nature genetics

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]

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, seeming to replace the more diverse** ***Shigella flexneri*** **in areas undergoing economic development and improvements in water quality[4–6]. Classical approaches have shown that** ***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 that diversified into several distinct lineages with unique characteristics. Our analysis suggests that the majority of this diversification occurred in Europe and was followed by more recent establishment of local pathogen populations on 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* that were isolated between 1943 and 2008 and spanned four continents (**Supplementary Table 1**). The *S. sonnei* genome is made up of a single circular chromosome and a plasmid, pINV B (ref. 2). We detected 10,111 chromosomal SNPs randomly distributed across the *S. sonnei* chromosome, with approximately 1 per 430 bp (0.23% nucleotide divergence) (**Supplementary Fig. 1**). To investigate the population structure of *S. sonnei*, we analyzed these chromosomal SNPs using multiple phylogenetic methods. Maximum-likelihood phylogenetic analysis (**Supplementary Fig. 2**), rooted using *Shigella* and *E. coli* outgroups (**Supplementary Table 2**), revealed a strong correlation between root-to-tip branch lengths and the known dates of isolation for the sequenced *S. sonnei*, which is indicative of rapid clock-like evolution in which substitution mutations occur at a regular rate (**Supplementary Fig. 3**). There seemed to be some variation in this mutation rate 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. Notably, this yielded the same tree topology as was generated with the maximum-likelihood analysis while also providing estimates of nucleotide substitution rates and divergence times for key *S. sonnei* lineages (**Fig. 1**). The phylogenies included four distinct *S. sonnei* lineages—three encompassing isolates spanning the 1940s through the 2000s and one comprising a single isolate from France. These lineages each had 100% maximum-likelihood bootstrap support and 100% Bayesian posterior support (BEAST) and were also recovered using a Bayesian clustering analysis (Online Methods). Although these lineages are uniquely characterized by hundreds of SNPs, they show 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 Fig. 2**, **Supplementary Table 3** and **Supplementary Note**). We estimated a mean substitution rate of 2.0×10^{-4} substitutions per site per year at the 10,111 chromosomal SNP loci (95% highest posterior density (HPD) 1.6×10^{-4} to 2.3×10^{-4}), corresponding to the accumulation of approximately 2.2 SNPs per chromosome per year (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 per site per year (95% HPD 5.2×10^{-7} to 6.7×10^{-7}), which likely represents

[1]Department of Microbiology and Immunology, University of Melbourne, Melbourne, Victoria, Australia. [2]The Hospital for Tropical Diseases, Wellcome Trust Major Overseas Programme, Oxford University Clinical Research Unit, Ho Chi Minh City, Vietnam. [3]Unité des Bactéries Pathogènes Entériques, Institut Pasteur, Paris, France. [4]Center for Infectious Disease Dynamics, Department of Biology, The Pennsylvania State University, University Park, Pennsylvania, USA. [5]Fogarty International Center, US National Institutes of Health, Bethesda, Maryland, USA. [6]Strathclyde Institute of Pharmacy and Biomedical Sciences, University of Strathclyde, Glasgow, UK. [7]Scottish Salmonella, Shigella and Clostridium difficile Reference Laboratory, Stobhill Hospital, 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, Republic of Korea. [10]Department of Genetics, Evolution and Bioagents, Biology Institute, Campinas State University (UNICAMP), Campinas, Brazil. [11]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK. Correspondence should be addressed to K.E.H. (kholt@unimelb.edu.au) or N.R.T. (nrt@sanger.ac.uk).

Received 31 January; accepted 5 July; published online 5 August 2012; doi:10.1038/ng.2369

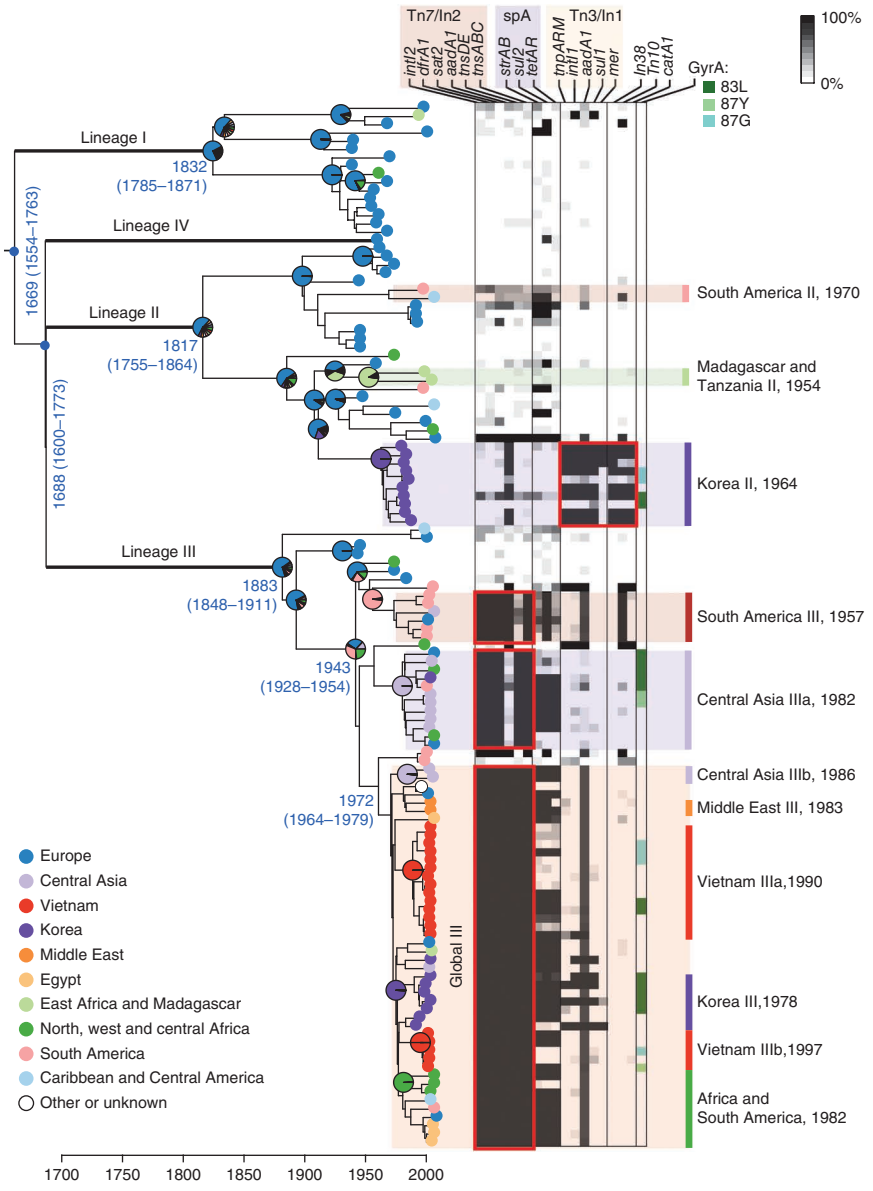

Figure 1 Bayesian maximum clade credibility phylogeny for *S. sonnei*. Left, branches defining major lineages are shown in bold (each with 100% posterior support). Pie charts indicate maximum-likelihood estimates for geographic origin of major nodes (legend at bottom). Divergence dates (median estimates and 95% HPD) are given in blue for major nodes. Right, the distribution of antimicrobial resistance determinants is indicated in the heatmap, reflecting the percentage of bases in each gene sequence that are covered by reads from each isolate (legend at top). GyrA mutations 83L, 87Y and 87G are indicated by color (inset legend). Likely MDR acquisition events are highlighted with red boxes. Geographically localized clonal expansions are highlighted and labeled with their median estimated divergence dates.

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} substitutions per site per year)[11] but lies between the rates estimated for *Yersinia pestis* (2×10^{-8}) (ref. 12) and *Staphylococcus aureus* (3×10^{-6}) (ref. 13). From BEAST analysis, we estimated that the most recent common ancestor (MRCA) of all contemporary *S. sonnei* existed less than 500 years ago (median calendar year for divergence date, 1669 C.E.; 95% HPD 1554–1763) (**Fig. 1**). Similarly, we estimated that the MRCA for each group in lineages I and II existed in the early nineteenth century and that all lineage III isolates descend from a hypothetical ancestor that existed approximately at the turn of the twentieth century (**Fig. 1**). Critically, these data indicate that, although the extant *S. sonnei* population descends from a single ancestor existing in the seventeenth century, by the late nineteenth 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 the association between phylogeny and geographic region)[14]. Notably, the European population shows the richest diversity, with isolates distributed across all four lineages (31% in lineage I, 35% in lineage II, 31% in lineage III and the sole lineage IV isolate) and occupying basal branches in each lineage (**Fig. 1**). In contrast, *S. sonnei* isolates from Asia, Africa and South or Central America were mainly from lineage III (70%, 67% and 73%, respectively), with fewer lineage II representatives (25%, 22% and 27%, respectively) and just two from lineage I (both from Africa). Furthermore, ancestral state reconstruction analysis indicated a >50% likelihood of a European common ancestor for each of lineages I, II and III (**Fig. 1**). The data also indicate that lineage III has been more successful at global dispersal than the other lineages, with only low numbers of lineage I or II detected outside of Europe (**Fig. 1**). A recently derived clade within lineage III (Global III, MRCA from 1972 (95% HPD 1964–1979)) has been particularly successful at global dissemination, comprising 49% of all isolates sampled since 1995 and detected in all regions represented in our collection (**Fig. 1**). Unlike the European isolates, isolates from non-European countries form tight, shallow-rooted phylogenetic clusters, consistent with and suggestive of contemporary dispersal (**Fig. 1**). In many cases, these clusters contain multiple isolates from the same country, indicating localized clonal expansions. For example, isolates from Korea formed two subclades within lineages II and III that likely represent separate introductions of *S. sonnei* into Korea during the 1960s and 1970s, each of which were followed by local clonal expansions. Similarly, isolates originating in Vietnam form two subclades, indicating that lineage III clones were established locally in Vietnam in the 1990s. At a regional level, a lineage III subclade seems to have been established in South America during the 1950s, to which isolates from Brazil and Peru can be traced, and this was followed by the dissemination of the Global III clade into Africa and South and Central America in the early 1980s.

Critically, the phylogeographic analysis indicates that all contemporary *S. sonnei* infections are caused by a small number of clones that have recently become globally dispersed (**Fig. 1**). The distribution of antimicrobial resistance genes and mutations within the

S. sonnei phylogeny suggests that selection for multiple drug resistance (MDR) had a pivotal role in driving this global dissemination (**Fig. 1**, **Supplementary Fig. 2** and **Supplementary Table 1**). In particular, the establishment of local *S. sonnei* lineage III populations outside of Europe is intimately associated with the carriage of transposon Tn*7* and class II integrons (In*2*) encoding resistance to multiple antimicrobials (**Fig. 1**). All three major lineage III subgroups carry a distinct In*2* variant, which is either encoded on a plasmid (South America III) or integrated into the chromosome adjacent to *glmS* (Central Asia IIIa and Global III), suggesting that independent acquisitions of the integron in each group during the 1960s and 1970s was followed by clonal expansion and subsequent international spread (**Fig. 1**). Studies from Europe, Asia, Africa, South America and Australia have reported a high prevalence of In*2*-bearing, MDR biotype g *S. sonnei*, often associated with local epidemics[15]. Our data show that biotype g classification is a marker for lineage III due to the presence of a conserved nonsense mutation in the rhamnose operon regulatory gene *rhaR* (**Supplementary Fig. 2**) and indicate that the global distribution of MDR biotype g, In*2*-bearing *S. sonnei* is the result of global dissemination of multiple In*2*-bearing subclades of lineage III *S. sonnei*. Half of the In*2*-bearing lineage III isolates also harbored the small MDR plasmid spA[2] containing the *tetAR*, *strAB* and *sul2* genes, which confer additional resistance to tetracycline, streptomycin and sulfonamides, respectively (**Fig. 1**). All quinolone-resistant isolates harbored one of three point mutations in the chromosomal *gyrA* gene (encoding DNA gyrase), known to confer quinolone resistance (**Fig. 1** and **Supplementary Table 1**; we detected no plasmid-mediated quinolone resistance genes). The distribution of *gyrA* mutations within the phylogeny shows that these resistance-conferring mutations have arisen independently on at least nine occasions in our *S. sonnei* collection, including the introduction of two separate mutations within the clonal Korea II group, which is indicative of unexpectedly strong selection for quinolone resistance, even in MDR isolates (**Fig. 1**). To investigate other signals of selection, we examined the clustering of SNPs within genes and chromosomal regions (**Supplementary Note**). We found evidence of phage and transposase insertions and a single case of homologous recombination affecting the *sitABCD* operon in isolate 31382, but we identified only two genes with mutations resulting in amino-acid variation that was significantly higher than expected under a random distribution of SNPs (adjusted P value < 0.05; see Online Methods). Neither of these genes (*rpoS* and *mreB*) encodes an extracellular protein, suggesting a lack of immune selection, in common with another human-restricted pathogen, *Salmonella* Typhi (which causes typhoid fever)[16]. However, we detected a large number of nonsynonymous SNPs and a high rate of nonsynonymous-to-synonymous substitutions per site (d_N/d_S) in the *acrD* (8 nonsynonymous SNPs; $d_N/d_S = 2.5$) and *acrB* (12 nonsynonymous SNPs; $d_N/d_S = 1.8$) genes encoding drug-efflux pumps. Currently, antimicrobial treatment is recommended for the management of dysentery[17] but may not substantially affect the resolution of *S. sonnei* or *S. flexneri* infections[18,19]. However, there is evidence that such treatment can prevent shedding of *S. sonnei* after the resolution of symptoms[20]. Thus, whereas antimicrobial resistance may have only minor implications for dysentery treatment, this phenotype may be important in sustaining *S. sonnei* transmission within human populations, and our data indicate that there is a strong selective pressure for its maintenance. It has been hypothesized that free-living amoebae may represent an environmental reservoir for *Shigella*, which are able to survive intracellularly in *Acanthamoeba*[21,22]. This could potentially provide another niche in which selective pressure for antibiotic resistance may be exerted, although intracellular *Shigella* are likely to be protected from most antibiotics by their amoebae hosts[23,24].

Previous studies have proposed that the acquisition of virulence plasmid pINV B, encoding the *Plesiomonas shigelloides*–related O antigen, was the defining event in the emergence of *S. sonnei*[25]. Unfortunately, the *S. sonnei* virulence plasmid is highly unstable in laboratory media and is commonly lost during subculturing[26]; as a consequence of this, less than half of our isolates yielded sufficient virulence plasmid sequence data for analysis (46 isolates with >10× read depth). Phylogenetic analysis of the available virulence plasmid sequences (of 84 SNPs) identified 3 distinct lineages (**Supplementary Fig. 4**). There was a close correspondence 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 after 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 geographic expansion of *S. sonnei* will be characterized by the introduction and expansion of novel clones moving into human populations with falling natural immunity that was previously obtained from exposure to *P. shigelloides*. Our model of recent dissemination out of Europe is very 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 up 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 the incidence of dysentery caused by *S. sonnei* in transitional countries. The combination of increased incidence and excessive antimicrobial resistance among globally disseminated *S. sonnei* populations indicates that a vaccine for *S. sonnei* 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, as all *S. sonnei* share a single O antigen that has proven to be a successful vaccine target[29]. Notably, the success of *S. sonnei* in the face of diminishing *S. flexneri* incidence suggests important epidemiological distinctions in the 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. TreeStat, http://tree.bio.ed.ac.uk/software/treestat/; Velvet, http://www.ebi.ac.uk/~zerbino/velvet/.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. The finished genome of *S. sonnei* 53G has been deposited at the European Molecular Biology Laboratory (EMBL) under accessions HE616528 (chromosome) and HE616529, HE616530, HE616531 and HE616532 (plasmids). Sequence reads for the 132 Illumina-sequenced *S. sonnei* have been deposited in the European Nucleotide Archive under accession ERP000182.

Note: Supplementary information is available in the online version of the paper.

ACKNOWLEDGMENTS

We thank M. Levine (University of Maryland School of Medicine) and C. Tang (University of Oxford) for their kind gift of *S. sonnei* strain 53G. 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). K.E.H. was supported by a Fellowship from the National Health & Medical Research Council (NHMRC) of Australia (628930); S.B. is supported by an Oak Foundation Fellowship through Oxford University (OAKF9) and by the Li Ka Shing foundation (LG13); F.X.W. was partially funded by the Institut de Veille Sanitaire; J.Y. was supported by a UK Medical Research Council (MRC) grant (G0800173); and D.W.K. was partially supported by grant RTI05-01-01 from the Korean Ministry of Knowledge and Economy (MKE).

AUTHOR CONTRIBUTIONS

K.E.H., N.R.T., E.C.H. and A.K. analyzed the data and performed phylogenetic analysis. N.R.T., G.D., J.Y., S.B., J.J.F., K.E.H. and J.P. were involved in the study design. F.-X.W., D.J.B., J.E.C., J.Y., V.S., D.W.K., S.Y.C., S.H.K., W.D.d.S. and D.J.P. were involved in isolate collection, DNA analysis and resistance phenotyping. K.E.H., S.B., N.R.T., G.D., A.K., E.C.H. and F.-X.W. contributed to manuscript writing.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/doifinder/10.1038/ng.2369.
Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

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. USA* **97**, 10567–10572 (2000).
2. Yang, F. *et al.* Genome dynamics and diversity of *Shigella* species, the etiologic agents of bacillary dysentery. *Nucleic Acids Res.* **33**, 6445–6458 (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**, 1126–1128 (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**, 651–666 (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**, 1413–1415 (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**, 2460–2467 (2011).
11. Mutreja, A. *et al.* Evidence for several waves of global transmission in the seventh cholera pandemic. *Nature* **477**, 462–465 (2011).
12. Morelli, G. *et al. Yersinia pestis* genome sequencing identifies patterns of global phylogenetic diversity. *Nat. Genet.* **42**, 1140–1143 (2010).
13. Harris, S.R. *et al.* Evolution of MRSA during hospital transmission and intercontinental spread. *Science* **327**, 469–474 (2010).
14. Parker, J., Rambaut, A. & Pybus, O.G. Correlating viral phenotypes with phylogeny: accounting for phylogenetic uncertainty. *Infect. Genet. Evol.* **8**, 239–246 (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**, 987–993 (2008).
17. World Health Organization. *Guidelines for the Control of Shigellosis, Including Epidemics Due to Shigella dysenteriae Type 1* (WHO Document Production Services, Geneva, 2005).
18. Christopher, P.R., David, K.V., John, S.M. & Sankarapandian, V. Antibiotic therapy for *Shigella* dysentery. *Cochrane Database Syst. Rev.* CD006784 (2010).
19. Vinh, H. *et al.* A multi-center randomized trial to assess the efficacy of gatifloxacin versus ciprofloxacin for the treatment of shigellosis in Vietnamese children. *PLoS Negl. Trop. Dis.* **5**, e1264 (2011).
20. Vinh, H. *et al.* Treatment of bacillary dysentery in Vietnamese children: two doses of ofloxacin versus 5-days nalidixic acid. *Trans. R. Soc. Trop. Med. Hyg.* **94**, 323–326 (2000).
21. Jeong, H.J. *et al. Acanthamoeba*: could it be an environmental host of *Shigella*? *Exp. Parasitol.* **115**, 181–186 (2007).
22. Saeed, A., Abd, H., Edvinsson, B. & Sandstrom, G. *Acanthamoeba castellanii* an environmental host for *Shigella dysenteriae* and *Shigella sonnei*. *Arch. Microbiol.* **191**, 83–88 (2009).
23. Winiecka-Krusnell, J. & Linder, E. Free-living amoebae protecting *Legionella* in water: the tip of an iceberg? *Scand. J. Infect. Dis.* **31**, 383–385 (1999).
24. Greub, G. & Raoult, D. Microorganisms resistant to free-living amoebae. *Clin. Microbiol. Rev.* **17**, 413–433 (2004).
25. Shepherd, J.G., Wang, L. & Reeves, P.R. Comparison of O-antigen gene clusters of *Escherichia coli* (*Shigella*) sonnei and *Plesiomonas shigelloides* O17: sonnei gained its current plasmid-borne O-antigen genes from *P. shigelloides* in a recent event. *Infect. Immun.* **68**, 6056–6061 (2000).
26. Sansonetti, P.J., Kopecko, D.J. & Formal, S.B. *Shigella sonnei* plasmids: evidence that a large plasmid is necessary for virulence. *Infect. Immun.* **34**, 75–83 (1981).
27. Van de Verg, L.L., Herrington, D.A., Boslego, J., Lindberg, A.A. & Levine, M.M. Age-specific prevalence of serum antibodies to the invasion plasmid and lipopolysaccharide antigens of *Shigella* species in Chilean and North American populations. *J. Infect. Dis.* **166**, 158–161 (1992).
28. Shimada, T. & Sakazaki, R. On the serology of *Plesiomonas shigelloides*. *Jpn. J. Med. Sci. Biol.* **31**, 135–142 (1978).
29. Kaminski, R.W. & Oaks, E.V. Inactivated and subunit vaccines to prevent shigellosis. *Expert Rev. Vaccines* **8**, 1693–1704 (2009).
30. Genobile, D. *et al.* An outbreak of shigellosis in a child care centre. *Commun. Dis. Intell.* **28**, 225–229 (2004).
31. Lewis, H.C. *et al.* Outbreaks of *Shigella sonnei* infections in Denmark and Australia linked to consumption of imported raw baby corn. *Epidemiol. Infect.* **137**, 326–334 (2009).
32. Cohen, D. *et al.* Reduction of transmission of shigellosis by control of houseflies (*Musca domestica*). *Lancet* **337**, 993–997 (1991).

ONLINE METHODS

Bacterial isolates and sequencing. Bacterial isolates analyzed in this study are detailed in **Supplementary Table 1**. DNA was prepared using the Wizard Genomic DNA kit (Promega) or phenol extraction. Index-tagged paired-end Illumina sequencing libraries were prepared using 1 of 12 unique indexing tags, as previously described[13]. These were combined into pools, with each pool containing 11 or 12 uniquely tagged libraries, which were sequenced on the Illumina Genome Analyzer GAII according to the 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 and NC_009345; plasmid pEG356: NC_013727) using Burrows-Wheeler Aligner (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 (chromosome: HE616528; plasmids: HE616529, HE616530, HE616531 and HE616532) were identified using the same mapping procedure to analyze reads simulated from the finished genome using the wgsim algorithm in SAMtools. 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 SNPs. 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 of ≤20, read depth of ≤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 identical tree topology with near-identical branch lengths (Pearson correlation coefficient = 0.995, $P < 1 \times 10^{-15}$), showing the robustness of the method and its independence from the choice of reference genome. The Ss046-mapped data were used for all analyses reported, as the Ss046 genome has been widely used in previous comparative studies, whereas 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 ≥10× and 137 kb 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 analyze 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 were removed from the alignment. The resulting alignment was further filtered to remove loci at which alleles were unknown for >40% of isolates (indicating that the site is not conserved), and a maximum-likelihood phylogeny was estimated using RAxML[35]. The BEAST package[8] was used for the Bayesian inference of phylogeny and divergence dates. Additionally, we used the BAPS program (Bayesian Analysis of Population Structure)[36] to examine clustering of the isolates on the basis of SNP data.

For maximum-likelihood analysis, RAxML was run ten times using the generalized time-reversible model with a Γ distribution to model site-specific rate variation (the GTR+Γ substitution model; GTRGAMMA in RAxML). We performed 1,000 bootstrap pseudo-replicate analyses to assess support for the maximum-likelihood phylogeny. The final result (**Supplementary Fig. 2**) is the tree with the highest likelihood across all ten runs, with maximum-likelihood estimates of branch length and confidence in major bipartitions calculated using the bootstrap values across all runs. This phylogeny was rooted using *E. coli* and *Shigella* outgroups (**Supplementary Table 2**).

Root-to-tip branches were extracted from the maximum-likelihood tree using the program TreeStat (see URLs). The relationships between root-to-tip distances, year of isolation and lineage were analyzed using linear regression. Plots and regression lines are shown in **Supplementary Figure 3**, along with Pearson correlation coefficients.

For BEAST analysis, we also used the GTR+Γ substitution model and defined tip dates as the year of isolation (restricting the analysis to those sequences with recorded dates). We performed multiple analyses using both constant size and Bayesian skyline demographic models in combination with either a strict molecular clock or a relaxed clock (uncorrelated log-normal 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 10 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 after burn-in, with all reported parameter estimates (medians and 95% HPDs) calculated using the Tracer v1.5 program. 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 models (Bayesian skyline) but with randomized tip dates (**Supplementary Fig. 5**). This randomization procedure produces a null set of tip-date and sequence correlations that may be analyzed to produce null substitution rate distributions, which can then be compared with empirical rate estimates.

Phylogeographic analysis. The geographic region of isolation of each *S. sonnei* was analyzed as a discrete character trait using two complementary methods. Phylogeographic analyses were performed using the 126 isolates that had complete information for both year and geographic region of isolation (**Supplementary Table 1**). First, the association between the phylogenetic relationships of *S. sonnei* isolates (inferred by BEAST) and their geographic region of isolation were tested using Bayesian Tip-association Significance (BaTS) software[14]. A random selection of 50,000 trees sampled during Bayesian phylogenetic analysis were used as input, and 1,000 randomizations were used to generate a null distribution for significance testing. Second, ancestral state reconstruction of the geographic origin of hypothetical common ancestors (internal nodes in the phylogeny) was performed using the ace function implemented in the ape package for R[37]. The percent probability estimates quoted and shown in pie charts in **Figure 1** are scaled likelihoods for the discrete character trait (region of isolation) at each node.

Gene content analysis. Each read set was assembled using the *de novo* short-read assembler Velvet[38] and Velvet Optimiser (see URLs). Contigs less than 100 bp in size were excluded from further analysis. The *S. sonnei* 53G genome and *de novo*–assembled contig sets were mapped iteratively to the pan-genome reference set (initialized as the concatenation of *S. sonnei* Ss046 chromosome and plasmids) using MUMmer (nucmer algorithm)[39]. At each iteration i, sequences not aligning to the current pan-genome P_{i-1} set were incorporated into an extended pan-genome P_i. The final pan-genome P was annotated using a combination of annotation transfer (for *S. sonnei* reference sequences) and *de novo* annotation using the RAST annotation server[40] for novel sequences assembled from reads. The latter included 1.67 Mb of sequence in 862 contigs in which 2,422 genes were annotated (incorporating 80.5% of bases), resulting in a total of 6,852 genes.

S. sonnei read sets were then aligned to the pan-genome using BWA[27] with default mapping parameters. A pileup was generated for each aligned read set using SAMtools[28] and was used to summarize, for each annotated gene in the pan-genome P, the coverage (percentage of bases covered) and presence of inactivating mutations (nonsense SNPs or non-triplet insertions and/or deletions (indels) resulting in frameshifts) in each genome. The results were used to identify genes whose presence or inactivation was associated with specific lineages (**Supplementary Fig. 6** and **Supplementary Note**).

doi:10.1038/ng.2369

Resistance gene analysis. The presence of resistance-conferring genes was initially determined from mapping data. The genetic context of resistance genes was examined by blastn search of each contig set with known resistance, transposase or integrase genes as query sequences. The resulting contigs were compared to the NCBI non-redundant nucleotide database to annotate the resistance genes and mobile elements. Mapping was then repeated using annotated mobile elements to generate the gene coverage maps shown in **Figure 1** and **Supplementary Figure 2**, which indicate the proportion of bases in each gene sequence that are covered by reads from each isolate (reference sequences are provided in **Supplementary Fig. 2**).

33. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
34. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
35. Stamatakis, A. RAxML-VI-HPC: maximum likelihood–based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**, 2688–2690 (2006).
36. Tang, J., Hanage, W.P., Fraser, C. & Corander, J. Identifying currents in the gene pool for bacterial populations using an integrative approach. *PLOS Comput. Biol.* **5**, e1000455 (2009).
37. Paradis, E., Claude, J. & Strimmer, K. APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* **20**, 289–290 (2004).
38. Zerbino, D.R. & Birney, E. Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Res.* **18**, 821–829 (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).