



**HAL**  
open science

## Global phylogeography and evolutionary history of *Shigella dysenteriae* type 1.

Elisabeth Njamkepo, Nizar Fawal, Alicia Tran-Dien, Jane Hawkey, Nancy Strockbine, Claire Jenkins, Kaiser A Talukder, Raymond Bercion, Konstantin Kuleshov, Renáta Kolínská, et al.

► **To cite this version:**

Elisabeth Njamkepo, Nizar Fawal, Alicia Tran-Dien, Jane Hawkey, Nancy Strockbine, et al.. Global phylogeography and evolutionary history of *Shigella dysenteriae* type 1.. *Nature Microbiology*, 2016, 1 (4), pp.16027. 10.1038/NMICROBIOL.2016.27 . pasteur-01422023

**HAL Id: pasteur-01422023**

**<https://pasteur.hal.science/pasteur-01422023v1>**

Submitted on 12 Mar 2019

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

1 **Global phylogeography and evolutionary history of *Shigella dysenteriae* type 1**

2

3 Elisabeth Njamkepo<sup>1\*</sup>, Nizar Fawal<sup>1\*</sup>, Alicia Tran-Dien<sup>1\*</sup>, Jane Hawkey<sup>2,3,4</sup>, Nancy

4 Strockbine<sup>5</sup>, Claire Jenkins<sup>6</sup>, Kaisar A. Talukder<sup>7</sup>, Raymond Bercion<sup>8,9</sup>, Konstantin

5 Kuleshov<sup>10</sup>, Renáta Kolínská<sup>11</sup>, Julie E. Russell<sup>12</sup>, Lidia Kaftyreva<sup>13</sup>, Marie Accou-

6 Demartin<sup>1</sup>, Andreas Karas<sup>14</sup>, Olivier Vandenberg<sup>15,16</sup>, Alison E. Mather<sup>17,18</sup>, Carl J.

7 Mason<sup>19</sup>, Andrew J. Page<sup>17</sup>, Thandavarayan Ramamurthy<sup>20</sup>, Chantal Bizet<sup>21</sup>, Andrzej

8 Gamian<sup>22</sup>, Isabelle Carle<sup>1</sup>, Amy Gassama Sow<sup>9</sup>, Christiane Bouchier<sup>23</sup>, Astrid Louise

9 Wester<sup>24</sup>, Monique Lejay-Collin<sup>1</sup>, Marie-Christine Fonkoua<sup>25</sup>, Simon Le Hello<sup>1</sup>, Martin J.

10 Blaser<sup>26</sup>, Cecilia Jernberg<sup>27</sup>, Corinne Ruckly<sup>1</sup>, Audrey Mérens<sup>28</sup>, Anne-Laure Page<sup>29</sup>,

11 Martin Aslett<sup>17</sup>, Peter Roggentin<sup>30</sup>, Angelika Fruth<sup>31</sup>, Erick Denamur<sup>32</sup>, Malabi

12 Venkatesan<sup>33</sup>, Hervé Bercovier<sup>34</sup>, Ladaporn Bodhidatta<sup>19</sup>, Chien-Shun Chiou<sup>35</sup>,

13 Dominique Clermont<sup>21</sup>, Bianca Colonna<sup>36</sup>, Svetlana Egorova<sup>13</sup>, Gururaja P. Pazhani<sup>20</sup>,

14 Analia V. Ezernitchi<sup>37</sup>, Ghislaine Guigon<sup>38</sup>, Simon R. Harris<sup>17</sup>, Hidemasa Izumiya<sup>39</sup>,

15 Agnieszka Korzeniowska-Kowal<sup>22</sup>, Anna Lutyńska<sup>40</sup>, Malika Gouali<sup>1</sup>, Francine

16 Grimont<sup>1</sup>, Céline Langendorf<sup>29</sup>, Monika Marejková<sup>41</sup>, Lorea A. M. Peterson<sup>42</sup>, Guillermo

17 Perez-Perez<sup>26</sup>, Antoinette Ngandjio<sup>25</sup>, Alexander Podkolzin<sup>10</sup>, Erika Souche<sup>43</sup>, Mariia

18 Makarova<sup>13</sup>, German A. Shipulin<sup>10</sup>, Changyun Ye<sup>44</sup>, Helena Žemličková<sup>11,45</sup>, Mária

19 Herpay<sup>46</sup>, Patrick A.D. Grimont<sup>1</sup>, Julian Parkhill<sup>17</sup>, Philippe Sansonetti<sup>47</sup>, Kathryn E.

20 Holt<sup>2,3</sup>, Sylvain Brisse<sup>38,48,49</sup>, Nicholas R. Thomson<sup>17,50</sup>, François-Xavier Weill<sup>1,17‡</sup>

21 \* Contributed equally

22 ‡ Corresponding author

23

24

- 25 1. Institut Pasteur, Unité des Bactéries Pathogènes Entériques, 75724 Paris Cedex  
26 15, France
- 27 2. Centre for Systems Genomics, University of Melbourne, Parkville, VIC 3010,  
28 Australia
- 29 3. Department of Biochemistry and Molecular Biology, Bio21 Molecular Science  
30 and Biotechnology Institute, University of Melbourne, Parkville, VIC 3010,  
31 Australia
- 32 4. School of Agriculture and Veterinary Science, University of Melbourne,  
33 Parkville, VIC 3010, Australia
- 34 5. Centers for Disease Control and Prevention, *Escherichia* and *Shigella* Reference  
35 Unit, Atlanta, GA 30333, United States of America
- 36 6. Public Health England, Gastrointestinal Bacteria Reference Unit, Colindale, NW9  
37 5HT, United Kingdom
- 38 7. icddr,b, Enteric and Food Microbiology Laboratory, Dhaka 1212, Bangladesh
- 39 8. Institut Pasteur de Bangui, BP 923, Bangui, République Centrafricaine
- 40 9. Institut Pasteur de Dakar, BP 220, Dakar, Senegal
- 41 10. Federal Budget Institute of Science, Central Research Institute for Epidemiology,  
42 Moscow 111123, Russian Federation
- 43 11. Czech National Collection of Type Cultures (CNCTC), National Institute of  
44 Public Health, Prague 10, Czech Republic
- 45 12. Public Health England, National Collection of Type Cultures, Porton Down, SP4  
46 0JG, United Kingdom
- 47 13. Pasteur Institute of St Petersburg, St Petersburg 197101, Russian Federation
- 48 14. Department of Medical Microbiology, University of KwaZulu-Natal, Durban  
49 4041, South Africa
- 50 15. Department of Microbiology, LHUB-ULB, Brussels University Hospitals  
51 Laboratory, 1000 Brussels, Belgium
- 52 16. Environmental Health Research Centre, Public Health School, Université Libre  
53 de Bruxelles, 1070 Brussels, Belgium
- 54 17. Wellcome Trust Sanger Institute, Cambridge, CB10 1SA, United Kingdom
- 55 18. Department of Veterinary Medicine, University of Cambridge, Cambridge, CB3  
56 0ES, United Kingdom
- 57 19. Armed Forces Research Institute of Medical Sciences (AFRIMS), Bangkok  
58 10400, Thailand
- 59 20. National Institute of Cholera and Enteric Diseases (NICED), Kolkata, West  
60 Bengal 700010, India
- 61 21. Institut Pasteur, Collection de l'Institut Pasteur (CIP), 75724 Paris Cedex 15,  
62 France
- 63 22. Polish Collection of Microorganisms, Institute of Immunology and Experimental  
64 Therapy, 53-114 Wroclaw, Poland
- 65 23. Institut Pasteur, Plate-forme Génomique (PF1), 75724 Paris Cedex 15, France
- 66 24. Department of Foodborne Infections, Norwegian Institute of Public Health,  
67 Nydalen 0403, Oslo, Norway
- 68 25. Centre Pasteur du Cameroun, BP 1274, Yaoundé, Cameroon
- 69 26. Departments of Medicine and Microbiology, New York University Langone  
70 Medical Center, New York, New York 10016, United States of America

- 71 27. Department of Diagnostics and Vaccinology, Public Health Agency of Sweden,  
72 17182 Solna, Sweden
- 73 28. Biology Department and Infection Control Unit, Bégin Military Hospital, 94160  
74 Saint-Mandé, France.
- 75 29. Epicentre, 75011 Paris, France
- 76 30. Institut für Hygiene und Umwelt, 20539 Hamburg, Germany
- 77 31. Division of Enteropathogenic Bacteria and Legionella, Robert Koch Institut,  
78 38855 Wernigerode, Germany
- 79 32. INSERM, IAME, UMR 1137; Univ Paris Diderot, IAME, UMR 1137, Sorbonne  
80 Paris Cité, 75018 Paris, France
- 81 33. Bacterial Diseases Branch, Walter Reed Army Institute of Research, Silver  
82 Spring, MD 20910, United States of America
- 83 34. Faculty of Medicine, Hebrew University of Jerusalem, Jerusalem 91120, Israel
- 84 35. Center of Research and Diagnostics, Centers for Disease Control, Taichung  
85 40855, Taiwan
- 86 36. Istituto Pasteur-Fondazione Cenci Bolognetti, Dipartimento di Biologia e  
87 Biotecnologie C Darwin, Sapienza Università di Roma, 00185, Roma, Italy
- 88 37. Central Laboratories, Ministry of Health, Jerusalem 91342, Israel
- 89 38. Institut Pasteur, Genotyping of Pathogens and Public Health Platform, 75724  
90 Paris Cedex 15, France
- 91 39. Department of Bacteriology I, National Institute of Infectious Diseases, Tokyo,  
92 162-8640, Japan
- 93 40. Department of Sera and Vaccines Evaluation, National Institute of Public Health–  
94 National Institute of Hygiene, 00-791 Warsaw, Poland
- 95 41. National Reference Laboratory for *E. coli* and *Shigella*, National Institute of  
96 Public Health, Prague 10, Czech Republic
- 97 42. National Microbiology Laboratory, Public Health Agency of Canada, Winnipeg,  
98 Manitoba R3E 3R2, Canada
- 99 43. Institut Pasteur, Bioinformatics platform, 75724 Paris Cedex 15, France
- 100 44. State Key Laboratory of Infectious Disease Prevention and Control, National  
101 Institute for Communicable Disease Control and Prevention, China CDC, Beijing  
102 102206, China
- 103 45. Department of Clinical Microbiology, Faculty of Medicine and University  
104 Hospital, Charles University, 500 05, Hradec Kralove, Czech Republic
- 105 46. Hungarian National Collection of Medical Bacteria, National Center for  
106 Epidemiology, H-1097 Budapest, Hungary
- 107 47. Institut Pasteur, Unité de Pathogénie Microbienne Moléculaire, 75724 Paris  
108 Cedex 15, France
- 109 48. Institut Pasteur, Microbial Evolutionary Genomics Unit, 75724 Paris Cedex 15,  
110 France
- 111 49. CNRS, UMR 3525, 75015, Paris, France
- 112 50. London School of Hygiene and Tropical Medicine, London, WC1E 7HT, United  
113 Kingdom
- 114

115

116 **ABSTRACT**

117

118 Together with plague, small-pox and typhus, epidemics of dysentery have been a major  
119 scourge of human populations for centuries<sup>1</sup>. A previous genomic study concluded that  
120 *Shigella dysenteriae* type 1 (Sd1), the epidemic dysentery bacillus, emerged and spread  
121 worldwide after World War (WW) I, with no clear pattern of transmission<sup>2</sup>. This is not  
122 consistent with the massive cyclic dysentery epidemics reported in Europe during the  
123 18th and 19th centuries<sup>1,3,4</sup> and the first isolation of Sd1 in Japan in 1897<sup>5</sup>. We report here  
124 a whole-genome analysis of 331 Sd1 isolates from around the world, collected between  
125 1915 and 2011, providing us with unprecedented insight into the historical spread of this  
126 pathogen. We show here that Sd1 has existed since at least the 18th century, and that it  
127 swept the globe at the end of the 19th century, diversifying into distinct lineages  
128 associated with WWI, WWII, and various conflicts or natural disasters across Africa,  
129 Asia, and Central America. We also provide a unique historical perspective on the  
130 evolution of antibiotic resistance over a 100-year period, beginning decades before the  
131 antibiotic era, and identify a prevalent multiple antibiotic-resistant lineage in South Asia  
132 that was transmitted in several waves to Africa, where it caused severe outbreaks of  
133 disease.

134

135 **TEXT**

136

137 January 2016 marks one hundred years since the invasion force from Britain,  
138 Australia, New Zealand and France withdrew from the Dardanelles, in the then Ottoman  
139 Empire, only eight months after landing. Most of the more than 120,000 casualties  
140 evacuated from the Gallipoli Peninsula were suffering from epidemic bacillary  
141 dysentery<sup>6</sup>, caused by *Shigella dysenteriae* type 1<sup>7,8</sup> (Sd1), a bacterium producing the  
142 powerful Shiga toxin. This human-adapted clone of *Escherichia coli*<sup>9</sup> was isolated for the  
143 first time by Kiyoshi Shiga during a dysentery outbreak in Japan, during which 90,000  
144 cases and 20,000 deaths occurred in the last six months of 1897 alone<sup>5</sup>. In the second half  
145 of the 20<sup>th</sup> century, large outbreaks of disease due to Sd1 were still being reported in  
146 Central America, with estimates of more than 500,000 cases and 20,000 deaths for the  
147 1969-1973 epidemic<sup>10,11</sup>, Africa, where there were an estimated 100,000 cases and 5-  
148 10,000 deaths in the 1979 epidemic<sup>12</sup>, and Asia<sup>13,14</sup>.

149 Very little is known about the origins, evolution and spread of this important  
150 human pathogen, including, in particular, the strains involved in the major outbreaks and  
151 the genetic relationships between them. We carried out a whole-genome sequence  
152 analysis on a set of Sd1 isolates selected from more than 35 international strain  
153 collections, to represent the widest possible temporal and geographic distribution of  
154 available isolates, to obtain a phylogenetic framework that was robust over time and  
155 space and to infer transmission dynamics. This unique collection included 325 isolates  
156 from 66 countries spanning four continents, collected between 1915 and 2011. Sixty-  
157 seven historical isolates collected between 1915 and 1960, including 14 isolates obtained

158 during World War I (WWI)<sup>15,16</sup>, were included in the collection, together with several  
159 isolates from each major outbreak reported since the 1960s. Short-read sequences from  
160 six Sd1 published genomes<sup>2</sup> were also included, with *S. flexneri*, *S. boydii*, *S. sonnei* and  
161 *Escherichia coli* genomes used as outgroups.

162         Single-nucleotide polymorphisms (SNPs) were detected by mapping short-read  
163 sequences against Sd1 reference genomes: Sd197<sup>17</sup>, which was isolated during an  
164 outbreak in China in the 1950s, and Sd1617<sup>18</sup>, which was isolated in Guatemala during  
165 the 1968-1969 epidemic. Maximum likelihood (ML) phylogenetic analysis was  
166 performed on 14,677 (mapping against Sd197) and 15,752 (mapping against Sd1617)  
167 chromosomal SNPs, which were randomly distributed over the non-repetitive non-  
168 recombinant core genome (85.6% of the Sd197 chromosome, Supplementary  
169 Information). Four genetic lineages (Fig. 1a, Supplementary Information) were identified.  
170 Lineage I contained only M115, which was isolated from a case in England in 1926.  
171 Lineage II contained mostly isolates collected in Europe between 1915 and 1958.  
172 Lineage III contained isolates from around the world and could be split into four  
173 sublineages with strong geographical affinities: IIIa in eastern and southeastern Asia  
174 (with isolates collected between 1927 and 1971), IIIb in Central America (1955-1992),  
175 IIIc in West Africa (1954-2006), and IIId in southern Asia and eastern Africa (1956-  
176 1977) and then in West Africa (1979-1998). Finally, lineage IV contained most of the  
177 Sd1 isolates obtained from the Indian subcontinent and Africa in the last few decades.

178         Ten of the 14 isolates (71%) amassed by Captain E.G.D. Murray during WWI  
179 belonged to the European lineage, lineage II, and most were isolated at the 2<sup>nd</sup> Western  
180 General Hospital, Manchester, which received many of the soldiers evacuated during the

181 Gallipoli campaign (Supplementary Fig. 1). The other four isolates belonged to three of  
182 the four sublineages of the global lineage, lineage III. None of the WWI isolates belonged  
183 to sublineage IIIId, which gave rise to the modern lineage, lineage IV.

184 The two candidate vaccine strains developed to date are derived from lineage III  
185 parental isolates (IIIb for parental strain Sd1617 of vaccine strain WRSd1<sup>19</sup> and IIIId for  
186 parental strain 7-87 of vaccine strain SC-599<sup>20</sup>).

187  
188 ML phylogenetic analysis revealed a strong correlation between root-to-tip branch  
189 lengths and the known years of isolation for the sequenced Sd1 isolates, indicating a  
190 clock-like evolution (Supplementary Fig. 2). We therefore used a Bayesian phylogenetic  
191 approach to provide estimates of the nucleotide substitution rates and divergence times of  
192 the different lineages for a spatially and temporally representative subset of 125 isolates  
193 (Fig. 2). We estimated the genome-wide substitution rate at  $8.7 \times 10^{-7}$  substitutions site<sup>-1</sup>  
194 year<sup>-1</sup> [95% credible interval (CI) =  $7.6 \times 10^{-7} - 9.9 \times 10^{-7}$ ], giving a most recent common  
195 ancestor (MRCA) for all the Sd1 in our collection dating from 1747 (95% CI, 1645 -  
196 1822). This finding is consistent with historical data from the 18<sup>th</sup> to mid-19<sup>th</sup> centuries,  
197 describing cyclic dysentery epidemics in Western and Northern Europe associated with  
198 extraordinarily high mortality rates. For example, the 1738-1742 and 1779 epidemics in  
199 France killed more than 200,000 people<sup>1</sup>, the 1770-1775 epidemic in Sweden killed  
200 almost 35,000 people (12% of all deaths during the period)<sup>3</sup>, and a large number of  
201 deaths from dysentery were also reported during the Irish Great Famine of 1846-1849<sup>4</sup>.  
202 The MRCA for all isolates other than M115 was dated to the mid-19<sup>th</sup> century (1853;  
203 95% CI 1831-1871), whereas the MRCAs for each of the sublineages of global lineage



204 III were estimated to have existed between 1889 (95% CI 1881-1897) and 1903 (95% CI  
205 1893-1913), indicating that this lineage spread worldwide over a period of less than two  
206 decades. This dating is also consistent with Shiga's observation that the dysentery  
207 outbreak of 1897 had begun in the late 1880s in the southern part of Japan<sup>21</sup>.

208 Our findings show that the global spread of Sd1 predates WWI. It therefore  
209 occurred earlier than for another *Shigella* serogroup, *S. sonnei*, which has been shown to  
210 have spread to other continents from Europe during the second half of the 20<sup>th</sup> century<sup>22</sup>.  
211 We cannot demonstrate causality between the spread of Sd1 and historical events on the  
212 basis of the results presented here, but the late 1800s coincided with a period of intense  
213 European emigration, the colonisation of various territories in Africa and Asia by  
214 European powers, facilitated by the opening of the Suez canal (1869) and the  
215 development of steamships.

216

217 Geographic and temporal analyses identified several intercontinental transmission  
218 events resulting in long-term establishment of the bacterium (Figs 1b, 1c, and 2).  
219 Transmission event T1 involved the European lineage II and led to an introduction of Sd1  
220 in Madagascar between 1915 (95% CI 1910-1921) and 1967 (95% CI 1956-1977), during  
221 French colonization. This is consistent with the first report, which unambiguously  
222 described Sd1 there in 1927<sup>23</sup>. Transmission event T2, involving eastern Asia and Poland,  
223 is estimated to have occurred between 1910 (95% CI 1899-1925) and 1944 (95% CI  
224 1942-1945). All other transmission waves originated in the Indian subcontinent and  
225 affected mostly East Africa. Two of these transmission waves, T5 and T8, led to major  
226 outbreaks; according to our estimates, T5 occurred between 1970 (95% CI 1963-1975)

227 and 1979 (95% CI 1976-1981). This dating is consistent with the first reported outbreak  
228 in the northeastern part of what is now the Democratic Republic of the Congo in 1979, 28  
229 years after the last isolation of Sd1 in Central Africa<sup>12</sup>. This epidemic then spread to the  
230 Great Lakes region, where it persisted until at least 1990<sup>12</sup>. T8 occurred between 1984  
231 (95% CI 1978-1987) and 1987 (95% CI 1985-1989), with a first reported outbreak in  
232 Zambia in 1990-1991<sup>12,24</sup>. The strain then rapidly spread across an Africa ravaged by  
233 civil unrest, war (e.g., Mozambique, Angola, Rwanda, Sierra Leone) and HIV  
234 infection<sup>12,24</sup> until 2011. With the exception of a localized outbreak in the northern part of  
235 the Central African Republic in 2004<sup>25</sup> caused by sublineage IIIc (see below), all other  
236 outbreaks in Africa since 1990 have been caused by lineage IV.

237

238         The high resolution of whole-genome sequence analysis (WGS) has significantly  
239 changed our understanding of the patterns of Sd1 transmission over time at a global scale.  
240 The classical molecular epidemiology tools (Supplementary Information) previously used  
241 were unable to unravel these patterns of transmission. Furthermore, a re-evaluation of  
242 two outbreaks that occurred in the Central African Republic in 2003-2004<sup>25</sup> that we had  
243 previously investigated by pulsed-field gel electrophoresis (PFGE), the current method of  
244 choice for subtyping Sd1, revealed a lack of correlation between PFGE and WGS data  
245 (Supplementary Fig. 3 and Supplementary Information). In particular, PFGE grouped the  
246 isolates from the two outbreaks closely together, whereas they actually belonged to two  
247 different lineages, IIIc and IV, separated by ~700 SNPs. By contrast, other African T8  
248 lineage IV isolates differing by 37 to 61 SNPs from the Central African Republic T8  
249 lineage IV outbreak isolates, formed a more distant group. Thus, PFGE cannot attribute

250 profiles from different apparently geographically restricted outbreaks to a single, longer  
251 epidemic, such as that associated with the T8 transmission wave in Africa. PFGE should,  
252 therefore, no longer be used for the assessment of phylogenetic relationships in Sd1.  
253 Instead, WGS provides a robust phylogenetic framework for the epidemiological tracking  
254 of this bacterium.

255

256           One key feature in the evolution of Sd1 is the acquisition and accumulation of  
257 antibiotic resistance genes (ARGs) (Figs 3, 4, Supplementary Fig. 4, and Supplementary  
258 Information). The first antibiotic-resistant Sd1 isolates in our collection were recovered in  
259 Asia and America during the 1960s and rapidly became predominant, such that  
260 susceptible isolates had become exceptional by 1991 (100%, [67/67] susceptible isolates,  
261 between 1915 and 1960 and <1% [1/123], between 1991 and 2011). Lineage IV, the most  
262 recent of the lineages identified, is the most affected by antibiotic resistance, but almost  
263 all the contemporary circulating strains from older lineages have also become resistant to  
264 multiple antibiotics. ARGs were acquired following the first use of antibiotics in clinical  
265 practice (Fig. 4b). The first ARGs identified in Sd1 were borne on small plasmids (<10  
266 kb), encoding resistance to streptomycin and sulfonamides. Larger plasmids (80-130 kb)  
267 of different types encoding additional resistance to tetracycline, chloramphenicol, and,  
268 for some plasmids, ampicillin (via the *bla*<sub>OXA-1</sub> or *bla*<sub>TEM-1</sub> genes) were then acquired in  
269 various geographic areas, from the mid-1960s to the 1980s. These plasmids belonged to  
270 the IncK and IncF groups in Asia and to the IncB/O group in Central America. The use of  
271 cotrimoxazole, beginning in the late 1960s, led to the acquisition of dihydrofolate  
272 reductase genes, mostly *dfrA1*, carried by 110-kb pST186 IncI1 and 30-kb IncX4

273 plasmids or by the Tn7 transposon inserted into the Sd1 chromosome close to the *glmS*  
274 gene, as observed for *S. sonnei*<sup>22</sup>. Since the 1990s, the principal structure associated with  
275 multidrug resistance in Sd1 has been a 66-kb genomic element called the *Shigella*  
276 resistance locus pathogenicity island (SRL-PAI)<sup>26</sup>. It was acquired four times in lineage  
277 IV (South Asia or the Middle East), once in sublineage IIIc (West Africa), and once in  
278 lineage II (Madagascar). Further evidence for the independent acquisition of the SRL-  
279 PAI is provided by the presence of slight differences between the different acquired SRL-  
280 PAIs (Supplementary Fig. 5). The SRL-A is very similar to the first SRL-PAI to be  
281 described in *S. flexneri*<sup>26</sup> and it was found exclusively in lineage IV. The SRL-B, found  
282 only in the lineage IV African T8 isolates, was probably derived from the SRL-A by  
283 insertion sequence (IS) *ISSdI*-mediated rearrangements rather than being independently  
284 acquired. The other SRL-PAI contained various insertions (group II introns, part of the  
285 *shf* operon, region replacing *orf47*) not present in SRL-A. Among the 149 isolates  
286 bearing the SRL-PAI, only two showed a partial deletion of the SRL-PAI, resulting in a  
287 loss of the antibiotic resistance cluster (i.e., the SRL *sensu stricto*). This structure is  
288 therefore quite stable over time, particularly in a bacterium containing hundreds of  
289 ISs<sup>17,18</sup>. This 66-kb element encodes resistance to ampicillin, streptomycin,  
290 chloramphenicol and tetracycline, with no more resistance than the previously circulating  
291 large plasmids. Its persistence may therefore be associated with a lower fitness cost and  
292 the presence of an *fec* operon for the capture of iron, serving as selective advantages<sup>26</sup>.  
293 Before the principal acquisition of the SRL-A, the closest ancestral group (consisting  
294 initially of South Asian and then South-East and Central Asian isolates), had acquired a  
295 chromosomally encoded transposon (Fig. 2, Supplementary Fig. 6). This 10-kb structure

296 encodes resistance to chloramphenicol and tetracycline. The structure of the double drug-  
297 resistance module is similar to that found in the SRL and to some previously circulating  
298 large multidrug resistance IncF plasmids, such as p3099-85 and p80-547. This recent  
299 trend towards acquiring ARG-containing genomic islands or chromosomally-encoded  
300 transposons rather than plasmids is also displayed by the 7th pandemic *V. cholerae*  
301 (SXT/R391) and *Salmonella enterica* serotype Typhi H58 (24-kb composite transposon)  
302 strains, which also originate from the Indian subcontinent<sup>27,28</sup>.

303 Resistance to nalidixic acid, a quinolone, mediated by point mutations in the DNA  
304 gyrase gene, *gyrA*, was acquired seven times in lineage IV Sd1 isolates from South Asia  
305 and Africa (Fig. 2) from the 1980s. The *gyrA* mutation leading to a serine-to-leucine  
306 substitution in the amino-acid sequence, S83L was the most frequently observed, but  
307 others, involving codon 87, such as D87G and D87Y, were observed in isolates from  
308 Central Africa and Thailand, respectively, during the 1990s. Interestingly, in the same  
309 geographic area of DRC and Rwanda in 1994, two different mutations were acquired  
310 (S83L and D87G). This may reflect the heavy use of nalidixic acid in the Rwandan  
311 refugee camps, which experienced outbreaks of disease caused by *Vibrio cholerae* O1  
312 and Sd1<sup>29</sup>.

313 Resistance to ciprofloxacin, a fluoroquinolone, mediated by a double mutation in  
314 *gyrA* (S83L and a second mutation in codon 87) and a mutation in the topoisomerase IV  
315 *parC* gene (S80I) was acquired only once, in a group of 20 isolates from the Indian  
316 subcontinent collected between 1995 and 2010 (Fig. 2). We observed no resistance to  
317 extended-spectrum cephalosporins, carbapenems or azithromycin in the isolates studied  
318 here, but the existence of such resistance is almost inevitable, as the area of circulation of

319 Sd1 overlaps with that of Enterobacteriaceae possessing mobile ARGs encoding  
320 resistance to the latest generation of antibiotics, such as NDM-1<sup>30</sup>. However, the dramatic  
321 decrease in Sd1 isolation reported since the turn of the century and not explained by the  
322 findings of this genomic study, may counterbalance these pessimistic predictions.

323

## 324 **METHODS**

325

### 326 **Bacterial isolates**

327

328 The Sd1 isolates analysed in this study are listed in Supplementary Table 1 and originated  
329 from the collections of the Centers for Disease Control and Prevention, Atlanta, GA,  
330 USA (*n*=56); Institut Pasteur, Paris, France (*n*=53); Public Health England, Colindale,  
331 UK (*n*=29); Iccdr,b, Dhaka, Bangladesh (*n*=29); Central Research Institute for  
332 Epidemiology, Moscow, Russian Federation (*n*=22); National Institute of Public Health,  
333 Prague, Czech Republic (*n*=19); Public Health England, Porton Down, UK (*n*=17); Iris-  
334 Lab, Brussels, Belgium (*n*=11); National Institute of Cholera and Enteric Diseases,  
335 Kolkata, India (*n*=8); Institut Pasteur de Bangui, Bangui, Central African Republic (*n*=7);  
336 Norwegian Institute of Public Health, Oslo, Norway (*n*=6); Hungarian National  
337 Collection of Medical Bacteria, Budapest, Hungary (*n*=6); Pasteur Institute of St  
338 Petersburg, St Petersburg, Russian Federation (*n*=5); National Institute of Public Health,  
339 Warsaw, Poland (*n*=5); Institut Pasteur de Dakar, Dakar, Senegal (*n*=4); New York  
340 University Langone Medical Center, New York, USA (*n*=4); Robert Koch Institut,  
341 Wernigerode, Germany (*n*=4); Institut für Hygiene und Umwelt, Hamburg, Germany

342 ( $n=3$ ); Bégin Military Hospital, Saint-Mandé, France ( $n=3$ ); IAME, Paris, France ( $n=3$ );  
343 Swedish Institute for Communicable Disease Control, Solna, Sweden ( $n=3$ ); Walter Reed  
344 Army Institute of Research, Silver Spring, MA, USA ( $n=3$ ); Epicentre, Maradi, Niger  
345 ( $n=2$ ); Polish Collection of Microorganisms, Wroclaw, Poland ( $n=2$ ); Ministry of Health,  
346 Jerusalem, Israel ( $n=2$ ); Centers for Disease Control, Taichung, Taiwan ( $n=2$ ); Centre  
347 Pasteur du Cameroun, Yaoundé, Cameroon ( $n=2$ ); National Institute of Infectious  
348 Diseases, Tokyo, Japan ( $n=1$ ); Public Health Agency of Canada, Winnipeg, Canada  
349 ( $n=1$ ); Istituto Pasteur-Fondazione Cenci Bolognetti, Rome, Italy ( $n=1$ ); Félix d'Hérelle  
350 reference center for bacterial viruses, Université Laval, Québec, Canada ( $n=1$ ); National  
351 Institute for Communicable Disease Control and Prevention, Beijing, China ( $n=1$ ).

352

353 Bacterial DNA samples were also received from the Armed Forces Research Institute of  
354 Medical Sciences, Bangkok, Thailand ( $n=10$ ).

355

356 The 18 Sd1 isolates from the E.G.D. Murray collection<sup>15,16,31</sup> included 14 isolates  
357 recovered during WWI and four isolates obtained between 1926 and 1930. The WWI  
358 isolates were obtained from different sources (Supplementary Fig. 1) and were stored at  
359 room temperature in Douglas digest agar slant glass tubes after sealing with a gas-air  
360 burner between August 1918 and October 1919. In 1980, the 18 tubes and the 680 other  
361 cultures of Enterobacteriaceae from the entire collection were shipped to the National  
362 Collection of Type Cultures (NCTC), Porton Down, UK, opened and freeze-dried.

363

364 It was confirmed that all the isolates included belonged to Sd1, by conventional methods  
365 and serotyping at the French National Reference Center for *E. coli*, *Shigella* and  
366 *Salmonella*, Institut Pasteur, Paris, as previously described<sup>32</sup>.

367

### 368 **Antibiotic susceptibility testing**

369

370 Antibiotic susceptibility was determined by disk diffusion on Mueller-Hinton (MH) agar  
371 in accordance with the guidelines of the Antibiogram Committee of the French Society  
372 for Microbiology (CA-SFM 2014) ([www.sfm-microbiologie.org/](http://www.sfm-microbiologie.org/)). The following  
373 antimicrobial drugs (Bio-Rad, Marnes-la-Coquette, France) were tested: amoxicillin,  
374 ceftriaxone, ceftazidime, streptomycin, kanamycin, amikacin, gentamicin, nalidixic acid,  
375 ofloxacin, ciprofloxacin, sulfonamides, trimethoprim, sulfamethoxazole-trimethoprim,  
376 chloramphenicol, tetracycline, and azithromycin. *Escherichia coli* CIP 76.24 (ATCC  
377 25922) was used as a control. For strains displaying resistance to either nalidixic acid or  
378 ciprofloxacin by the disk diffusion method, this resistance was confirmed by  
379 determination of the minimal inhibitory concentration (MIC) with the corresponding  
380 Etest strips (bioMérieux, Marcy L'Etoile, France). The MICs of azithromycin and  
381 nitrofurantoin were determined by Etests for 30 isolates chosen on the basis of resistance  
382 phenotype, and year and country of isolation.

383

### 384 **Determination of the mutator phenotype of strain M115**

385



386 The mutation rate of M115 was estimated by monitoring the capacity of this strain to  
387 generate mutations conferring resistance to rifampin in two independent experiments  
388 including duplicates, as previously described<sup>33</sup>. *E. coli* strain ECOR48 (CIP 106023) was  
389 used as a strong mutator positive control<sup>34</sup>, the Sd1 97-13397 isolate was used as a  
390 putative strong mutator isolate (deletion of the *mutS* gene), Sd1 M116 and Sd197 were  
391 used as putative normomutator isolates (integrity of the *mutS*, *mutH*, *mutL* and *uvrD*  
392 methyl-directed mismatch repair genes).

393

#### 394 **Total DNA extraction**

395

396 Total DNA was extracted with the InstaGene matrix kit (Bio-Rad) for the PCR  
397 identification of antibiotic resistance genes, the Wizard Genomic DNA Kit (Promega,  
398 Madison, WI, USA) for multilocus sequence typing and Illumina sequencing and the  
399 phenol chloroform method<sup>35</sup> for Illumina sequencing and PacBio sequencing.

400

#### 401 **Multi-locus sequence typing**

402

403 Conventional multi-locus sequence typing (MLST) was performed on a subset of 33 Sd1  
404 isolates, as previously described<sup>36</sup>. Sequencing was performed at the *Plateforme de*  
405 *Génotypage des Pathogènes et Santé Publique*, PF8 (Institut Pasteur). The nucleotide  
406 sequences and deduced protein sequences were analysed with EditSeq and Megalign  
407 software (DNASTAR, Madison, WI, USA). The BLASTN program of NCBI was used  
408 for database searches (<http://www.ncbi.nlm.nih.gov/BLAST/>).

409

#### 410 **PCR identification of antibiotic resistance genes**

411

412 The *bla*<sub>TEM</sub>, *bla*<sub>SHV</sub>, *bla*<sub>OXA-1</sub>, *cat1*, *sul1*, *dfrA1*, and *aadA1* resistance genes and the class  
413 1 and 2 integron gene cassettes were amplified by PCR, as previously described<sup>37</sup>.

414

415 The presence of the *Shigella* resistance locus pathogenicity island (SRL-PAI) was  
416 assessed by PCR, as previously described<sup>38</sup>. The structure of the SRL-PAI was assessed  
417 by PCR mapping with the primers described or with new primers designed on the basis of  
418 GenBank accession no. AF326777. Amplicons not of the expected size were sequenced.

419

#### 420 **Plasmid analyses**

421

422 Plasmids were obtained from *E. coli* transconjugants or transformants, as previously  
423 described<sup>37</sup>, except that ampicillin (50 mg/L) or chloramphenicol (20 mg/L) was used as  
424 a selective agent.

425

426 Plasmid size was determined in parental and transconjugant or transformants strains by  
427 S1 nuclease treatment and pulsed-field gel electrophoresis, as previously described<sup>37</sup>.

428 PCR-based replicon-typing analysis was performed as previously described<sup>39</sup>.

429

430 Eight 30-130 kb plasmids conferring antimicrobial resistance were sequenced. Plasmid  
431 DNA was extracted with the Large-Construct Kit (Qiagen, Courtaboeuf, France) and

432 sequenced through services provided by GATC Biotech (Konstanz, Germany), using  
433 shotgun sequencing runs on a 454/Roche GS FLX Analyzer (Roche, Basel, Switzerland).  
434 The resulting sequences were assembled into a unique scaffold. Gap closure was carried  
435 out by PCR followed by Sanger DNA sequencing with the Big Dye® Terminator V3.1  
436 Cycle Sequencing Kit (Applied Biosystems, Foster City, CA, USA) and a 96-capillary  
437 3730xl DNA Analyzer (Applied Biosystems), by Eurofins MGW Operon (Cochin  
438 Platform, Paris, France). Automatic annotation was performed with the RAST<sup>40</sup>  
439 server (<http://rast.nmpdr.org/>), followed by manual inspection and correction. The  
440 sequences obtained have been deposited in GenBank under the accession numbers  
441 KT754160 (p80-547), KT754161 (pCAR10), KT754162 (pBU53M1), KT754163  
442 (pA5468), KT754164 (p3099-85), KT754165 (p93-531-1), KT754166 (p92-9000),  
443 KT754167 (p69-3818).

444

#### 445 **Whole-genome sequencing**

446

447 High-throughput genome sequencing was carried out at the genomics platform of the  
448 Pasteur Institute, GATC Biotech, Beckman Coulter Genomics (Danvers, MA, USA) or at  
449 the Wellcome Trust Sanger Institute, on Illumina platforms generating 100 to 146 bp  
450 paired-end reads, yielding a mean of 206-fold coverage (minimum 37-fold, maximum  
451 990-fold) (Supplementary Table 2). Short-read sequence data were submitted to the  
452 European Nucleotide Archive (ENA) (<http://www.ebi.ac.uk/ena>) and the genome  
453 accession numbers are provided in Supplementary Table 1.

454

455 We optimised the resolution of the chromosome-encoded antibiotic resistance structures  
456 and ensured that representative isolates from the various lineages were included, by  
457 sequencing 10 isolates on the PacBIO RS II platform (Pacific Biosciences, CA, USA), as  
458 previously described<sup>28</sup>. The PacBio data were submitted to the ENA and the genome  
459 accession numbers are provided in Supplementary Table 1.

460

#### 461 **Other studied genomes**

462

463 Sd1 strain Sd197<sup>17</sup> was used as the reference genome. A second Sd1 genome Sd1617<sup>18</sup>  
464 was used as a second reference genome, to confirm the population structure found with  
465 Sd197.

466

467 Short-read sequences from the following six Sd1 genomes published by Rohmer *et al.*<sup>2</sup>  
468 were downloaded from the ENA and included in this study: 2735 (USA, 1974,  
469 SRR765065), 91R17 (Guatemala, 1991, SRR765098), 91R14 (Guatemala, 1991,  
470 SRR765104), DH03 (Central African Republic, 1996, SRR765110), DH05 (Central  
471 African Republic, 1996, SRR765112), and DH06 (Central African Republic, 1996,  
472 SRR765113).

473

474 The following genomes were used as outgroups: *E. coli* O157:H7 strain Sakai (GenBank  
475 accession no. NC\_002695), *E. coli* strain K-12 MG1655 (GenBank accession no.  
476 NC\_000913), *S. flexneri* type 2a strain 2457T (GenBank accession no. AE014073), *S.*

477 *boydii* strain Sb227 (GenBank accession no. NC\_007613), and *S. sonnei* strain Ss046  
478 (GenBank accession no. NC\_007384).

479

#### 480 **Read alignment and SNP detection**

481

482 For the analysis of single-nucleotide polymorphisms (SNPs), Illumina-generated paired-  
483 end reads and the simulated paired-end reads from publicly available assembled  
484 genomes, were mapped to the reference genome of Sd1 strain Sd197, including the  
485 chromosome (CP000034) and plasmids pSD1\_197 (CP000035) and pSD197\_spA  
486 (CP000640), with SMALT (version 0.7.4)  
487 (<http://www.sanger.ac.uk/resources/software/smalt/> as previously described<sup>28</sup>.

488

#### 489 ***De novo* assembly**

490

491 The reads for each strain were assembled *de novo* with Velvet<sup>41</sup> version 1.2.09, with  
492 parameters optimised with VelvetOptimiser version 2.2.5  
493 (<https://github.com/tseemann/VelvetOptimiser>). They were scaffolded with SSPACE<sup>42</sup>  
494 version v2.0. The gaps were closed with GapFiller<sup>43</sup> version 1.11, and the sequences were  
495 annotated with Prokka<sup>44</sup> version 1.5, as previously described<sup>28</sup>. CLC Assembly Cell  
496 version 4.2.0 (CLC bio, Aarhus, Denmark) was also used to investigate antibiotic  
497 resistance determinants.

498

#### 499 **Phylogenetic analyses**

500

501 The maximum likelihood (ML) phylogenetic tree shown in Supplementary Fig. 7 was  
502 built from a 140,385-chromosomal SNP alignment generated by `snp_sites` software  
503 ([https://github.com/sanger-pathogens/snp\\_sites](https://github.com/sanger-pathogens/snp_sites)) from all 331 short-read sequences, plus  
504 Sd1 genomes Sd197 (used as a reference) and Sd1617, together with the six *E. coli* and  
505 *Shigella* sp. genomes used as outgroups. RAxML<sup>45</sup> version 7.8.6 was used with the  
506 generalised time-reversible model and a Gamma distribution to model site-specific rate  
507 variation (the GTR+ $\Gamma$  substitution model; GTRGAMMA in RAxML). Support for the  
508 ML phylogeny was assessed by 100 bootstrap pseudo-analyses of the alignment data, and  
509 the final tree was visualised in FigTree version 1.4.2  
510 (<http://tree.bio.ed.ac.uk/software/figtree/>).

511

512 The ML phylogenetic trees shown in Figs 1a, 3a, 3c, Supplementary Figs 1a, 3b, 4, 9, 11  
513 and 14 were built from a 14,677-chromosomal SNP alignment of all 331 Sd1 short-read  
514 sequences, plus Sd1 genome Sd197, used as the reference. Repetitive regions (within the  
515 chromosome, between the chromosome and the virulence plasmid (VP) or the SRL-PAI)  
516 were removed manually with the Artemis<sup>46</sup> genome browser. Recombinogenic regions  
517 were also removed with the Gubbins<sup>47</sup> software. The remaining 14,677 chromosomal  
518 SNPs were randomly distributed along the non-repetitive non-recombinant core genome  
519 (3,750,125 bp), with a spacing of about one SNP per 256 bp or a nucleotide divergence of  
520 0.39% (Supplementary Fig. 12). RAxML version 7.8.6 (GTRGAMMA substitution  
521 model) was used to construct the tree. We performed 500 bootstrap pseudoreplicate  
522 analyses to assess support for the ML phylogeny. The tree was rooted on M115, which

523 was shown to be the most closely related to the ancestral strain of Sd1 by two different  
524 approaches (ML and Bayesian) and was visualised with MEGA<sup>48</sup> version 6, iTOL<sup>49,50</sup> or  
525 FigTree version 1.4.2.

526

527 The ML phylogenetic trees shown in Supplementary Figs 10 and 11 were built from a  
528 15,752-chromosomal SNP alignment of all 331 Sd1 short-read sequences, plus Sd1  
529 genome Sd1617, used as the reference. The method used was similar to that described  
530 above, except that the repetitive regions were not removed manually and phylogenetic  
531 support was assessed by 100 bootstrap pseudo-analyses.

532

533 The VP phylogenetic tree shown in Supplementary Fig. 15 was constructed similarly,  
534 from the 226 plasmid-containing isolates (> 90% coverage at read depth > 10x), based on  
535 290 SNPs randomly distributed along the non-repetitive non-recombinant pSD1\_197  
536 sequence (99,704 bp, 54.6% of pSD1\_197). The tree was unrooted.

537

### 538 **Phylogenetic clustering**

539

540 We clustered the isolates of Sd1 into various lineages by eye and by applying hierarchical  
541 Bayesian analysis of population structure (BAPS)<sup>51</sup> software to the 14,677-chromosomal  
542 SNP alignment. Five iterations ( $L$  value) were run with a maximum cluster number ( $K$   
543 value) of 6 or 10 and three iterations were run with  $K=6$ .

544

### 545 **Temporal analysis**

546

547 We investigated the temporal signal in the ML phylogeny for Sd1, using Path-O-Gen  
548 (<http://tree.bio.ed.ac.uk/software/pathogen/>). The relationships between root-to-tip  
549 distances, year of isolation and lineage were analysed by linear regression methods.

550

551 We used Bayesian Evolutionary Analysis by Sampling Trees (BEAST)<sup>52</sup> version 1.8 to  
552 date the important nodes. The analyses were conducted on a subsample of 125 isolates  
553 from across the ML tree, covering the full temporal and geographic range of this  
554 pathogen. The concatenated 10,798 chromosomal SNP alignments of these 125 strains  
555 were subjected to multiple BEAST analyses with both constant-size and Bayesian skyline  
556 population size change models, in combination with either a strict molecular clock or a  
557 relaxed clock, to identify the best-fit model<sup>22,53</sup>. For the BEAST analysis, the GTR+ $\Gamma$   
558 substitution model was selected and tip dates were defined as the year of isolation. For all  
559 model combinations, three independent chains of 100 million generations each were run  
560 to ensure convergence, with sampling every 1,000 iterations. Convergence and effective  
561 sample size (ESS) values were inspected using Tracer<sup>52</sup> version 1.5. A marginal  
562 likelihood estimation was carried out, with path sampling and stepping stone sampling  
563 for each run that had converged, to compare the different combinations of clock and tree  
564 models<sup>54,55</sup>. The marginal likelihood estimation was then used to determine which model  
565 gave the best fit, by calculating the Bayes Factor. The relaxed, uncorrelated lognormal  
566 clock model, which allows evolutionary rates to vary among the branches of the tree  
567 together with the skyline demographic model proved a much better fit for the data, as  
568 found previously for *S. sonnei*<sup>22</sup> and *S. flexneri*<sup>53</sup>. The parameter and tree estimates of the



569 three runs were combined with LogCombiner<sup>52</sup> version 1.7.5, with the first 20% of states  
570 in each chain removed as burn-in. Maximum clade credibility (MCC) trees were  
571 generated with TreeAnnotator<sup>52</sup> version 1.7.5 on the combined files, and visualised with  
572 FigTree version 1.4.2. Estimates are reported as median values with the 95% highest  
573 posterior density (HPD, hereafter referred to as the credible interval). The Bayesian  
574 skyline plot was calculated and visualised with Tracer<sup>52</sup> version 1.5, to investigate  
575 changes in the effective population size of Sd1 over time. To confirm the dating  
576 estimates, ten other random subsamples were generated from clusters calculated using the  
577 Prosperi method<sup>56</sup> (code here:  
578 [http://figshare.com/articles/clustertree.R\\_Code\\_for\\_clustering\\_phylogenetic\\_trees/97225](http://figshare.com/articles/clustertree.R_Code_for_clustering_phylogenetic_trees/97225))  
579 with a threshold of 0.03. All singleton isolates were included (n=86) and one isolate from  
580 each of the 33 clusters was randomly selected to generate the ten subsamples. These  
581 alignments were analysed in BEAST using the same model and showed similar dating for  
582 each of the lineages (Supplementary Table 2).

583

#### 584 **Genetic analyses**

585

586 *In silico* MLST was then carried out by MLST version 1.8  
587 (<https://cge.cbs.dtu.dk/services/MLST/>) on assembled sequences for all the dataset. New  
588 alleles were confirmed by Sanger sequencing and submitted to the MLST database  
589 website (<http://mlst.warwick.ac.uk/mlst/>).

590

591 The presence and type of antibiotic resistance genes (ARGs) or ARG-containing  
592 structures (Fig. 3b and Supplementary Fig. 4) were determined with ResFinder<sup>57</sup> version  
593 2.1 (<https://cge.cbs.dtu.dk/services/ResFinder/>), BLAST analysis against defined  
594 reference sequences (plasmids or chromosomally encoded structures), PlasmidFinder<sup>58</sup>  
595 version 1.3 (<https://cge.cbs.dtu.dk/services/PlasmidFinder/>), Plasmid MLST  
596 locus/sequence definitions database (<http://pubmlst.org/plasmid/>), and pMLST version  
597 1.2 (<https://cge.cbs.dtu.dk/services/pMLST/>) on CLC or Velvet assemblies. The new  
598 alleles and STs of IncI<sup>59</sup> and IncN<sup>60</sup> plasmids have been deposited in the PubMLST  
599 database (<http://pubmlst.org/plasmid/>). The presence of mutations in the quinolone-  
600 resistance determining region of the DNA gyrase and topoisomerase IV genes was  
601 determined manually on *de novo* assembled sequences. PacBio sequences were used to  
602 analyse the structure of the SRL-PAI variants and the composite transposon inserted into  
603 the chromosome in genome CDC 87-3330. The *in silico* results were compared with PCR  
604 data, when available.

605

## 606 **Pan-genome analysis**

607

608 Roary<sup>61</sup> version 3.2.4 was used on Velvet-annotated assemblies, to construct a pan-  
609 genome. The pan-genome analysis identified genome 2735<sup>2</sup> as an outlier. Further  
610 investigation revealed an extreme AT bias, therefore this sample was excluded from  
611 subsequent analyses. A more sensitive annotation was performed on the resulting clusters  
612 of proteins with InterPro<sup>62</sup>, to provide Gene Ontology<sup>63</sup> classifications for each gene.

613

614 **REFERENCES**

- 615 1. Kohn G.C (ed). Encyclopedia of Plague and Pestilence. New York: Facts on File  
616 New York (1995).
- 617 2. Rohmer, L. *et al.* Genomic analysis of the emergence of 20th century epidemic  
618 dysentery. *BMC Genomics*. **15**, 355 (2014).
- 619 3. Castenbrandt, H. K. A forgotten plague : dysentery in Sweden, 1750–1900.  
620 *Scand. J. Hist.* **39**, 612-639 (2014).
- 621 4. Creighton, C. A History of Epidemics in Britain. Volume II : from the Extinction  
622 of Plague to the Present Time. Cambridge University Press (1894).
- 623 5. Shiga, K. Ueber den Erreger der Dysenterie in Japan. Vorläufige Mitteilung.  
624 *Zentralbl. Bakteriol. Microbiol. Hyg.* **23**, 599-600 (1898).
- 625 6. Manson-Bahr, P. H. Dysentery and diarrhoea in wartime. *Br. Med. J.* **2**, 346-348  
626 (1942).
- 627 7. Ledingham, J. C. & Penfold, W. J. Recent bacteriological experiences with  
628 typhoidal disease and dysentery. *Br. Med. J.* **2**, 704-711 (1915).
- 629 8. Tribondeau, L. & Fichet, M. Note sur les dysenteries des Dardanelles. *Ann. Inst.*  
630 *Pasteur (Paris)* **30**, 357-362 (1916).
- 631 9. Pupo, G. M., Lan, R. & Reeves, P. R. Multiple independent origins of *Shigella*  
632 clones of *Escherichia coli* and convergent evolution of many of their  
633 characteristics. *Proc. Natl Acad. Sci. U S A* **97**, 10567-10572 (2000).
- 634 10. Mata, L. J., Gangarosa, E. J, Cáceres, A., Perera D.R. & Mejicanos M. L.  
635 Epidemic Shiga bacillus dysentery in Central America. I. Etiologic investigations  
636 in Guatemala, 1969. *J. Infect. Dis.* **122**, 170-180 (1970).

- 637 11. Parsonnet, J. *et al.* *Shigella dysenteriae* type 1 infections in US travellers to  
638 Mexico, 1988. *Lancet* **2**, 543-545 (1989).
- 639 12. Cobra, C. & Sack, D. A. The Control of Epidemic Dysentery in Africa: Overview,  
640 Recommendations, and Checklists. SD Publication Series. Technical Paper No.  
641 37. Washington, DC : USAID. Bureau for Africa. Office of sustainable  
642 development (1996) accessed October 1 2015 at  
643 [http://rportal.net/library/content/usaaid-afr-sd-publications-series/usaaid-afr-sd-  
health/the-control-of-epidemic-dysentery-in-africa-overview-recommendations-  
and-checklists/at\\_download/file](http://rportal.net/library/content/usaaid-afr-sd-publications-series/usaaid-afr-sd-<br/>644 health/the-control-of-epidemic-dysentery-in-africa-overview-recommendations-<br/>645 and-checklists/at_download/file)
- 646 13. Rahaman, M. M., Khan, M. M., Aziz, K. M., Islam, M. S. & Kibriya A. K. An  
647 outbreak of dysentery caused by *Shigella dysenteriae* type 1 on a coral island in  
648 the Bay of Bengal. *J. Infect. Dis.* **132**, 15-19 (1975).
- 649 14. Taylor, D. N. *et al.* Introduction and spread of multi-resistant *Shigella dysenteriae*  
650 1 in Thailand. *Am. J. Trop. Med. Hyg.* **40**, 77-85 (1989).
- 651 15. Murray, G. R. E. More on bacterial longevity: the Murray collection. *ASM News*  
652 **51**, 261–262 (1985).
- 653 16. Baker, K. S. *et al.* The Murray collection of pre-antibiotic era Enterobacteriaceae:  
654 a unique research resource. *Genome Med.* **7**, 97 (2015).
- 655 17. Yang, F. *et al.* Genome dynamics and diversity of *Shigella* species, the etiologic  
656 agents of bacillary dysentery. *Nucleic Acids Res.* **33**, 6445-58 (2005).
- 657 18. Vongsawan, A. A. *et al.* The genome of *Shigella dysenteriae* strain Sd1617  
658 comparison to representative strains in evaluating pathogenesis. *FEMS Microbiol.*  
659 *Lett.* **362**, pii: fnv011 (2015).

- 660 19. McKenzie, R. *et al.* Safety and immunogenicity of WRSd1, a live attenuated  
661 *Shigella dysenteriae* type 1 vaccine candidate. *Vaccine* **26**, 3291-3296 (2008).
- 662 20. Launay, O. *et al.* Safety and immunogenicity of SC599, an oral live attenuated  
663 *Shigella dysenteriae* type-1 vaccine in healthy volunteers: results of a Phase 2,  
664 randomized, double-blind placebo-controlled trial. *Vaccine* **27**, 1184-1191 (2009).
- 665 21. Shiga, K. Observations on the epidemiology of dysentery in Japan. *Philipp. J. Sci.*  
666 **1**, 485-500 (1906).
- 667 22. Holt, K. E. *et al.* *Shigella sonnei* genome sequencing and phylogenetic analysis  
668 indicate recent global dissemination from Europe. *Nat. Genet.* **44**, 1056-10569  
669 (2012).
- 670 23. Robic, J. Une épidémie de dysenterie bacillaire à Madagascar (1927-1928). *Bull.*  
671 *Soc. Path. Exot.* **21**, 709-713 (1928).
- 672 24. Guerin, P. J., Grais, R. F., Rottingen, J. A., Valleron, A. J. & Shigella Study  
673 Group. Using European travellers as an early alert to detect emerging pathogens  
674 in countries with limited laboratory resources. *BMC Public Health.* **7**, 8 (2007).
- 675 25. Bercion, R. *et al.* Molecular epidemiology of multidrug-resistant *Shigella*  
676 *dysenteriae* type 1 causing dysentery outbreaks in Central African Republic,  
677 2003-2004. *Trans R. Soc. Trop. Med. Hyg.* **100**, 1151-1158 (2006).
- 678 26. Luck, S. N., Turner, S. A., Rajakumar, K., Sakellaris, H. & Adler, B. Ferric  
679 dicitrate transport system (Fec) of *Shigella flexneri* 2a YSH6000 is encoded on a  
680 novel pathogenicity island carrying multiple antibiotic resistance genes. *Infect.*  
681 *Immun.* **69**, 6012-6021 (2001).

- 682 27. Mutreja, A. *et al.* Evidence for several waves of global transmission in the  
683 seventh cholera pandemic. *Nature*. **477**, 462-465 (2011).
- 684 28. Wong, V. K. *et al.* Phylogeographical analysis of the dominant multidrug-  
685 resistant H58 clade of *Salmonella* Typhi identifies inter- and intracontinental  
686 transmission events. *Nat. Genet.* **47**, 632-639 (2015).
- 687 29. Islam, M.S., *et al.* Microbiological investigation of diarrhoea epidemics among  
688 Rwandan refugees in Zaire. *Trans. R. Soc. Trop. Med. Hyg.* **89**, 506 (1995).
- 689 30. Kumarasamy, K. K. *et al.* Emergence of a new antibiotic resistance mechanism in  
690 India, Pakistan, and the UK: a molecular, biological, and epidemiological study.  
691 *Lancet Infect. Dis.* **10**, 597-602 (2010).
- 692 31. Murray, E. G. D. An attempt at classification of *Bacillus dysenteriae*, based upon  
693 an examination of the agglutinating properties of fifty-three strains. *J. R. Army*  
694 *Med. Corps* **31**, 257-271 (1918).
- 695 32. Langendorf, C. *et al.* Enteric bacterial pathogens in children with diarrhea in  
696 Niger: diversity and antimicrobial resistance. *PLoS One* **10**, e0120275 (2015).
- 697 33. Taddei, F., Matic, I. & Radman, M. cAMP-dependent SOS induction and  
698 mutagenesis in resting bacterial populations. *Proc. Natl Acad. Sci. USA* **92**,  
699 11736-11740 (1995).
- 700 34. Picard, B. *et al.* Mutator natural *Escherichia coli* isolates have an unusual  
701 virulence phenotype. *Infect. Immun.* **69**, 9-14 (2001).
- 702 35. Grimont, F. & Grimont, P. A. D. Determination of rDNA gene restriction  
703 patterns. *Methods Mol. Biol.* **46**, 181-200 (1995).

- 704 36. Wirth, T. *et al.* Sex and virulence in *Escherichia coli*: an evolutionary perspective.  
705 Mol. Microbiol. **60**, 1136-1151 (2006).
- 706 37. Fabre, L *et al.* Chromosomal integration of the extended-spectrum beta-lactamase  
707 gene *bla*<sub>CTX-M-15</sub> in *Salmonella enterica* serotype Concord isolates from  
708 internationally adopted children. *Antimicrob. Agents Chemother.* **53**, 1808-1816  
709 (2009).
- 710 38. Turner, S. A., Luck, S. N., Sakellaris, H., Rajakumar, K. & Adler, B. Molecular  
711 epidemiology of the SRL pathogenicity island. *Antimicrob. Agents Chemother.*  
712 **47**, 727-734 (2003).
- 713 39. Carattoli, A. *et al.* Identification of plasmids by PCR-based replicon typing. *J.*  
714 *Microbiol. Methods* **63**, 219–228 (2005).
- 715 40. Aziz, R.K. *et al.* The RAST server: rapid annotation using subsystems  
716 technology. *BMC Genomics* **9**, 75 (2008)
- 717 41. Zerbino, D.R. & Birney, E. Velvet: algorithms for *de novo* short read assembly  
718 using de Bruijn graphs. *Genome Res.* **18**, 821-9 (2008).
- 719 42. Boetzer, M., Henkel, C. V., Jansen, H. J., Butler, D. & Pirovano, W. Scaffolding  
720 pre-assembled contigs using SSPACE. *Bioinformatics* **27**, 578-579 (2011).
- 721 43. Boetzer, M. & Pirovano, W. Toward almost closed genomes with GapFiller.  
722 *Genome Biol.* **13**, R56 (2012).
- 723 44. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**,  
724 2068-2069 (2014).

- 725 45. Stamatakis, A. RAxML-VI-HPC: maximum likelihood-based phylogenetic  
726 analyses with thousands of taxa and mixed models. *Bioinformatics* **22**, 2688-2690  
727 (2006).
- 728 46. Carver, T. *et al.* Artemis and ACT: viewing, annotating and comparing sequences  
729 stored in a relational database. *Bioinformatics*. **24**, 2672-6 (2008).
- 730 47. Croucher, N. J. *et al.* Rapid phylogenetic analysis of large samples of recombinant  
731 bacterial whole genome sequences using Gubbins. *Nucleic Acids Res.* **43**, e15  
732 (2015).
- 733 48. Tamura, K., Stecher, G., Peterson, D., Filipski A. & Kumar, S. MEGA6:  
734 Molecular Evolutionary Genetics Analysis version 6.0. *Mol. Biol. Evol.* **30**, 2725-  
735 2729 (2013).
- 736 49. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL): an online tool for  
737 phylogenetic tree display and annotation. *Bioinformatics* **23**, 127-8 (2007).
- 738 50. Letunic, I. & Bork, P. Interactive Tree Of Life v2: online annotation and display  
739 of phylogenetic trees made easy. *Nucleic Acids Res.* **39**, W475-8 (2011).
- 740 51. Cheng, L., Connor, T. R., Siren, J., Aanensen, D. M. & Corander, J. Hierarchical  
741 and spatially explicit clustering of DNA sequences with BAPS software. *Mol.*  
742 *Biol. Evol.* **30**, 1224-1228 (2013).
- 743 52. Drummond, A. J. & Rambaut, A. BEAST: Bayesian evolutionary analysis by  
744 sampling trees. *BMC Evol. Biol.* **7**, 214 (2007).
- 745 53. Connor, T.R. *et al.* Species-wide whole genome sequencing reveals historical  
746 global spread and recent local persistence in *Shigella flexneri*. *eLife* **4**, e07335  
747 (2015).



- 748 54. Baele, G. *et al.* Improving the accuracy of demographic and molecular clock  
749 model comparison while accommodating phylogenetic uncertainty. *Mol. Biol.*  
750 *Evol.* **29**, 2157-2167 (2012).
- 751 55. Baele, G., Li, W. L., Drummond, A. J., Suchard, M. A. & Lemey, P. Accurate  
752 model selection of relaxed molecular clocks in bayesian phylogenetics. *Mol. Biol.*  
753 *Evol.* **30**, 239-243 (2013).
- 754 56. Prosperi, M. C. *et al.* A novel methodology for large-scale phylogeny partition.  
755 *Nat. Commun.* **2**, 321 (2011).
- 756 57. Zankari, E. *et al.* Identification of acquired antimicrobial resistance genes. *J.*  
757 *Antimicrob. Chemother.* **67**, 2640-2644 (2012).
- 758 58. Carattoli, A., *et al.* *In silico* detection and typing of plasmids using PlasmidFinder  
759 and plasmid multilocus sequence typing. *Antimicrob. Agents Chemother.* **58**,  
760 3895-3903 (2014).
- 761 59. García-Fernández, A. *et al.* Multilocus sequence typing of IncI1 plasmids  
762 carrying extended-spectrum beta-lactamases in *Escherichia coli* and *Salmonella*  
763 of human and animal origin. *J. Antimicrob. Chemother.* **61**, 1229-1233 (2008).
- 764 60. García-Fernández, A. *et al.* Multilocus sequence typing of IncN plasmids. *J.*  
765 *Antimicrob. Chemother.* **66**, 1987-1991 (2011).
- 766 61. Page, A. J. *et al.* Roary: Rapid large-scale prokaryote pan genome analysis.  
767 *Bioinformatics* **31**, 3691-3693 (2015).
- 768 62. Mitchell, A. *et al.* The InterPro protein families database: the classification  
769 resource after 15 years. *Nucleic Acids Res.* **43**, D213-D221 (2015).

770 63. Gene Ontology Consortium. Gene Ontology Consortium: going forward. *Nucleic*  
771 *Acids Res.* **43**, D1049-D1056 (2015).

772

### 773 **SUPPLEMENTARY INFORMATION**

774 Supplementary Information is linked to the online version of the paper at

775 [www.nature.com/nature](http://www.nature.com/nature).

776

777 The authors declare no competing financial interests.

778

779 Correspondence and requests for materials should be addressed to F.-X.W.

780 (fxweill@pasteur.fr)

781

### 782 **ACKNOWLEDGEMENTS**

783

784 This study was supported by the Institut Pasteur and the Institut Pasteur International

785 Network, the Institut de Veille Sanitaire, the French government's Investissement

786 d'Avenir programme, Laboratoire d'Excellence 'Integrative Biology of Emerging

787 Infectious Diseases' (grant number ANR-10-LABX-62-IBEID), the Fondation « Le

788 Roch-Les Mousquetaires », the Canetti family through the Georges, Jacques et Elias

789 Canetti Award 2013, the Wellcome Trust through grant 098051 to the Sanger Institute,

790 the NHMRC of Australia (grant 1061409 to K.E.H), the Victorian Life Sciences

791 Computation Initiative (VLSCI) (grant VR0082) and the Indian Council of Medical

792 Research, New Delhi, India. We thank A. Dautry-Versat, A. P. Pugsley, C. Bréchet and J.

793 Savall for their support; T. Hieu, C. Soto, E. Bourreterre and B. Faye for technical  
794 assistance; Z. Szabó, D. Tremblay for providing isolates; L. R. Hiltzik, N. Baldwin and  
795 C. Mackenzie for their searches of the archives; M. Toucas, H. d’Hauteville, E. Aldová,  
796 S. Formal, and A.T. Maurelli for information about isolates; D. Nedelec for helpful  
797 discussion, I. Gut, M. Gut, L. Ma, D. Harris, K. Oliver, and the sequencing teams at the  
798 Institut Pasteur and Wellcome Trust Sanger Institute for sequencing the samples. The  
799 views expressed in this publication are those of the authors and do not reflect the views of  
800 the US Department of the Army or Department of Defense.  
801 The funders had no role in study design, data collection and analysis, decision to publish,  
802 or preparation of the manuscript.

803

#### 804 **AUTHOR CONTRIBUTIONS**

805

806 R.B., P.A.D.G., S.B., N.R.T and F.-X.W. designed the study. N.S., C.J., K.A.T., R.B.,  
807 K.K., R.K., J.E.R., L.K., A.K., O.V., C.J.M., T.R., C. Bizet, A.G.S, A.G., A.L.W., M.-  
808 C.F., S.L.H., M.J.B, C.J., A.M., A.-L.P., P.R., A.F., E.D., M.V., H.B., M.H., P.A.D.G.,  
809 P.S., L.B., C.-S.C., D.C., B.C., S.E., G.P.P., A.V.E., H.I., A.K.-K., A.L., M.G., F.G.,  
810 C.L., M.M., L.A.M.P, G.P.-P., A.P., G.A.S., D.T., C.Y., H.Z., P.S. and F.-X.W. selected  
811 and provided characterized isolates and their epidemiological information. E.N.-N., M.L.-  
812 C., I. C., C.R., A.T.-D., M. A.-D. and L.B. did the phenotypic experiments and DNA  
813 extractions. A.E.M. and S.R.H provided guidance for genomic analyses. C. Bouchier  
814 performed the whole-genome sequencing. M.A. processed the short reads. E.N.-N., N.F.,  
815 K.K., S. B., K.E.H, J.H, A.J.P., G.G., E.S., and F.-X.W. analysed the genomic sequence

816 data. F.-X.W. wrote the manuscript with major contributions from A.E.M., P.A.D.G.,  
817 E.D., J.P., P.S., K.E.H., S.B. and N.R.T. All authors contributed to manuscript editing.  
818 F.-X.W. oversaw the project.

819

## 820 **AUTHOR INFORMATION**

821

822 Short-read sequences have been deposited at EBI-ENA, under study accession numbers  
823 PRJEB10304, PRJEB2846 and PRJEB3255. PacBio sequences have been deposited at  
824 EBI-ENA, under study accession number PRJEB7928. Plasmid, SRL-PAI, and Tn87-  
825 3330 sequences have been deposited in GenBank, under accession numbers KT754160–  
826 KT754167, KT777637–KT777641, and KT777642, respectively.

827

828 Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints)

829

## 830 **TABLE**

831 None

832

## 833 **FIGURE LEGENDS**

834

835 **Figure 1. Geographic distribution and transmission patterns of *Shigella dysenteriae***  
836 **type 1 genetic lineages. a**, Maximum likelihood (ML) phylogeny of the 332 genomes  
837 studied, showing the four lineages, I to IV, and the four sublineages of lineage III: IIIa to  
838 IIIId. The tree was rooted on M115, the most closely related to the *S. dysenteriae* type 1

839 ancestral strain. The tips of the tree are coloured to indicate the continent on which the  
840 infection occurred. T1 to T8 indicate intercontinental transmission events. **b**, Geographic  
841 presence (circles), inferred arrivals (thick arrows) and principal long-distance  
842 transmission events (thin arrows) of lineages I to III based on phylogeographic analysis.  
843 Intercontinental transmission events are indicated by the letter T. The date ranges shown  
844 for transmission events are the median values for the MRCA (taken from BEAST) with  
845 the first number indicating the median MRCA of the transmitted strains, and the second  
846 number indicating the median MRCA of the transmitted strains and their closest relative  
847 from the source location. **c**, Geographic presence (circles, thunderbolts) and  
848 intercontinental transmission events of lineage IV based on phylogeographic analysis.  
849 Isolate assignment to the corresponding transmission event is indicated by coloured  
850 halos.

851

852 **Figure 2. Timed phylogeny of a subsample of 125 *Shigella dysenteriae* type 1 isolates.**

853 **a**, Bayesian skyline plot showing temporal changes since 1747 in effective population  
854 size (black curve) with 95% confidence intervals (cyan). World War I (WWI) is indicated  
855 by a red bar. **b**, Maximum clade credibility tree produced using BEAST (lognormal  
856 relaxed clock model; Bayesian skyline) also presenting information about the ortho-  
857 nitrophenyl- $\beta$ -galactoside (ONPG) test. Resistance to nalidixic acid (NAL<sup>R</sup>) is indicated  
858 by a purple circle and resistance to ciprofloxacin (CIP<sup>R</sup>) is indicated by a purple triangle.  
859 Acquisition of the antibiotic resistance element, *Shigella* resistance locus pathogenicity  
860 island (SRL-PAI), is indicated by a black thunderbolt. Acquisition of the resistance  
861 transposon (Tn87-3330), originally found in isolate CDC 87-3330, is indicated by an

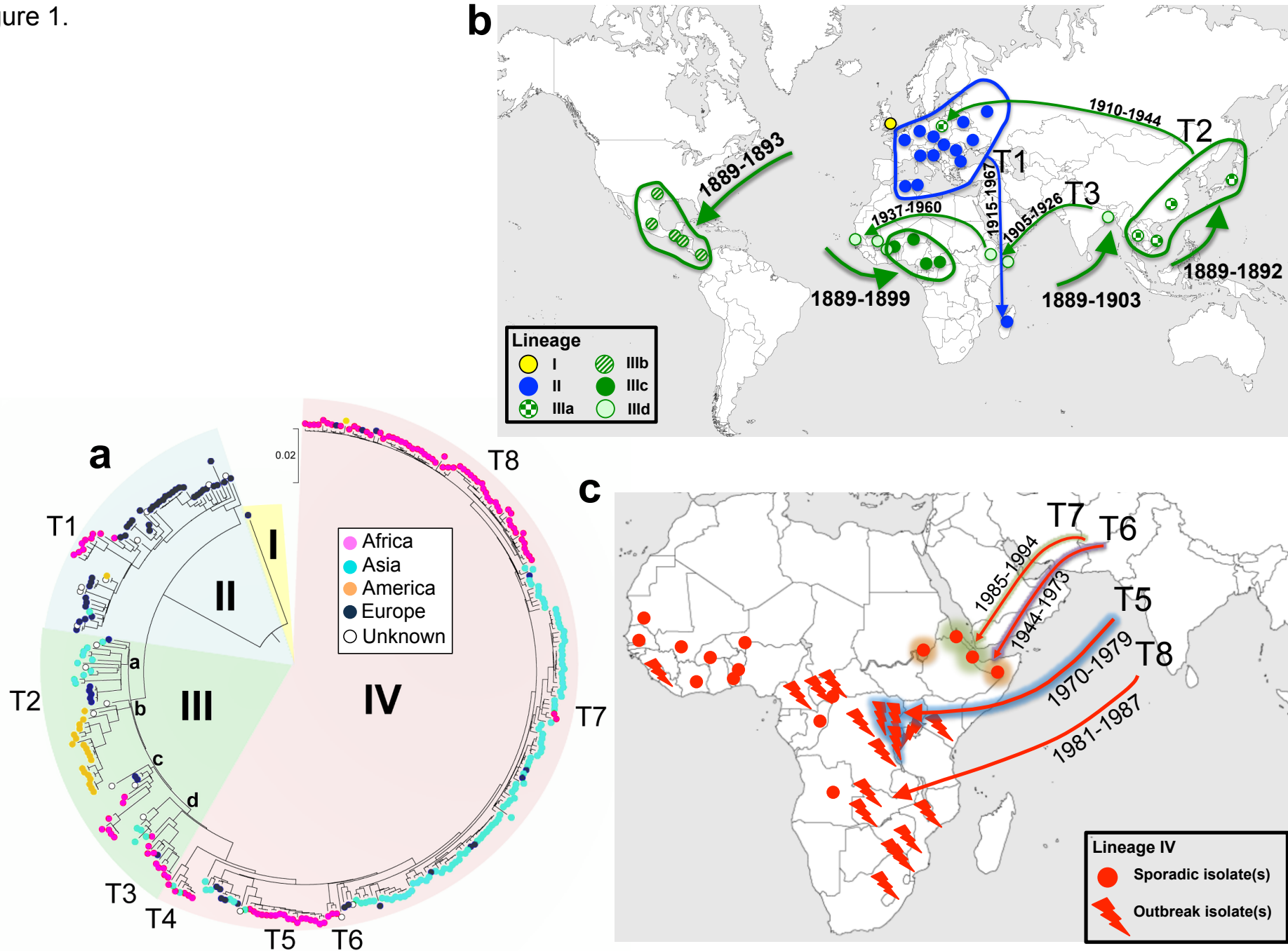
862 orange thunderbolt. T1 to T8 indicate intercontinental transmission events. Estimated  
863 dates for the intercontinental transmission events are provided in dataset S7 of  
864 Supplementary Table 2.

865

866 **Figure 3. Phenotypic and genetic characterization of antibiotic resistance in *Shigella***  
867 ***dysenteriae* type 1. a**, Resistance phenotype for eight antibiotics (ampicillin, AMP;  
868 streptomycin, STR; sulfonamides, SUL; trimethoprim, TMP; chloramphenicol, CHL;  
869 tetracycline, TET; nalidixic acid, NAL; and ciprofloxacin, CIP), according to the lineages  
870 (I to IV) defined on the basis of the maximum likelihood (ML) phylogeny (as in Fig. 1a).  
871 Resistance is indicated in red and susceptibility in grey, whereas no antibiotic  
872 susceptibility data is indicated in white. **b**, Principal genetic structures bearing antibiotic  
873 resistance genes (ARGs) as a function of genetic lineage (defined by ML phylogeny),  
874 time period and geography. A more detailed figure is provided in Supplementary Fig. 4.  
875

876 **Figure 4. Evolution of antibiotic resistance of *Shigella dysenteriae* type 1. a**, Change  
877 in the number of antibiotic resistance genes (ARGs) per isolate over time. The  
878 logarithmic trendline and the correlation coefficient of determination ( $R^2$ ) are shown in  
879 red. **b**, Timeline of the first detection of the main ARGs in our collection. The antibiotics  
880 (AMP, ampicillin; STR, streptomycin; SUL, sulfonamides; TMP, trimethoprim; CHL,  
881 chloramphenicol; TET, tetracycline; NAL, nalidixic acid; and CIP, ciprofloxacin) for  
882 which the ARGs convey resistance to are indicated. Asteriks indicate the mutation of  
883 chromosomal genes of the core genome.

Figure 1.







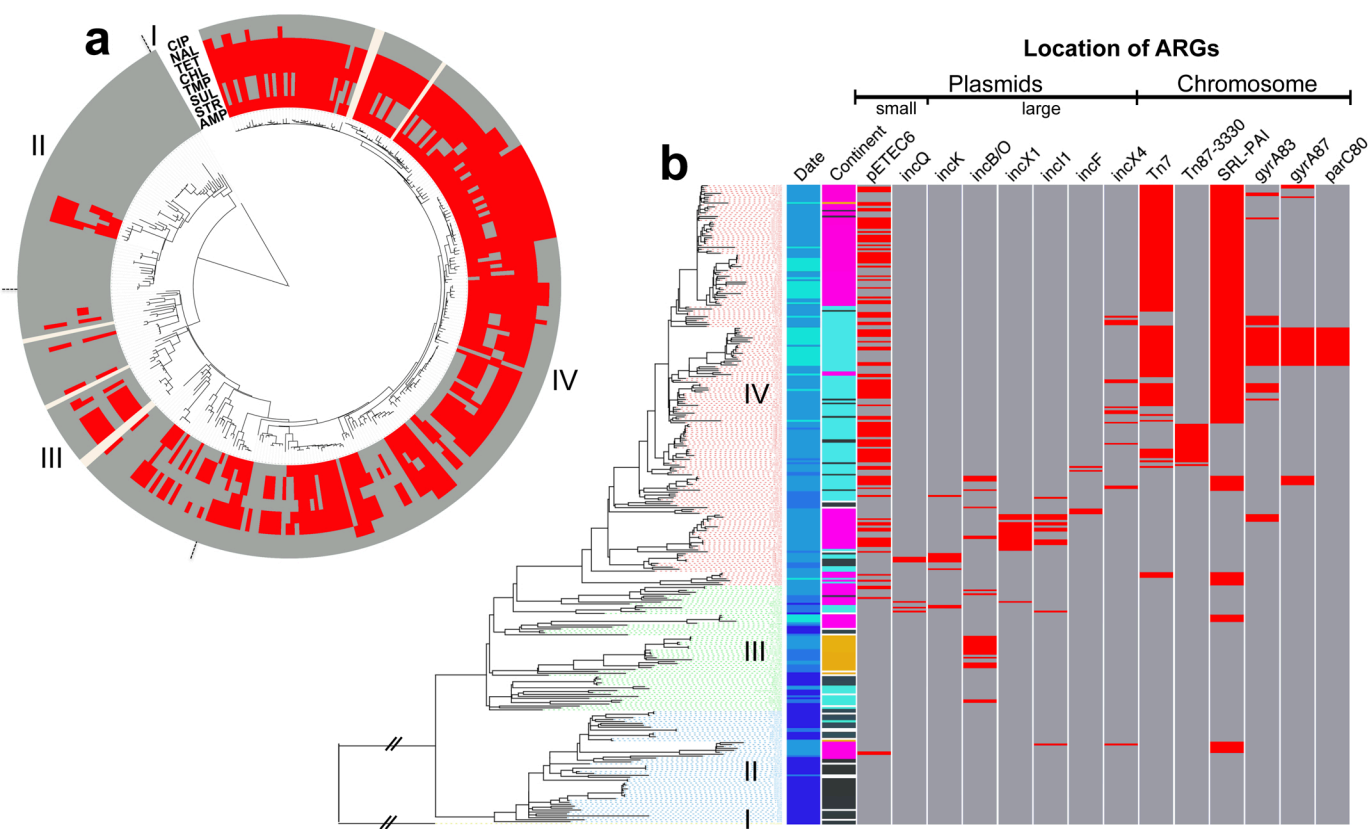


Figure 3.

**Keys**

(Date) 1915-1959 1960-1979 1980-1999 2000-2011

(Continent) Africa America Asia Europe

(Others) Presence Absence

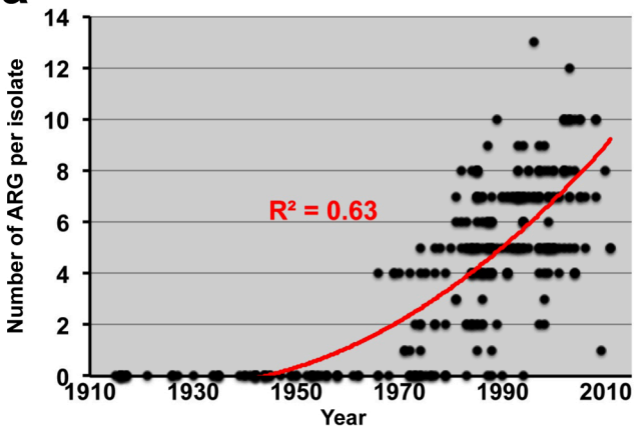
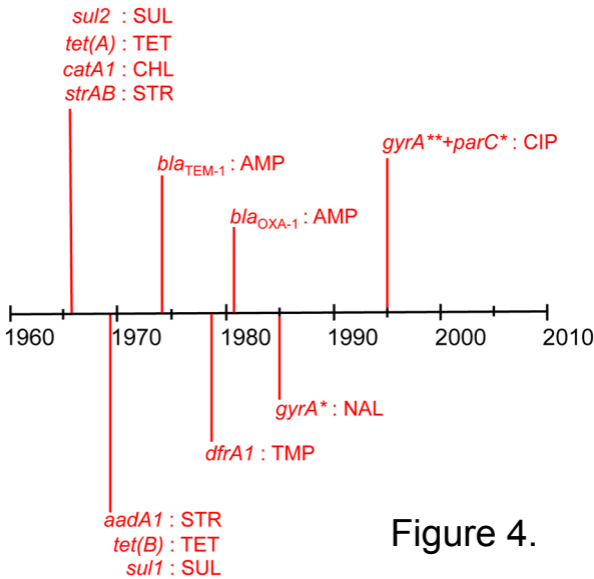
**a****b**

Figure 4.