be linked to the web version of your manuscript.