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

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

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

Elisabeth Njamkepo[1]*, Nizar Fawal[1]*, Alicia Tran-Dien[1]*, Jane Hawkey[2,3,4], Nancy Strockbine[5], Claire Jenkins[6], Kaisar A. Talukder[7], Raymond Bercion[8,9], Konstantin Kuleshov[10], Renáta Kolínská[11], Julie E. Russell[12], Lidia Kaftyreva[13], Marie Accou-Demartin[1], Andreas Karas[14], Olivier Vandenberg[15,16], Alison E. Mather[17,18], Carl J. Mason[19], Andrew J. Page[17], Thandavarayan Ramamurthy[20], Chantal Bizet[21], Andrzej Gamian[22], Isabelle Carle[1], Amy Gassama Sow[9], Christiane Bouchier[23], Astrid Louise Wester[24], Monique Lejay-Collin[1], Marie-Christine Fonkoua[25], Simon Le Hello[1], Martin J. Blaser[26], Cecilia Jernberg[27], Corinne Ruckly[1], Audrey Mérens[28], Anne-Laure Page[29], Martin Aslett[17], Peter Roggentin[30], Angelika Fruth[31], Erick Denamur[32], Malabi Venkatesan[33], Hervé Bercovier[34], Ladaporn Bodhidatta[19], Chien-Shun Chiou[35], Dominique Clermont[21], Bianca Colonna[36], Svetlana Egorova[13], Gururaja P. Pazhani[20], Analia V. Ezernitchi[37], Ghislaine Guigon[38], Simon R. Harris[17], Hidemasa Izumiya[39], Agnieszka Korzeniowska-Kowal[22], Anna Lutyńska[40], Malika Gouali[1], Francine Grimont[1], Céline Langendorf[29], Monika Marejková[41], Lorea A. M. Peterson[42], Guillermo Perez-Perez[26], Antoinette Ngandjio[25], Alexander Podkolzin[10], Erika Souche[43], Mariia Makarova[13], German A. Shipulin[10], Changyun Ye[44], Helena Žemličková[11,45], Mária Herpay[46], Patrick A.D. Grimont[1], Julian Parkhill[17], Philippe Sansonetti[47], Kathryn E. Holt[2,3], Sylvain Brisse[38,48,49], Nicholas R. Thomson[17,50], François-Xavier Weill[1,17]‡

* Contributed equally

‡ Corresponding author

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

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

**ABSTRACT**

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

**TEXT**

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

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

158 during World War I (WWI)[15,16], 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[2] 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[17], which was isolated during an
164 outbreak in China in the 1950s, and Sd1617[18], 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$^{nd}$ Western
180 General Hospital, Manchester, which received many of the soldiers evacuated during the

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

The two candidate vaccine strains developed to date are derived from lineage III parental isolates (IIIb for parental strain Sd1617 of vaccine strain WRSd1[19] and IIId for parental strain 7-87 of vaccine strain SC-599[20]).

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

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

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

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

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

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

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

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

plasmids or by the Tn*7* transposon inserted into the Sd1 chromosome close to the *glmS* gene, as observed for *S. sonnei*[22]. Since the 1990s, the principal structure associated with multidrug resistance in Sd1 has been a 66-kb genomic element called the *Shigella* resistance locus pathogenicity island (SRL-PAI)[26]. It was acquired four times in lineage IV (South Asia or the Middle East), once in sublineage IIIc (West Africa), and once in lineage II (Madagascar). Further evidence for the independent acquisition of the SRL-PAI is provided by the presence of slight differences between the different acquired SRL-PAIs (Supplementary Fig. 5). The SRL-A is very similar to the first SRL-PAI to be described in *S. flexneri*[26] and it was found exclusively in lineage IV. The SRL-B, found only in the lineage IV African T8 isolates, was probably derived from the SRL-A by insertion sequence (IS) IS*Sd1*-mediated rearrangements rather than being independently acquired. The other SRL-PAI contained various insertions (group II introns, part of the *shf* operon, region replacing orf47) not present in SRL-A. Among the 149 isolates bearing the SRL-PAI, only two showed a partial deletion of the SRL-PAI, resulting in a loss of the antibiotic resistance cluster (i.e., the SRL *sensu stricto*). This structure is therefore quite stable over time, particularly in a bacterium containing hundreds of ISs[17,18]. This 66-kb element encodes resistance to ampicillin, streptomycin, chloramphenicol and tetracycline, with no more resistance than the previously circulating large plasmids. Its persistence may therefore be associated with a lower fitness cost and the presence of an *fec* operon for the capture of iron, serving as selective advantages[26]. Before the principal acquisition of the SRL-A, the closest ancestral group (consisting initially of South Asian and then South-East and Central Asian isolates), had acquired a 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[27,28].

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[29].

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

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

## METHODS

### Bacterial isolates

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

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

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

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

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

**Antibiotic susceptibility testing**

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

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

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[33]. *E. coli* strain ECOR48 (CIP 106023) was
389 used as a strong mutator positive control[34], 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[35] 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[36]. 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*$_{TEM}$, *bla*$_{SHV}$, *bla*$_{OXA-1,}$ *cat1*, *sul1*, *dfrA1*, and *aadA1* resistance genes and the class

413 1 and 2 integron gene cassettes were amplified by PCR, as previously described[37].

414

415 The presence of the *Shigella* resistance locus pathogenicity island (SRL-PAI) was

416 assessed by PCR, as previously described[38]. 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[37], 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[37].

428 PCR-based replicon-typing analysis was performed as previously described[39].

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

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

**Whole-genome sequencing**

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

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[28]. 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[17] was used as the reference genome. A second Sd1 genome Sd1617[18]
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*.[2]
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.*

*boydii* strain Sb227 (GenBank accession no. NC_007613), and *S. sonnei* strain Ss046 (GenBank accession no. NC_007384).

**Read alignment and SNP detection**

For the analysis of single-nucleotide polymorphisms (SNPs), Illumina-generated paired-end reads and the simulated paired-end reads from publicly available assembled genomes, were mapped to the reference genome of Sd1 strain Sd197, including the chromosome (CP000034) and plasmids pSD1_197 (CP000035) and pSD197_spA (CP000640), with SMALT (version 0.7.4) (http://www.sanger.ac.uk/resources/software/smalt/ as previously described[28].

***De novo* assembly**

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

**Phylogenetic analyses**

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

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

was shown to be the most closely related to the ancestral strain of Sd1 by two different approaches (ML and Bayesian) and was visualised with MEGA[48] version 6, iTOL[49,50] or FigTree version 1.4.2.

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

The VP phylogenetic tree shown in Supplementary Fig. 15 was constructed similarly, from the 226 plasmid-containing isolates (> 90% coverage at read depth > 10x), based on 290 SNPs randomly distributed along the non-repetitive non-recombinant pSD1_197 sequence (99,704 bp, 54.6% of pSD1_197). The tree was unrooted.

**Phylogenetic clustering**

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

**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)[52] 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[22,53]. For the BEAST analysis, the GTR+⌈
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[52] 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[54,55]. 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*[22] and *S. flexneri*[53]. The parameter and tree estimates of the

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

**Genetic analyses**

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

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

**Pan-genome analysis**

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

## REFERENCES


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

11. Parsonnet, J. *et al*. *Shigella dysenteriae* type 1 infections in US travellers to Mexico, 1988. *Lancet* **2,** 543-545 (1989).

12. Cobra, C. & Sack, D. A. The Control of Epidemic Dysentery in Africa: Overview, Recommendations, and Checklists. SD Publication Series. Technical Paper No. 37. Washington, DC : USAID. Bureau for Africa. Office of sustainable development (1996) accessed October 1 2015 at http://rmportal.net/library/content/usaid-afr-sd-publications-series/usaid-afr-sd-health/the-control-of-epidemic-dysentery-in-africa-overview-recommendations-and-checklists/at_download/file

13. Rahaman, M. M., Khan, M. M., Aziz, K. M., Islam, M. S. & Kibriya A. K. An outbreak of dysentery caused by *Shigella dysenteriae* type 1 on a coral island in the Bay of Bengal. *J. Infect. Dis.* **132,** 15-19 (1975).

14. Taylor, D. N. *et al.* Introduction and spread of multi-resistant *Shigella dysenteriae* 1 in Thailand. *Am. J. Trop. Med. Hyg.* **40,** 77-85 (1989).

*15.* Murray, G. R. E. More on bacterial longevity: the Murray collection. *ASM News* **51,** 261–262 (1985).

*16.* Baker, K. S. *et al.* The Murray collection of pre-antibiotic era Enterobacteriaceae: a unique research resource. *Genome Med.* **7,** 97 (2015).

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

18. Vongsawan, A. A. *et al.* The genome of *Shigella dysenteriae* strain Sd1617 comparison to representative strains in evaluating pathogenesis. *FEMS Microbiol. Lett.* **362,** pii: fnv011 (2015).

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

27. Mutreja, A. *et al.* Evidence for several waves of global transmission in the seventh cholera pandemic. *Nature.* **477,** 462-465 (2011).

28. Wong, V. K. *et al.* Phylogeographical analysis of the dominant multidrug-resistant H58 clade of *Salmonella* Typhi identifies inter- and intracontinental transmission events. *Nat. Genet.* **47,** 632-639 (2015).

29. Islam, M.S., *et al*. Microbiological investigation of diarrhoea epidemics among Rwandan refugees in Zaire. *Trans. R. Soc. Trop. Med. Hyg*. **89,** 506 (1995).

30. Kumarasamy, K. K. *et al.* Emergence of a new antibiotic resistance mechanism in India, Pakistan, and the UK: a molecular, biological, and epidemiological study. *Lancet Infect. Dis.* **10,** 597-602 (2010).

31. Murray, E. G. D. An attempt at classification of *Bacillus dysenteriae*, based upon an examination of the agglutinating properties of fifty-three strains. *J. R. Army Med. Corps* **31,** 257-271 (1918).

32. Langendorf, C. *et al.* Enteric bacterial pathogens in children with diarrhea in Niger: diversity and antimicrobial resistance. *PLoS One* **10,** e0120275 (2015).

33. Taddei, F., Matic, I. & Radman, M. cAMP-dependent SOS induction and mutagenesis in resting bacterial populations. *Proc. Natl Acad. Sci. USA* **92,** 11736-11740 (1995).

34. Picard, B. *et al*. Mutator natural *Escherichia coli* isolates have an unusual virulence phenotype. *Infect. Immun.* **69,** 9-14 (2001).

35. Grimont, F. & Grimont, P. A. D. Determination of rDNA gene restriction patterns. *Methods Mol. Biol.* **46,** 181-200 (1995).

36. Wirth, T. *et al.* Sex and virulence in *Escherichia coli*: an evolutionary perspective. Mol. Microbiol. **60,** 1136-1151 (2006).

37. Fabre, L *et al*. Chromosomal integration of the extended-spectrum beta-lactamase gene $bla_{\text{CTX-M-15}}$ in *Salmonella enterica* serotype Concord isolates from internationally adopted children. *Antimicrob. Agents Chemother*. **53,** 1808-1816 (2009).

38. Turner, S. A., Luck, S. N., Sakellaris, H., Rajakumar, K. & Adler, B. Molecular epidemiology of the SRL pathogenicity island. *Antimicrob. Agents Chemother.* **47,** 727-734 (2003).

39. Carattoli, A. *et al*. Identification of plasmids by PCR-based replicon typing. *J. Microbiol. Methods* **63,** 219–228 (2005).

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

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

42. Boetzer, M., Henkel, C. V., Jansen, H. J., Butler, D. & Pirovano, W. Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* **27,** 578-579 (2011).

43. Boetzer, M. & Pirovano, W. Toward almost closed genomes with GapFiller. *Genome Biol.* **13,** R56 (2012).

44. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30,** 2068-2069 (2014).

45. Stamatakis, A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22,** 2688-2690 (2006).

46. Carver, T. *et al.* Artemis and ACT: viewing, annotating and comparing sequences stored in a relational database. *Bioinformatics.* **24,** 2672-6 (2008).

47. Croucher, N. J. *et al*. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Res.* **43,** e15 (2015).

48. Tamura, K., Stecher, G., Peterson, D., Filipski A. & Kumar, S. MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Mol. Biol. Evol.* **30,** 2725-2729 (2013).

49. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics* **23,** 127-8 (2007).

50. Letunic, I. & Bork, P. Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy. *Nucleic Acids Res.* **39,** W475-8 (2011).

51. Cheng, L., Connor, T. R., Siren, J., Aanensen, D. M. & Corander, J. Hierarchical and spatially explicit clustering of DNA sequences with BAPS software. *Mol. Biol. Evol.* **30,** 1224-1228 (2013).

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

53. Connor, T.R. *et al.* Species-wide whole genome sequencing reveals historical global spread and recent local persistence in *Shigella flexneri*. *eLife* **4,** e07335 (2015).

54. Baele, G. *et al.* Improving the accuracy of demographic and molecular clock model comparison while accommodating phylogenetic uncertainty. *Mol. Biol. Evol.* **29,** 2157-2167 (2012).

55. Baele, G., Li, W. L., Drummond, A. J., Suchard, M. A. & Lemey, P. Accurate model selection of relaxed molecular clocks in bayesian phylogenetics. *Mol. Biol. Evol.* **30,** 239-243 (2013).

56. Prosperi, M. C. *et al.* A novel methodology for large-scale phylogeny partition. *Nat. Commun.* **2,** 321 (2011).

57. Zankari, E. *et al.* Identification of acquired antimicrobial resistance genes. *J. Antimicrob. Chemother*. **67,** 2640-2644 (2012).

58. Carattoli, A., *et al. In silico* detection and typing of plasmids using PlasmidFinder and plasmid multilocus sequence typing. *Antimicrob. Agents Chemother.* **58,** 3895-3903 (2014).

59. García-Fernández, A. *et al*. Multilocus sequence typing of IncI1 plasmids carrying extended-spectrum beta-lactamases in *Escherichia coli* and *Salmonella* of human and animal origin. *J. Antimicrob. Chemother.* **61,** 1229-1233 (2008).

60. García-Fernández, A. *et al*. Multilocus sequence typing of IncN plasmids. *J. Antimicrob. Chemother*. **66,** 1987-1991 (2011).

61. Page, A. J. *et al*. Roary: Rapid large-scale prokaryote pan genome analysis. *Bioinformatics* **31,** 3691-3693 (2015).

62. Mitchell, A. *et al*. The InterPro protein families database: the classification resource after 15 years. *Nucleic Acids Res.* **43,** D213-D221 (2015).

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

## SUPPLEMENTARY INFORMATION

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

The authors declare no competing financial interests.

Correspondence and requests for materials should be addressed to F.-X.W. (fxweill@pasteur.fr)

## ACKNOWLEDGEMENTS

This study was supported by the Institut Pasteur and the Institut Pasteur International Network, the Institut de Veille Sanitaire, the French government's Investissement d'Avenir programme, Laboratoire d'Excellence 'Integrative Biology of Emerging Infectious Diseases' (grant number ANR-10-LABX-62-IBEID), the Fondation « Le Roch-Les Mousquetaires », the Canetti family through the Georges, Jacques et Elias Canetti Award 2013, the Wellcome Trust through grant 098051 to the Sanger Institute, the NHMRC of Australia (grant 1061409 to K.E.H), the Victorian Life Sciences Computation Initiative (VLSCI) (grant VR0082) and the Indian Council of Medical Research, New Delhi, India. We thank A. Dautry-Versat, A. P. Pugsley, C. Bréchot and J.

Savall for their support; T. Hieu, C. Soto, E. Bourreterre and B. Faye for technical assistance; Z. Szabó, D. Tremblay for providing isolates; L. R. Hiltzik, N. Baldwin and C. Mackenzie for their searches of the archives; M. Toucas, H. d'Hauteville, E. Aldová, S. Formal, and A.T. Maurelli for information about isolates; D. Nedelec for helpful discussion, I. Gut, M. Gut, L. Ma, D. Harris, K. Oliver, and the sequencing teams at the Institut Pasteur and Wellcome Trust Sanger Institute for sequencing the samples. The views expressed in this publication are those of the authors and do not reflect the views of the US Department of the Army or Department of Defense.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

## AUTHOR CONTRIBUTIONS

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., 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.-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., 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., 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 and provided characterized isolates and their epidemiological information. E.N.-N., M.L.-C., I. C., C.R., A.T.-D., M. A.-D. and L.B. did the phenotypic experiments and DNA extractions. A.E.M. and S.R.H provided guidance for genomic analyses. C. Bouchier performed the whole-genome sequencing. M.A. processed the short reads. E.N.-N., N.F., K.K., S. B., K.E.H, J.H, A.J.P., G.G., E.S., and F.-X.W. analysed the genomic sequence

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

## AUTHOR INFORMATION

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



## TABLE

None

## FIGURE LEGENDS

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

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

**Figure 2. Timed phylogeny of a subsample of 125 *Shigella dysenteriae* type 1 isolates.** **a,** Bayesian skyline plot showing temporal changes since 1747 in effective population size (black curve) with 95% confidence intervals (cyan). World War I (WWI) is indicated by a red bar. **b,** Maximum clade credibility tree produced using BEAST (lognormal relaxed clock model; Bayesian skyline) also presenting information about the ortho-nitrophenyl-β-galactoside (ONPG) test. Resistance to nalidixic acid ($NAL^R$) is indicated by a purple circle and resistance to ciprofloxacin ($CIP^R$) is indicated by a purple triangle. Acquisition of the antibiotic resistance element, *Shigella* resistance locus pathogenicity island (SRL-PAI), is indicated by a black thunderbolt. Acquisition of the resistance 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 ciprofloxaxin, 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.

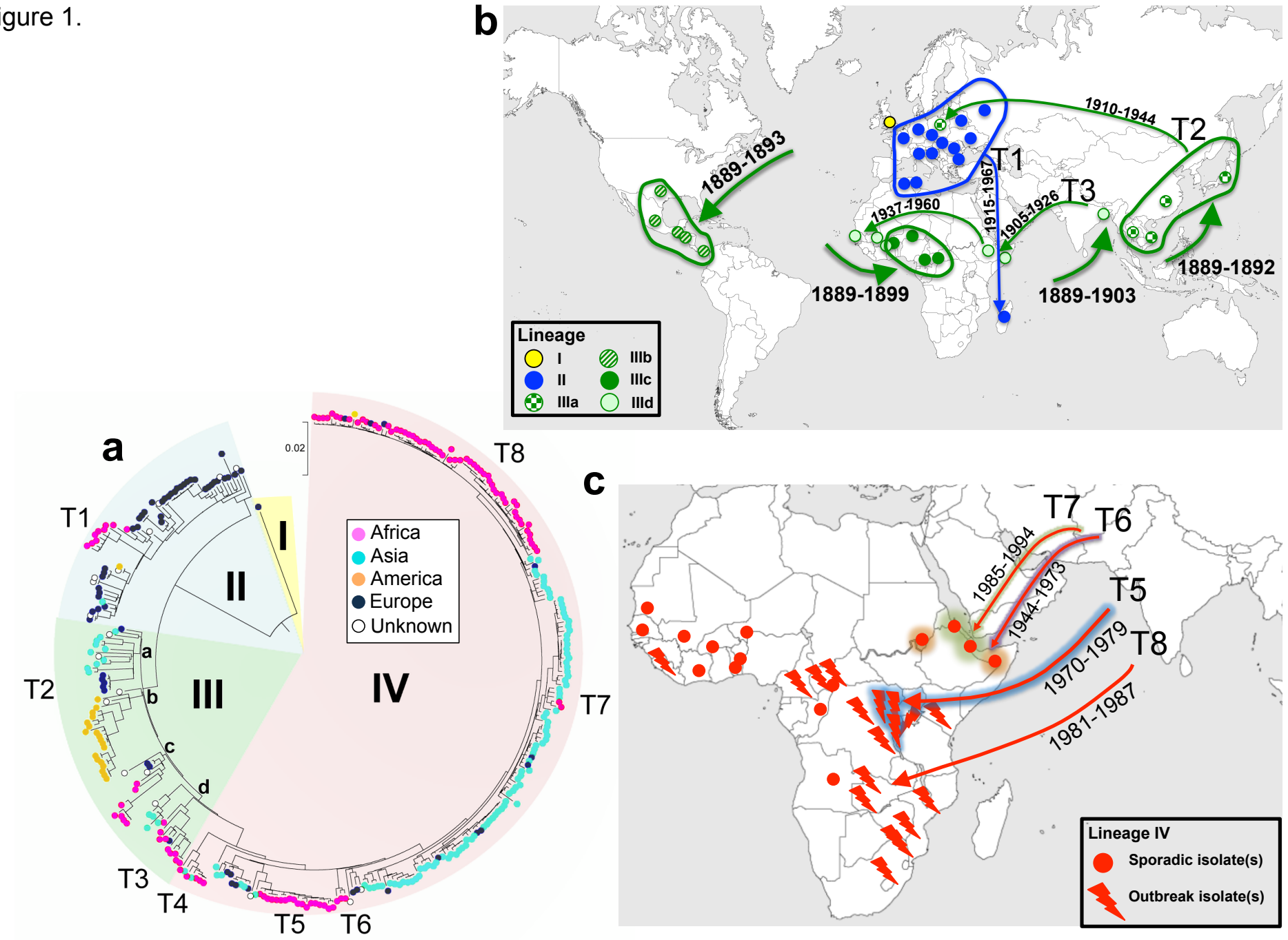
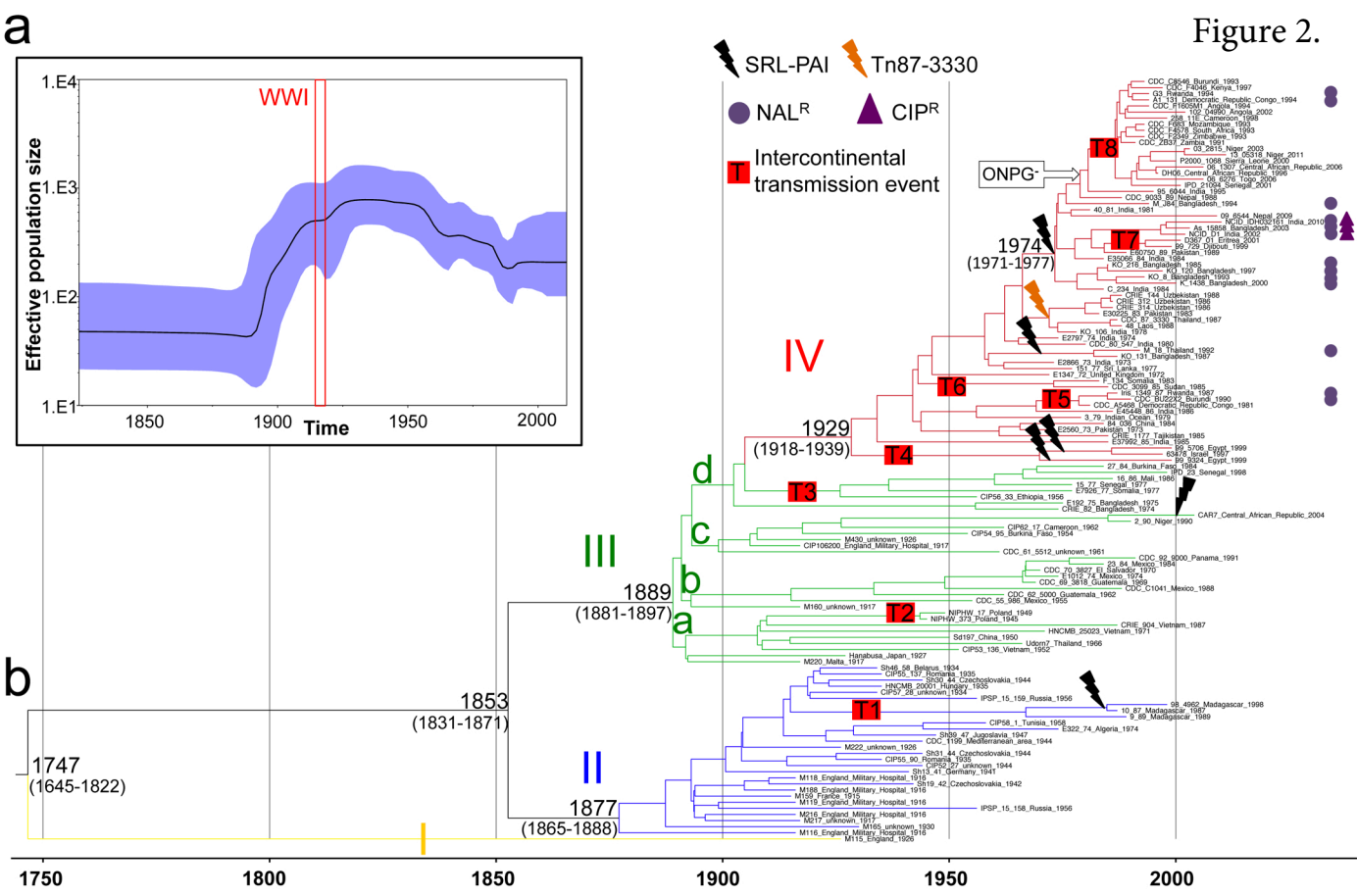

a
0.02
T1
T2
T3
T4
T5
T6
T7
T8
I
II
III
IV
a
b
c
d
Africa
Asia
America
Europe
Unknown
b
1889-1893
1910-1944
T1
T2
T3
1915-1967
1937-1960
1905-1926
1889-1899
1889-1903
1889-1892
Lineage
I
II
IIIa
IIIb
IIIc
IIId
c
T7
T6
T5
T8
1985-1994
1944-1973
1970-1979
1981-1987
Lineage IV
Sporadic isolate(s)
Outbreak isolate(s)

Figure 2.


a
b
SRL-PAI
Tn87-3330
NAL^R
CIP^R
T
Intercontinental transmission event
ONPG-
WWI
Effective population size
1.E4
1.E3
1.E2
1.E1
1850
1900
1950
2000
Time
T8
T7
T6
T5
T4
T3
T2
T1
IV
III
II
d
c
b
a
1974
(1971-1977)
1929
(1918-1939)
1889
(1881-1897)
1853
(1831-1871)
1877
(1865-1888)
1747
(1645-1822)
1750
1800
1850
1900
1950
2000

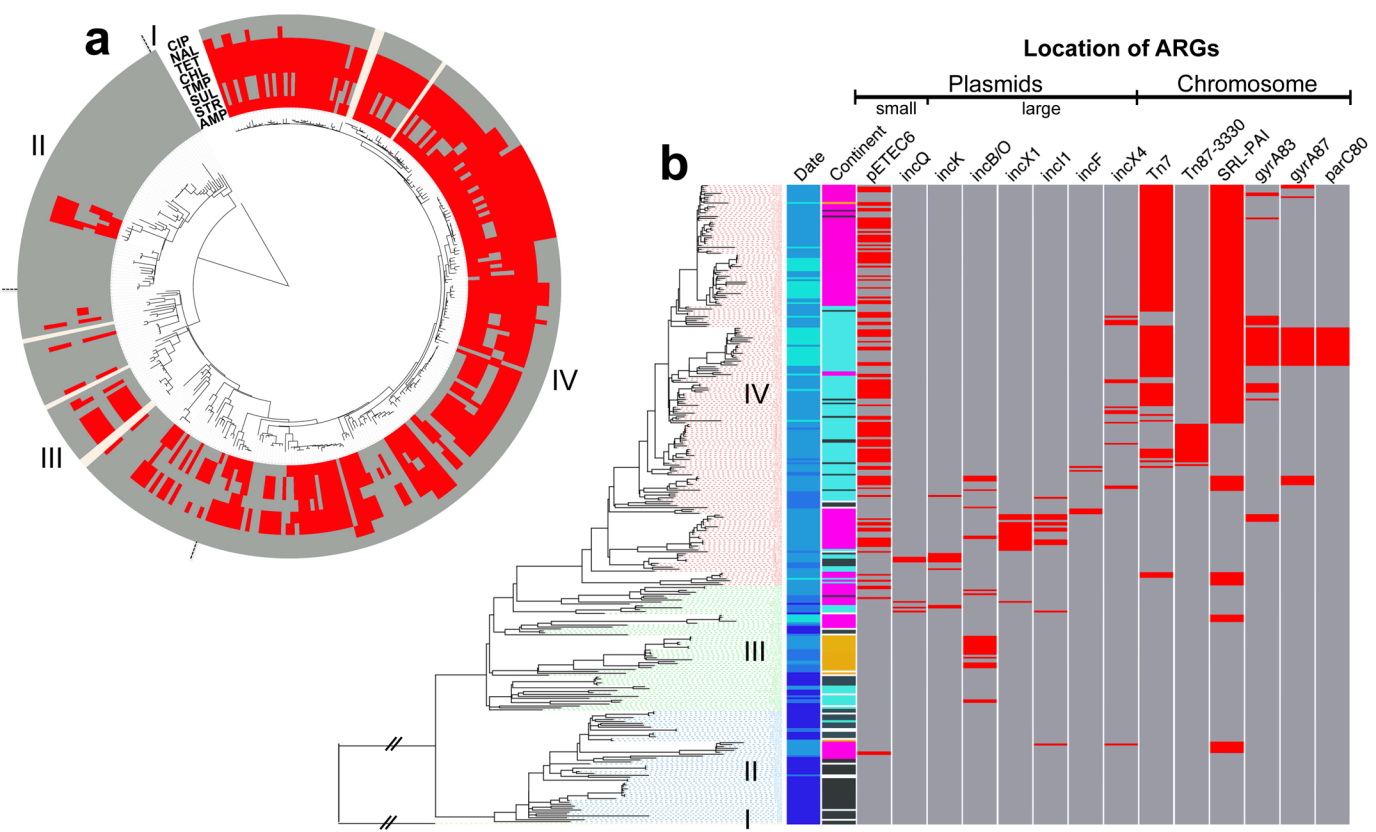

a
I
II
III
IV
CIP
NAL
TET
CHL
TMP
SUL
STR
AMP
b
Location of ARGs
Plasmids
Chromosome
small
large
Date
Continent
pETEC6
incQ
incK
incB/O
incX1
incI1
incF
incX4
Tn7
Tn87-3330
SRL-PAI
gyrA83
gyrA87
parC80
IV
III
II
I
Keys
(Date) 1915-1959 1960-1979 1980-1999 2000-2011
(Continent) Africa America Asia Europe
(Others) Presence Absence

Figure 3.

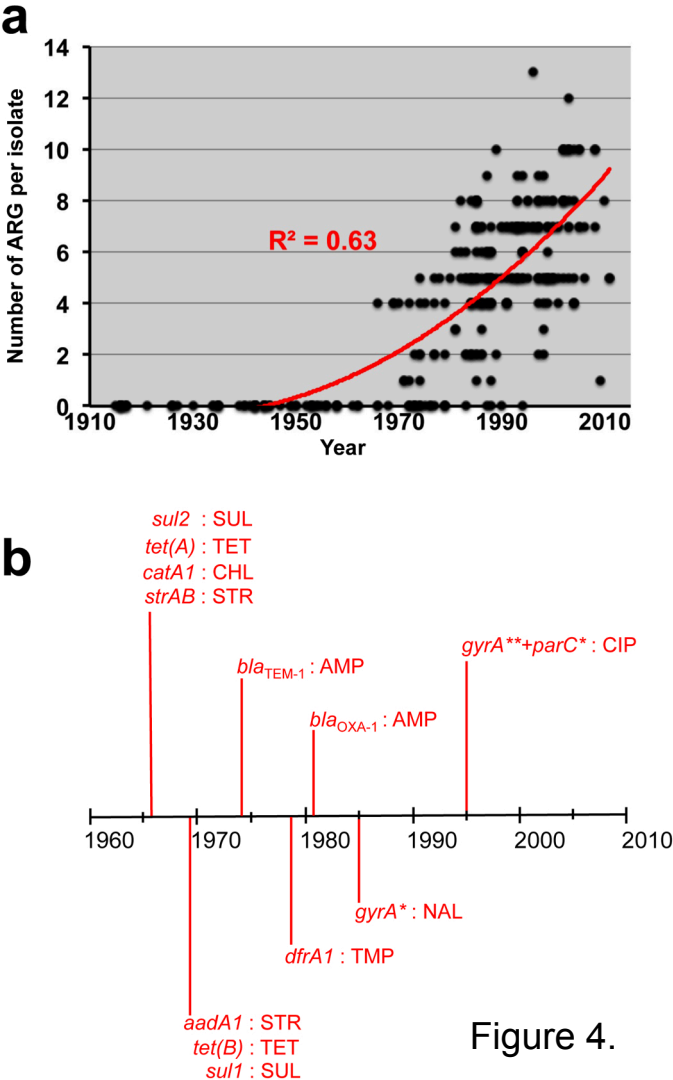

a
Number of ARG per isolate
14
12
10
8
6
4
2
0
R² = 0.63
1910
1930
1950
1970
1990
2010
Year
b
sul2 : SUL
tet(A) : TET
catA1 : CHL
strAB : STR
blaTEM-1 : AMP
blaOXA-1 : AMP
gyrA**+parC* : CIP
1960
1970
1980
1990
2000
2010
gyrA* : NAL
dfrA1 : TMP
aadA1 : STR
tet(B) : TET
sul1 : SUL

Figure 4.