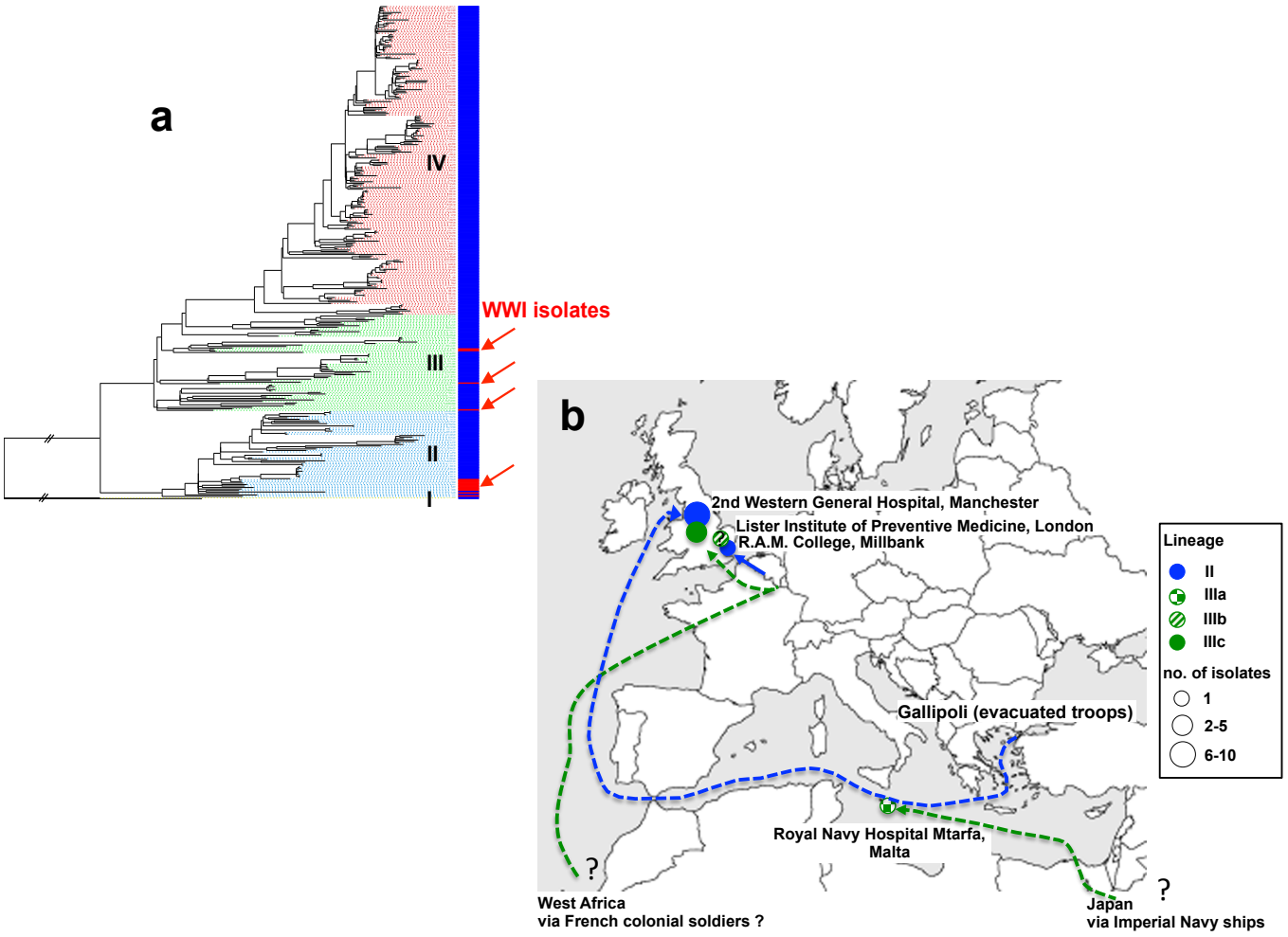
Supplementary Information

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

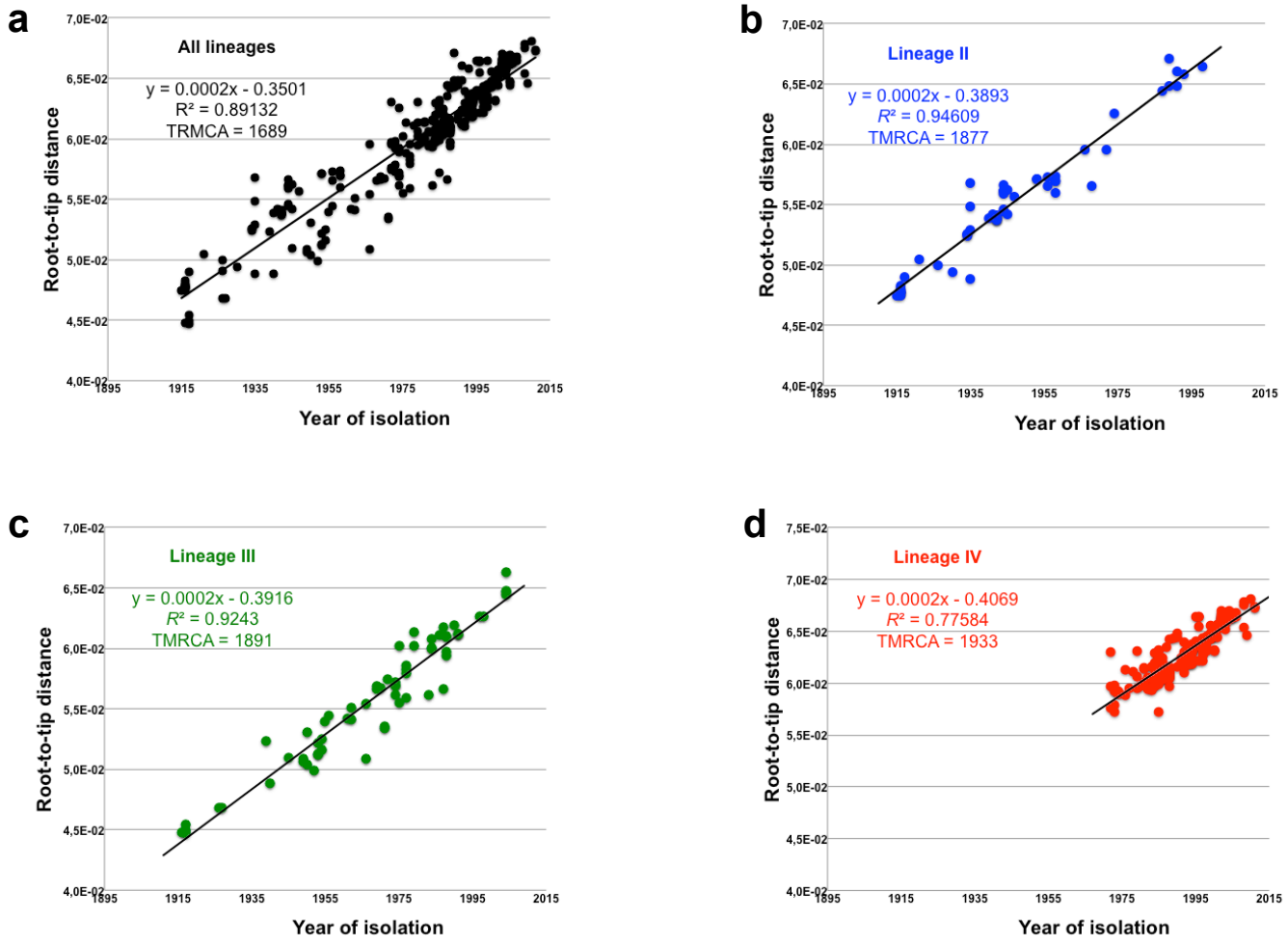
Elisabeth Njamkepo, Nizar Fawal, Alicia Tran-Dien, Jane Hawkey, Nancy Strockbine, Claire Jenkins, Kaisar A. Talukder, Raymond Bercion, Konstantin Kuleshov, Renáta Kolínská, Julie E. Russell, Lidia Kaftyreva, Marie Accou-Demartin, Andreas Karas, Olivier Vandenberg, Alison E. Mather, Carl J. Mason, Andrew J. Page, Thandavarayan Ramamurthy, Chantal Bizet, Andrzej Gamian, Isabelle Carle, Amy Gassama Sow, Christiane Bouchier, Astrid Louise Wester, Monique Lejay-Collin, Marie-Christine Fonkoua, Simon Le Hello, Martin J. Blaser, Cecilia Jernberg, Corinne Ruckly, Audrey Mérens, Anne-Laure Page, Martin Aslett, Peter Roggentin, Angelika Fruth, Erick Denamur, Malabi Venkatesan, Hervé Bercovier, Ladaporn Bodhidatta, Chien-Shun Chiou, Dominique Clermont, Bianca Colonna, Svetlana Egorova, Gururaja P. Pazhani, Analia V. Ezernitchi, Ghislaine Guigon, Simon R. Harris, Hidemasa Izumiya, Agnieszka Korzeniowska-Kowal, Anna Lutyńska, Malika Gouali, Francine Grimont, Céline Langendorf, Monika Marejková, Lorea A. M. Peterson, Guillermo Perez-Perez, Antoinette Ngandjio, Alexander Podkolzin, Erika Souche, Mariia Makarova, German A. Shipulin, Changyun Ye, Helena Žemličková, Mária Herpay, Patrick A.D. Grimont, Julian Parkhill, Philippe Sansonetti, Kathryn E. Holt, Sylvain Brisse, Nicholas R. Thomson, François-Xavier Weill.



Supplementary Fig. 1

Source, phylogenetic clustering, and putative origins of the 14 World War I (WWI) isolates of the Murray collection.

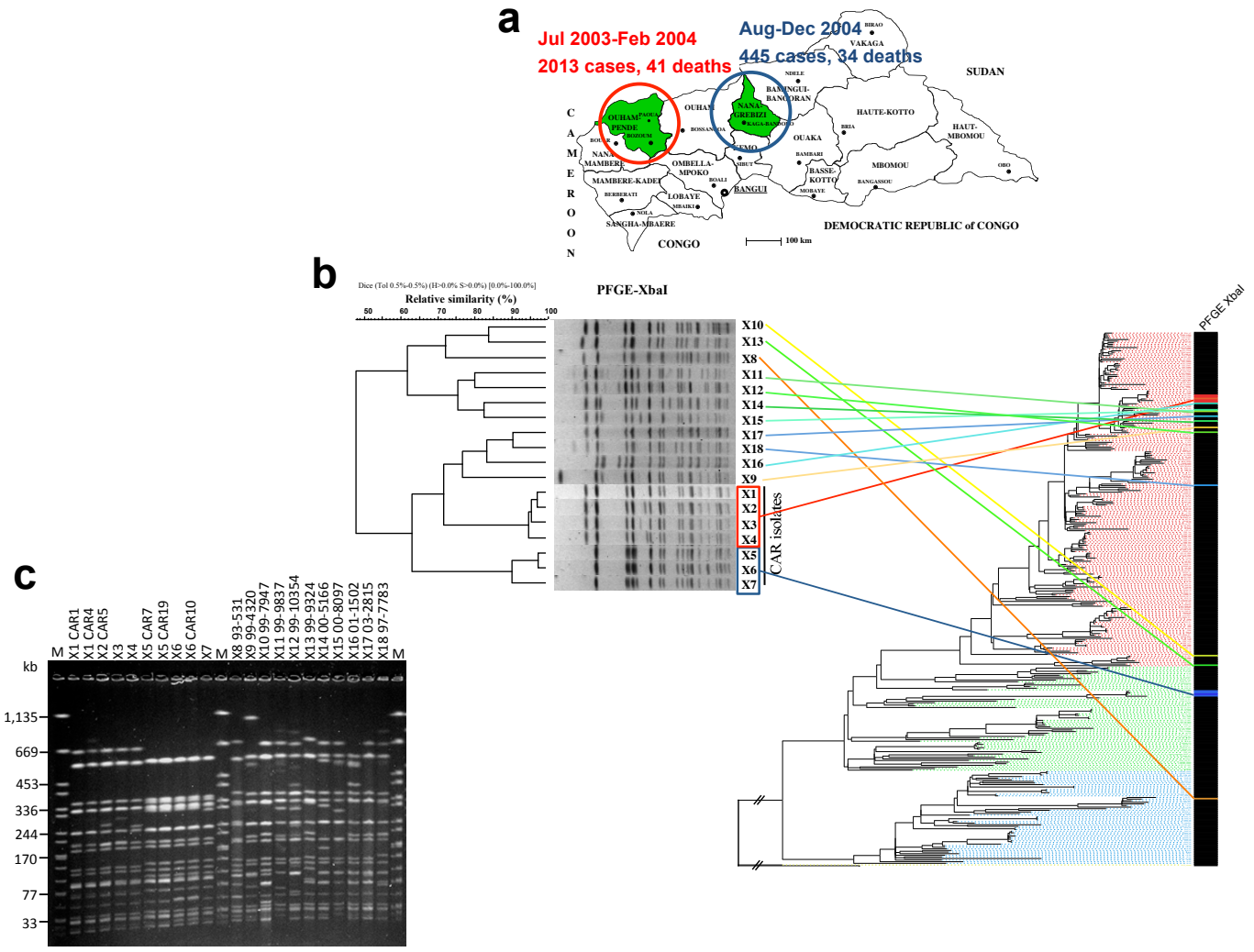
a, The phylogeny shown is the maximum likelihood phylogeny based on 14,677 chromosomal SNPs (reference genome Sd197). **b,** Source of the strains with hypotheses concerning their geographic origin.



Supplementary Fig. 2

Year of isolation vs. root-to-tip distances extracted by Path-O-Gen from an ML phylogeny.

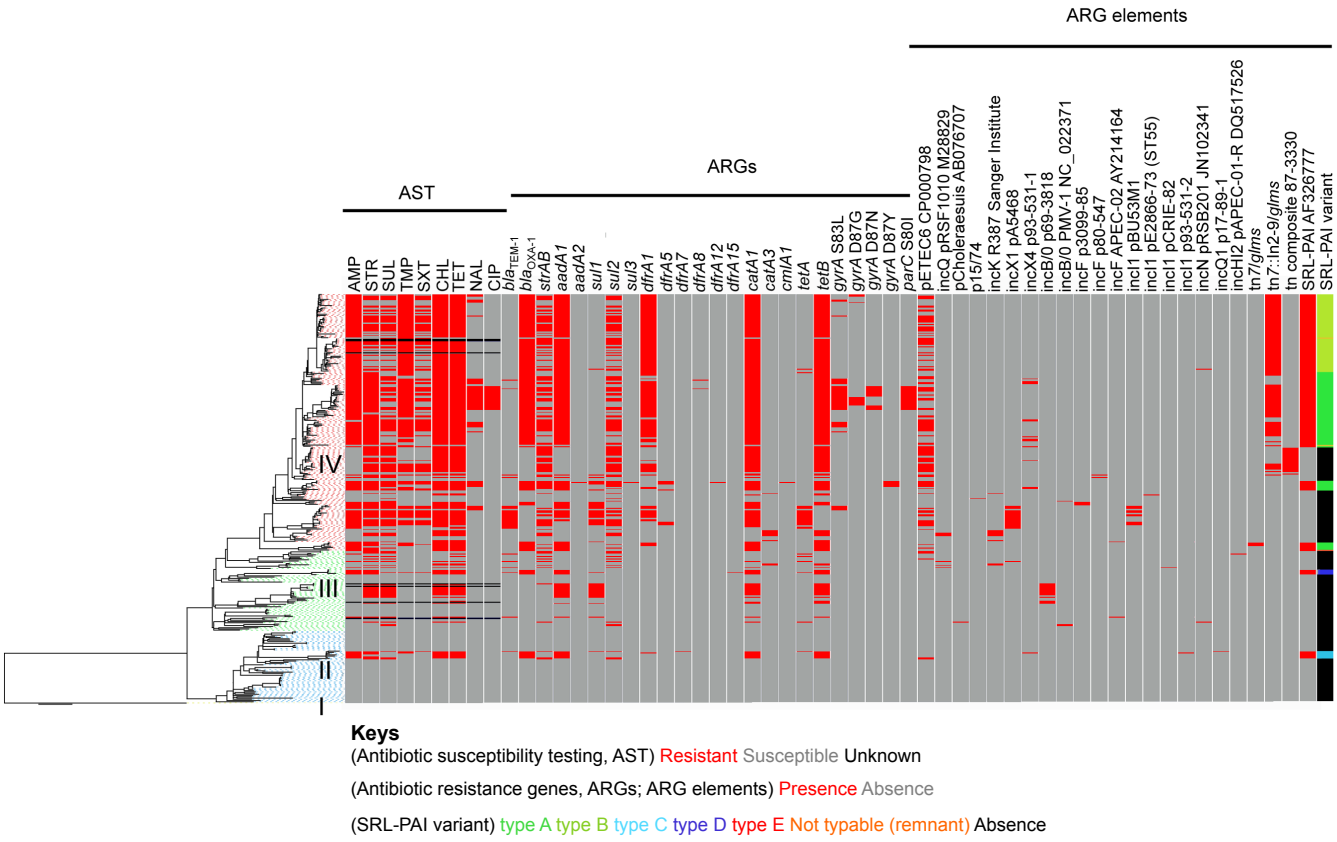
Linear regression line, slope, R^2 correlation coefficient, and time to the most recent common ancestor (TMRCA) are indicated for the whole dataset (panel a) and separately for lineages II to IV (panels b to d). Isolate CDC 3036-94, which was probably acquired during laboratory contamination with an old collection strain, was excluded from the analysis. The maximum likelihood (ML) phylogeny used is based on 14,677 chromosomal SNPs (reference genome Sd197).



Supplementary Fig. 3

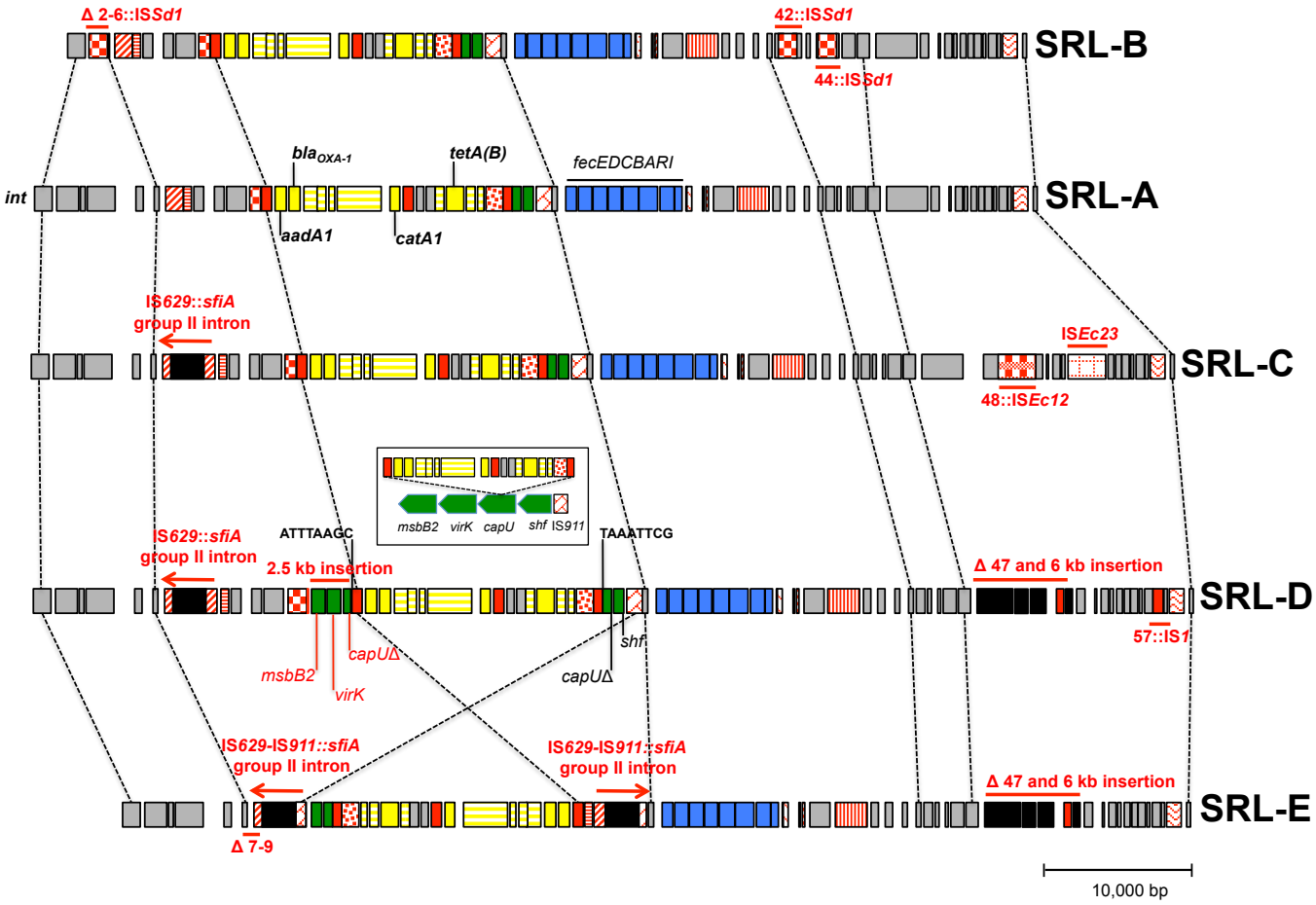
Correlation between pulsed-field gel electrophoresis (PFGE) data and genomic sequences from isolates recovered during two outbreaks in the Central African Republic in 2003-2004.

a, Time span, location, morbidity and mortality of the two outbreaks in the Central African Republic (CAR), according to Bercion *et al.*[25]. **b,** For each isolate analysed by *Xba*I–pulsed-field gel electrophoresis (PFGE), the position within the maximum likelihood phylogenetic tree (reference genome Sd197) is shown. The dendrogram was generated using BioNumerics version 4.0 (Applied Maths, Sint-Martens-Latem, Belgium) and shows the results of cluster analysis on the basis of *Xba*I–PFGE fingerprinting. Similarity analysis was performed using the Dice coefficient, and clustering analysis was performed by using the unweighted pair-group method with arithmetic averages (UPGMA). **c,** Original PFGE gel showing the different *Xba*I–profiles (X1 to X18). The isolates that were whole-genome sequenced are named. *Salmonella enterica* serotype Braenderup H9812 was used as a molecular size marker (M).



Supplementary Fig. 4
Complete antibiotic resistance data.

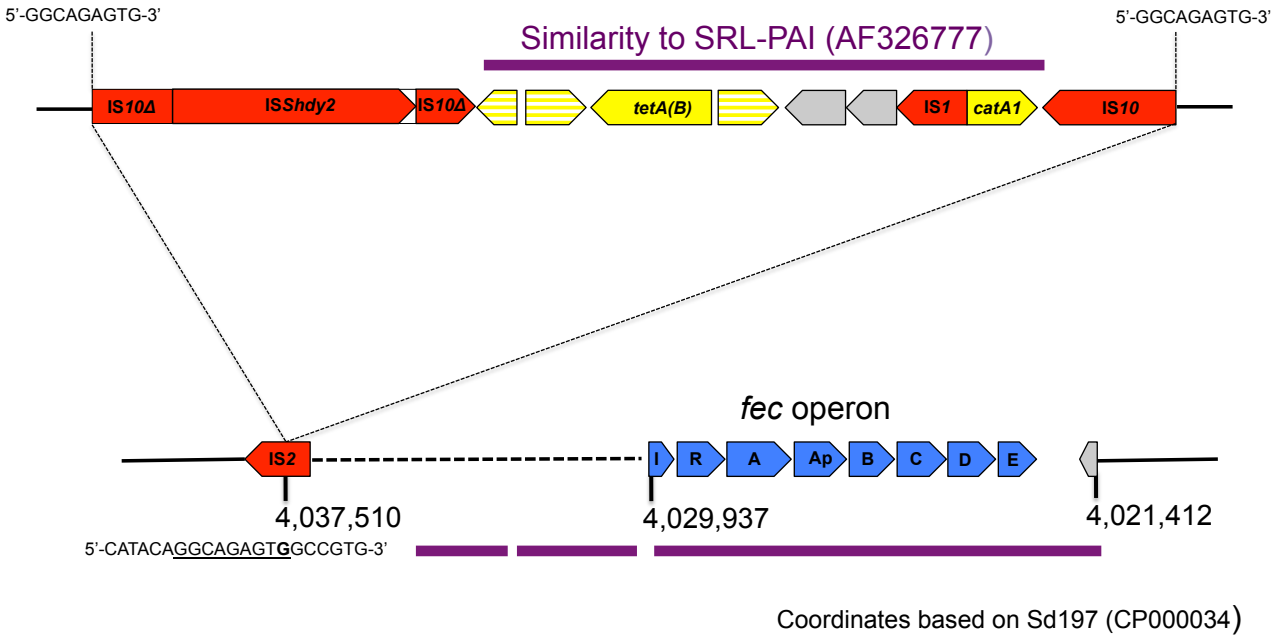
The antibiotic susceptibility testing data (AST), the presence of the different antibiotic resistance genes (ARGs), and ARG-bearing structures are shown for each isolate, according to its position in the phylogeny (maximum likelihood phylogeny based on 14,677 chromosomal SNPs after mapping against the reference genome Sd197). Abbreviations: AMP, ampicillin; STR, streptomycin; SUL, sulfonamides; TMP, trimethoprim; SXT, cotrimoxazole; CHL, chloramphenicol; TET, tetracycline; NAL, nalidixic acid; and CIP, ciprofloxacin. GenBank accession numbers are given after the ARG element name. The sequence of R387 was found at the Wellcome Trust Sanger Institute website. Additional ARG elements were identified in this study. When known, plasmid incompatibility group (inc) is given before the plasmid name.



Supplementary Fig. 5

Structure of the different variants of the *Shigella* resistance locus (SRL) pathogenicity island (SRL-PAI).

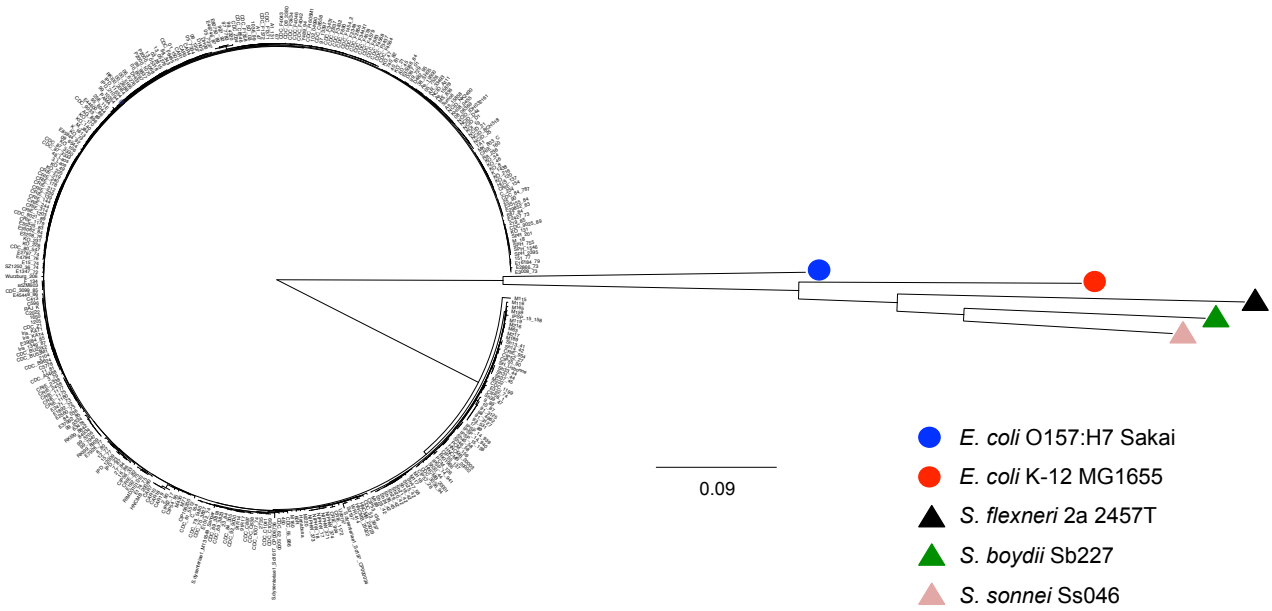
The data were extracted from the PacBio sequences of isolates 40-81 (SRL-A), CDC ZB4 (SRL-B), 17/89 (SRL-C), CAR10 (SRL-D), and 99-9324 (SRL-E). All the *Shigella* resistance locus pathogenicity islands (SRL-PAIs) are inserted into the *serW* tRNA gene. Boxes containing yellow patterns correspond to antibiotic resistance genes. Boxes containing red patterns correspond to insertion sequences. Red labeling indicates insertions/deletions related to SRL-A. The inset shows the structure of the putative ancestral *shf* locus (as observed in various virulence plasmids from *Shigella* spp., including pSD1_197) before the insertion of the SRL which is shown above the *shf* locus. The 8-bp inverted repeats at both ends of the SRL are shown.



Supplementary Fig. 6

Structure of the chromosomally encoded transposon found in the CDC 87-3330 isolate.

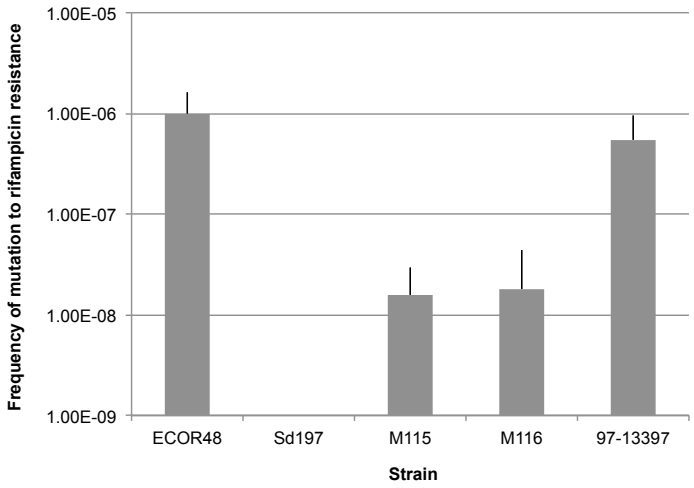
The transposon is shown in the upper part of the figure. Its structure is based on the PacBio sequence of isolate CDC 87-3330. Antibiotic resistance genes are boxed in yellow and insertion sequences are boxed in red. The 8-bp direct repeats at the two ends of the transposon are shown. The chromosomal location of the transposon is shown in the bottom part of the figure, using coordinates based on the *S. dysenteriae* type 1 reference genome Sd197. Regions of similarity to the SRL-PAI are also indicated in purple.



Supplementary Fig. 7

Circular maximum likelihood phylogeny of the sequenced *S. dysenteriae* type 1 genomes rooted on non *S. dysenteriae* genomes.

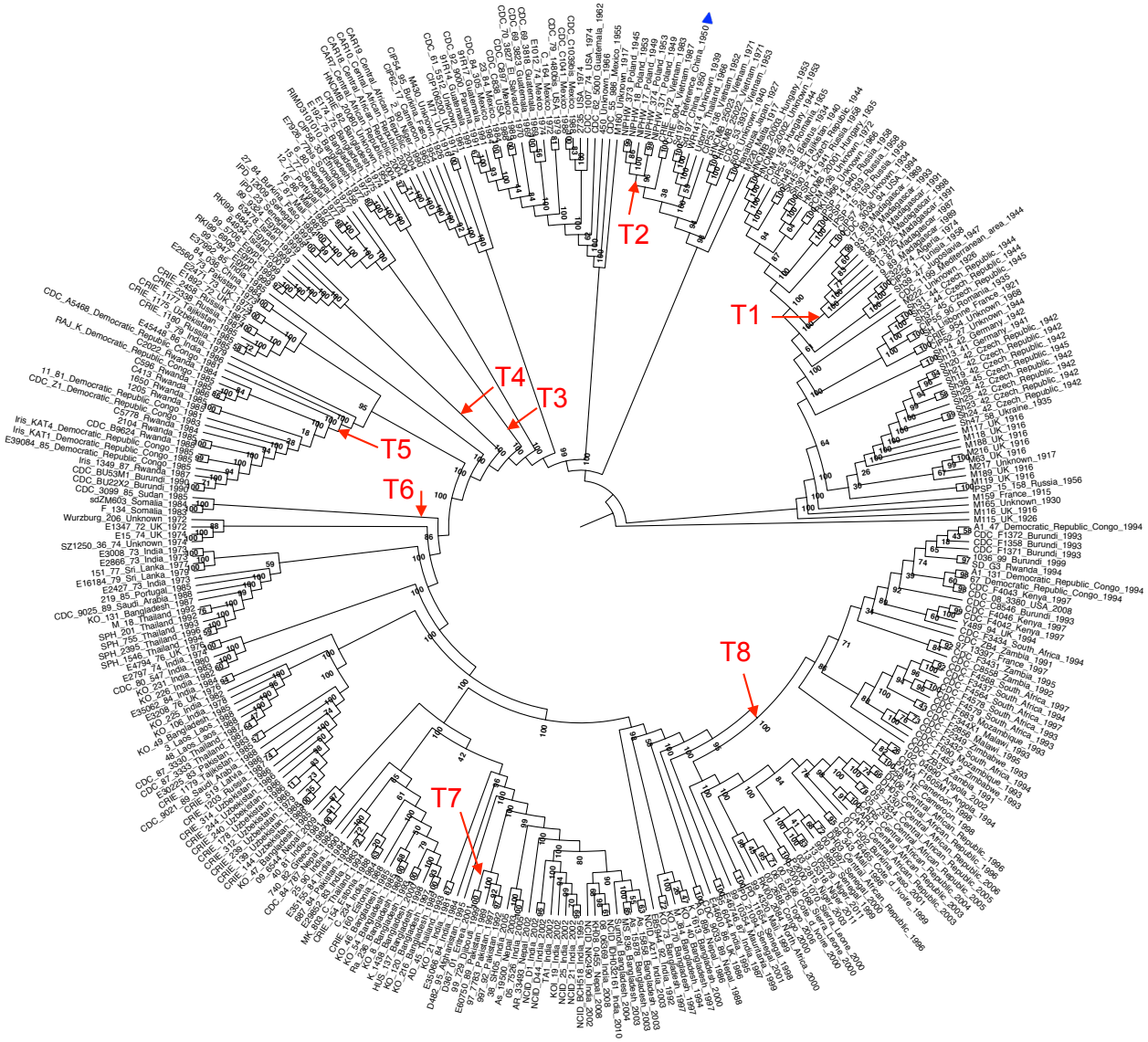
The tree is based on 140,385 SNPs called after mapping against Sd197. The working names of the genomes (See Supplementary Table 1 for the correspondence with isolate names) are shown. The scale corresponds to 12,634 SNPs.



Supplementary Fig. 8

Capacity of M115 to generate mutations conferring resistance to rifampicin.

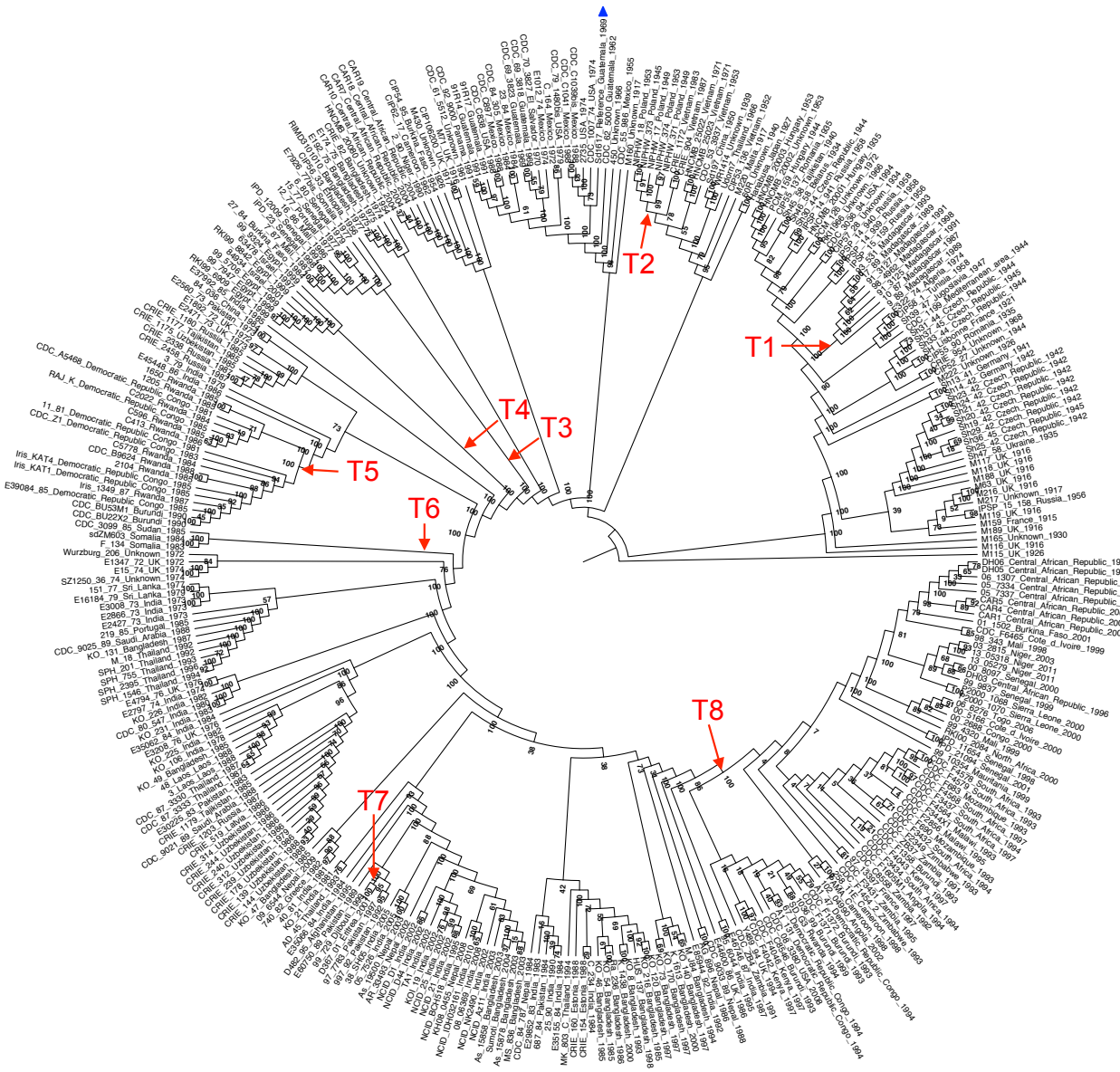
The results are presented as mean values (± standard errors) for two independent experiments. The *E. coli* ECOR48 (CIP 106023) strain was used as a strong mutator positive control, the *S. dysenteriae* type 1 97-13397 isolate was used as a putative strong mutator isolate (deletion of the *mutS* gene), and the *S. dysenteriae* type 1 isolates M116 and Sd197 were used as putative normomutator isolates (integrity of the *mutS*, *mutH*, *mutL* and *uvrD* methyl-directed mismatch repair genes).



Supplementary Fig. 9

Circular maximum likelihood phylogeny of the entire set of sequenced *S. dysenteriae* type 1 genomes.

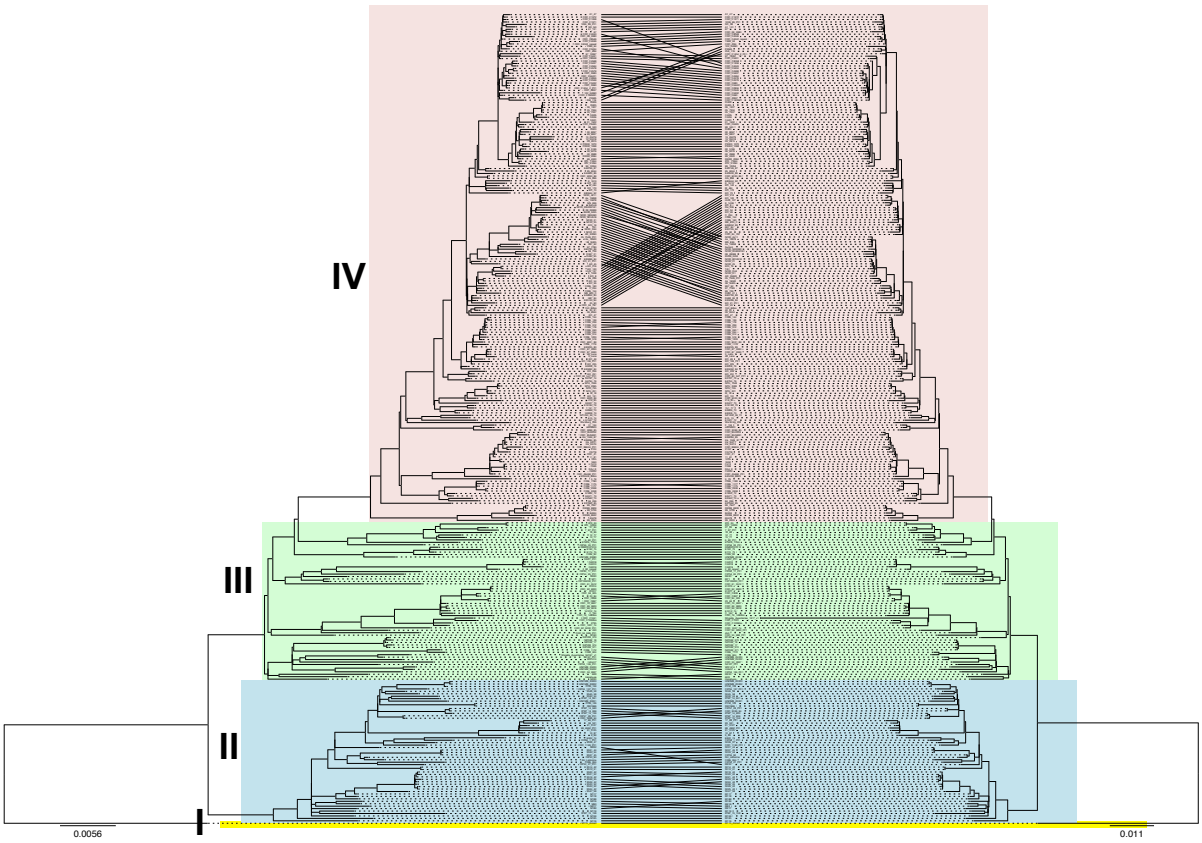
The tree is based on 14,677 SNPs called after mapping against Sd197 (marked with a blue triangle). The tree was rooted on M115, which is most closely related to the *S. dysenteriae* type 1 ancestral strain. The intercontinental transmission events (T1 to T8) and the bootstrap values are shown.



Supplementary Fig. 10

Circular maximum likelihood phylogeny of the entire set of sequenced *S. dysenteriae* type 1 genomes plus the newly published reference genome Sd1617.

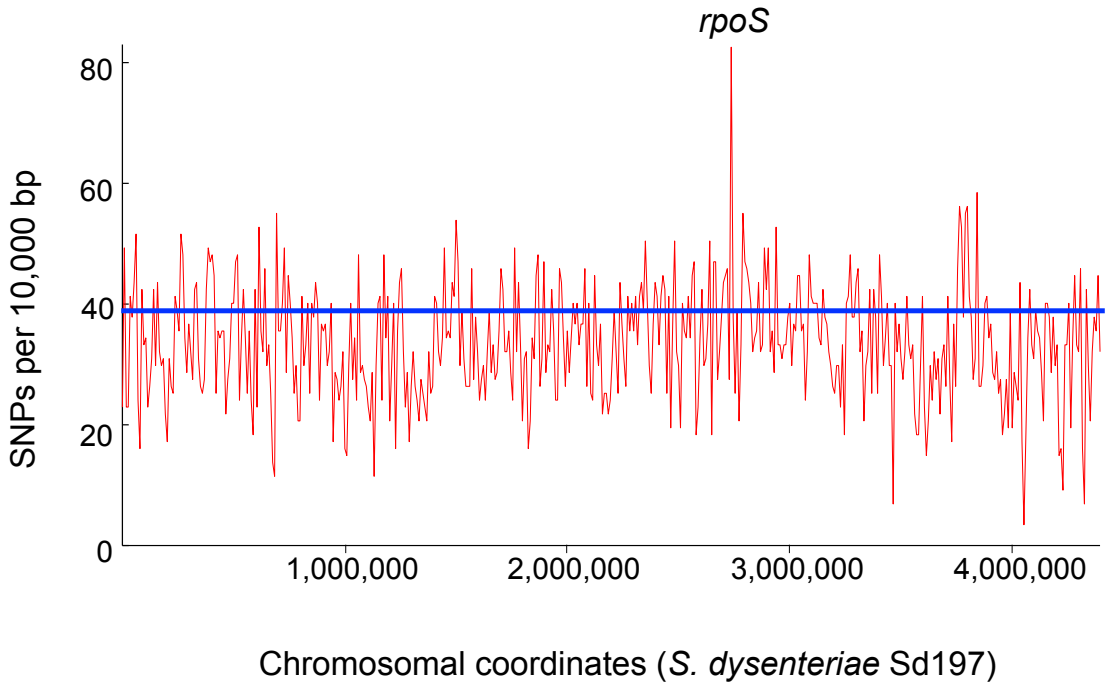
The tree is based on 15,752 SNPs called after mapping against Sd1617 (marked with a blue triangle). The tree was rooted on M115, which is most closely related to the *S. dysenteriae* type 1 ancestral strain. The intercontinental transmission events (T1 to T8) and the bootstrap values are shown.



Supplementary Fig. 11

Comparison of the maximum likelihood trees obtained after mapping against two different *S. dysenteriae* type 1 reference genomes.

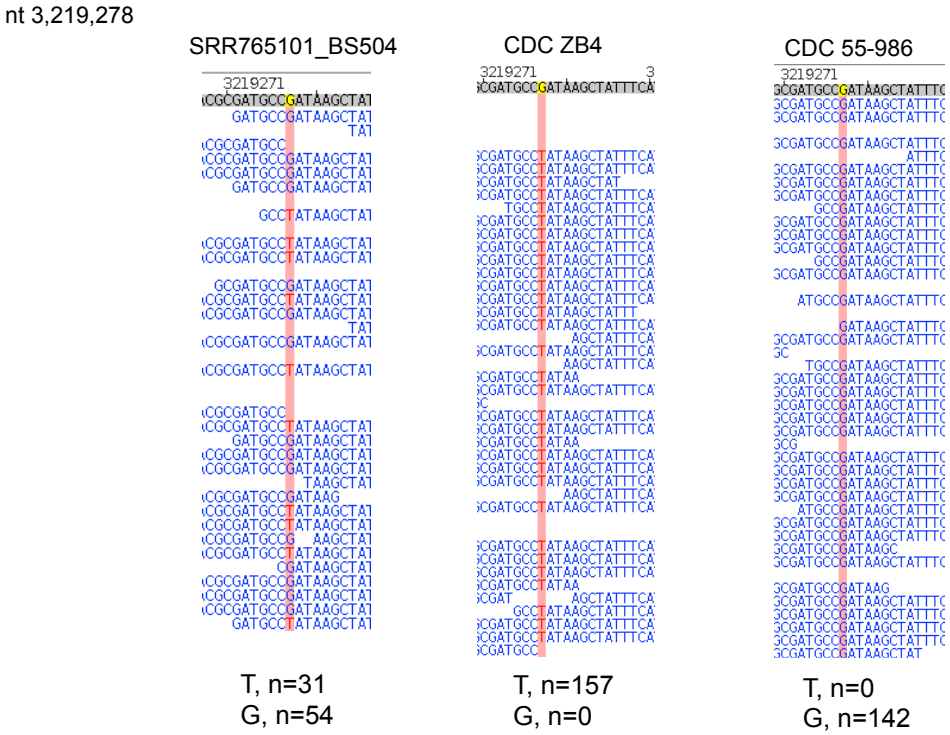
The tree on the left is based on 14,677 SNPs called after mapping against Sd197 (lineage IIIa). The tree on the right is based on 15,752 SNPs called after mapping against Sd1617 (lineage IIIb). Coloured boxes mark each of the lineages (I, II, III, IV), respectively; yellow, blue, green, red.



Supplementary Fig. 12

Distribution of SNPs with respect to the *S. dysenteriae* type 1 reference genome Sd197.

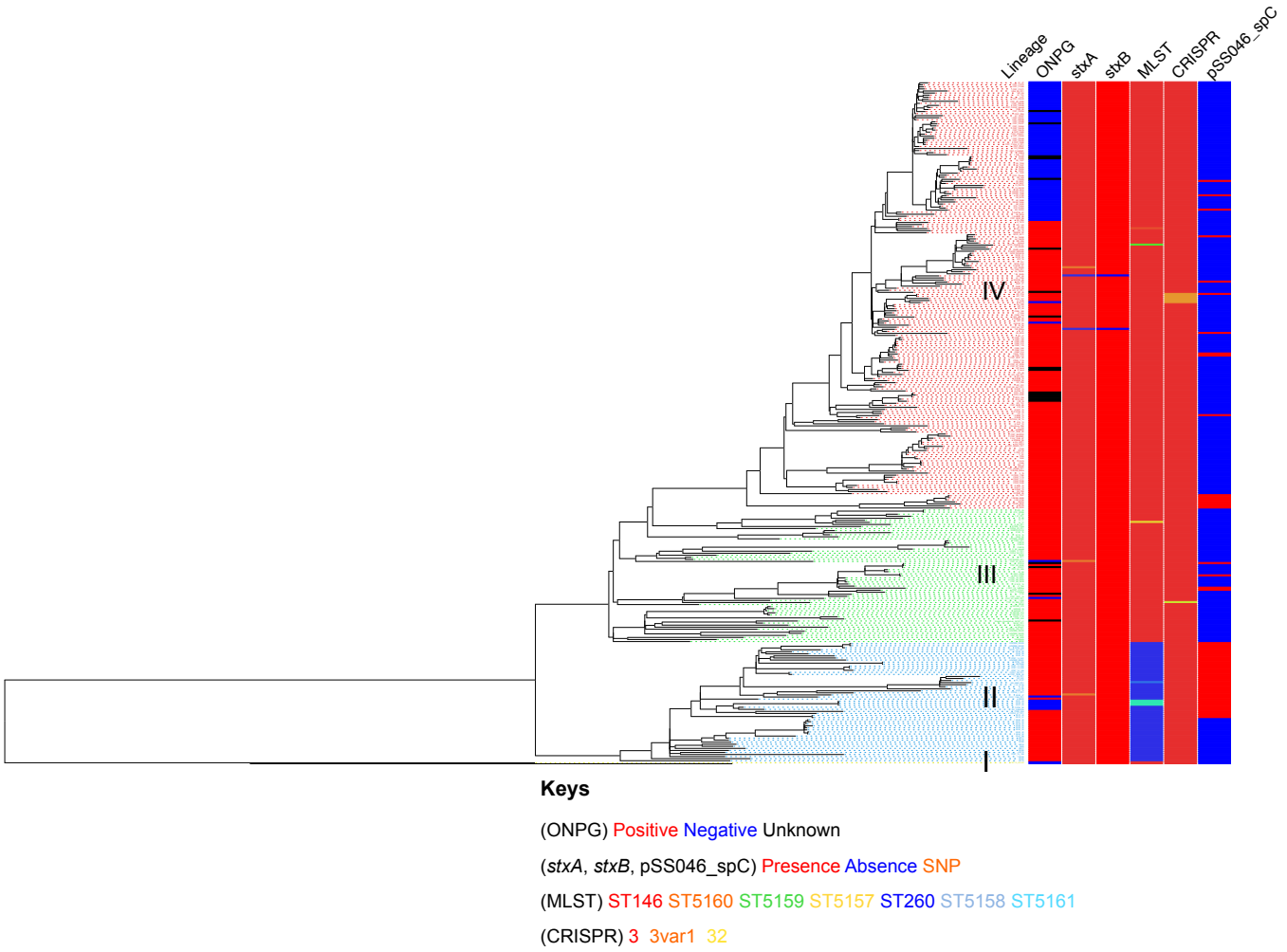
SNP counts per 10,000 bp window are plotted on the *y*-axis. The blue line indicates the mean rate of 39 SNPs per 10,000 bp (or 1 SNP per 256 bp). The peak is due to the *rpoS* gene, which contains 40 SNPs.



Supplementary Fig. 13

Distribution of nucleotides at a key SNP position for the BS504 genome.

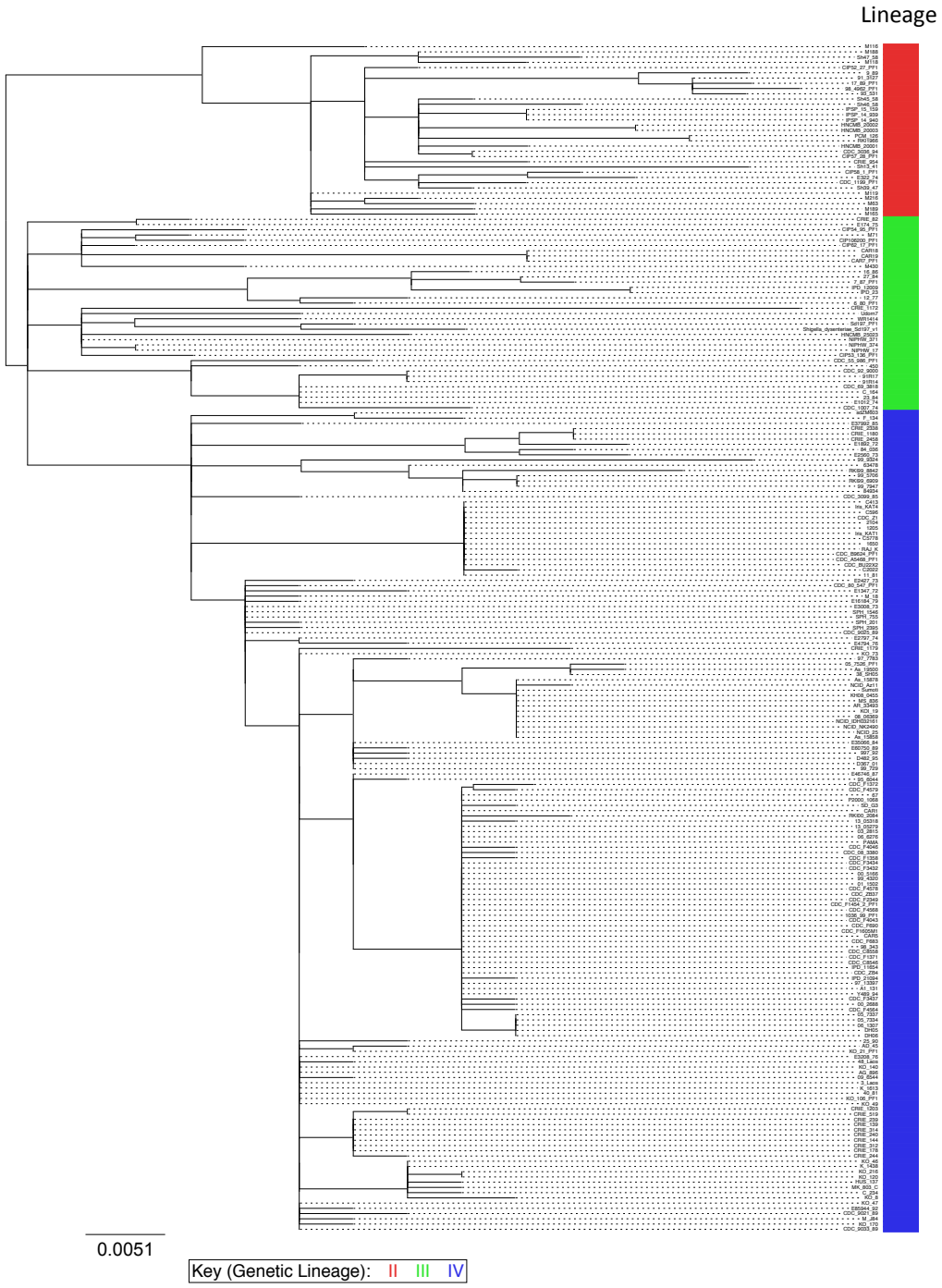
The SNP at position 3,219,278 (T/G) is present in all the African lineage IV isolates of transmission wave T8, but not in other isolates. In the study by Rohmer *et al.*[2], the nucleotide called for this SNP in BS504 was G, despite a frequency of T residues of 36.5% at this position. The read mapping visualised by Artemis[46] at this position is also shown for our isolates CDC ZB4 (the original name of BS504, which was isolated in Zambia in 1991) and CDC 55-986 (Mexico, 1955).



Supplementary Fig. 14

Phenotypic and molecular markers from existing typing and subtyping schemes.

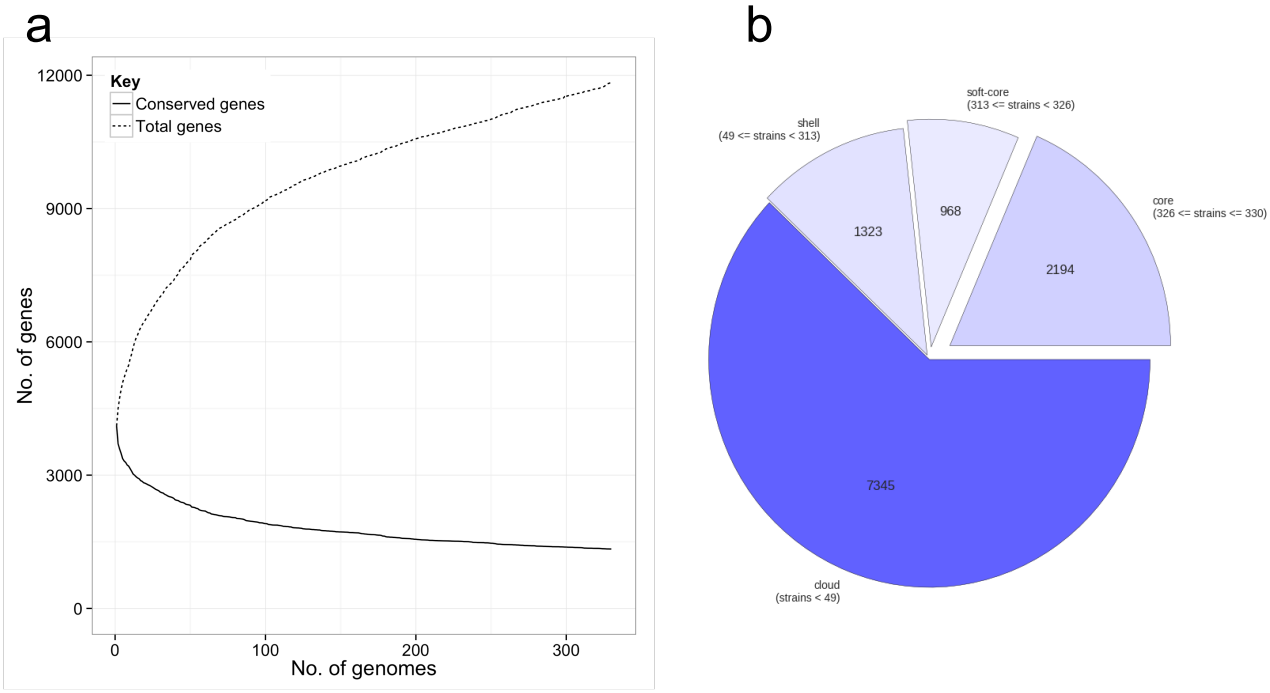
The data are correlated with the maximum likelihood phylogeny based on 14,677 chromosomal SNPs (reference genome Sd197). The four genetic lineages (I, II, III, IV) are indicated by colour, respectively; yellow, blue, green, red. Columns next to the tree show the markers analysed. From left to right: ortho-nitrophenyl-β-galactoside test (ONPG); presence of Shiga toxin genes (*stxA* and *stxB*); presence of plasmid pSS046_spC (pSS046_spC); multilocus sequence type (MLST); and CRISPR spacer content (CRISPR) – see inset legend. The genetic data were obtained principally from whole genome sequences.



Supplementary Fig. 15

Maximum likelihood phylogenetic tree for *S. dysenteriae* type 1 virulence plasmid pSD1_197.

This unrooted tree was constructed from the 226 plasmid-containing isolates. The genetic lineage determined from the maximum likelihood phylogeny for the 14,677 chromosomal SNPs (reference genome Sd197) is indicated in the column on the right.



Supplementary Fig. 16

Pan-genome analysis.

a, Changes to the total genes (dotted) and conserved genes (solid) in the pan-genome with the addition of genomes. **b,** Breakdown of the frequency of genes within isolates where the categories are defined as: core, genes contained in nearly all isolates (>=99%); soft core, genes contained in 95%-99% of the isolates; cloud, genes present only in a few isolates (15%-95%); shell, the remaining genes, present in several isolates (<15%).

Supplementary Tables

Supplementary Table 1 – Details of *Shigella dysenteriae* type 1 isolates and genomes used in this study.

The following are shown: year and country of isolation, epidemiological information; lineage, biotype, antibiotic resistance phenotypes, source and EBI-ENA accession numbers (sheet no. 1); spatiotemporal distribution (sheet no. 2); MLST, CRISPR type, presence/absence of *stxA* and *stxB* genes (sheet no. 3); *gyrA* SNPs and presence/absence of resistance-associated genes and structures (sheet no. 4); distribution of the lineages over time (sheet no. 5).

See separate Excel file

Supplementary Table 2 – Whole-genome sequences, SNPs and phylogenetic data used in this study.

The following are shown: mapping statistics (sheet no. 1); assembly statistics (sheet no. 2); SNPs used for phylogeny (sheet no. 3); pairwise SNP distance between the isolates (sheet no. 4); summary of Bayesian models used for analyses with BEAST (sheet no. 5); date estimates for the main lineages (sheet no. 6); date estimates for the intercontinental transmission events (sheet no. 7); Gene Ontology functions of the 5,630 annotated accessory genes (sheet no. 8).

See separate Excel file

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

Identifier	DNA sequence
A1-var1	CAAGTGATATCCATCATCGCATCCAGTGCGCC
6	CAGCGTCAGGCGTGAAATCTCACCGTCGTTGC
24	TCGGTTCAGGCGTTGCAAACCTGGCTACCGGG
115	TCGGTTCAGGCGTTGCAAACCTAGCTACCGGG
21	GTAGTCCATCATTCCACCTATGTCTGAACTCC

Supplementary Discussion

SNP variation in* S. dysenteriae *type 1

The Maximum likelihood (ML) phylogenetic analysis (RAxML) with *non-S. dysenteriae* outgroups and the BEAST analysis (Fig. 2 and Supplementary Fig. 7) showed that M115 is most closely related to the ancestral *Shigella dysenteriae* type 1 (Sd1) strain. This isolate, which differed from others by more than 1,200 SNPs, was the only representative of lineage I. Otherwise, M115 displayed all the characters of Sd1 in terms of biotyping, serotyping, MLST, CRISPR typing and the presence of the Shiga toxin genes. To confirm that the divergence of M115 was not due to laboratory contamination or a hypermutator phenotype, it was even sequenced a second time from a separate DNA extract, after serial dilution, to ensure that the DNA came from a single colony. The two genomic sequences obtained were identical. M115 displayed no modifications to genes involved in the DNA repair system (*mutS*, *mutH*, *mutL*, and *uvrD*), and it did not have

the hypermutator phenotype (Supplementary Fig. 8). The various ML trees were therefore rooted on M115.

The topology of the two ML trees obtained after mapping short-read sequences against the sequences of the Sd1 reference genomes Sd197 and Sd1617 were similar (Supplementary Figs. 9-11). We therefore used the 14,677 chromosomal SNPs randomly distributed over the non-repetitive non-recombinant core genome detected after mapping against Sd197 (Supplementary Fig. 12).

All BAPS runs converged on three sequence clusters corresponding to lineages II to IV, concordant with the results obtained for clustering by eye, except for M115. This unique lineage I isolate was consistently added to lineage II by BAPS, despite differing from the other isolates in this lineage by more than 1,200 SNPs and having been shown to be the isolate most closely related to the ancestral Sd1 strain by BEAST analysis (see above).

The mean intra-clade pairwise SNP variation within lineages was 275 (minimum 1-maximum 517) for the 58 lineage II isolates, 417 (0-677) for the 64 lineage III isolates, and 192 (0-485) for the 208 lineage IV isolates. Within lineage IV, the isolates from the outbreak in Central Africa in the 1980s (18 T5 isolates isolated from three countries between 1981 and 1990) differed by a mean of 35 SNPs (1-76), whereas the isolates from the emerging ortho-nitrophenyl-β-galactoside negative (ONPG-) African strain (63 T8 isolates isolated from 22 countries between 1991 and 2011) differed by a mean of 56 SNPs (0-133).

We assessed the influence of storage conditions on the rare isolates from a single outbreak recovered by different groups with different culture preservation methods. The CDC A5468 and 11-81 isolates, which originated from the Democratic Republic of the Congo (DRC) and were obtained in 1981, were stored at -80°C or freeze-dried. They differed by only 10 SNPs. The Iris Kat-4 isolate was isolated in DRC in 1985 and stored at -80°C, whereas the E39084/85 isolate was collected in Rwanda in 1985 and stored at room temperature as a stab culture on Dorset egg medium. They also differed by only 10 SNPs. Indirectly, the case of the CDC 3036-94 isolate (see below) also argues for the stability of the SNP pattern. We cannot rule out some differential SNP evolution due to the number of passages before preservation and the mode of storage, but the limited SNP variation observed for 30 year-old isolates and the consistency of the phylogeographic grouping suggest that this was not a major issue in this study.

Other information revealed by the genome sequencing data

Over the last 10 years, we have tried unsuccessfully to find the original cultures established by Shiga in Japan in 1897[5] and Kruse in Germany in 1900[64]. It seems likely that they have been lost or destroyed. However, on the basis of our phylogeographic data, it seems likely that Shiga's isolate would have belonged to sublineage IIIa and that Kruse's isolate would have belonged to lineage II.

In 1986-1987, a multidrug resistant Sd1 strain caused outbreaks in north-eastern Thailand (and also in Laos, according to our isolates) after a lull of 20 years[14]. Based on the distance of this region from India and Bangladesh, and the plasmid and antibiotic

susceptibility patterns of the isolates, the authors concluded that this strain was unlikely to have originated from the Indian subcontinent. However, our results suggest that the sequences of the outbreak isolates were derived from those of bacteria isolated in India from 1978 to 1982, and in Bangladesh in 1985. Furthermore, all these isolates had the same antibiotic resistance structures (i.e., two chromosomally encoded transposons, which are the composite transposon shown in Supplementary Fig. 6 and Tn*7*).

During the early 1990s, a nalidixic acid-resistant strain caused further outbreaks in Thailand, this time close to the Burmese border[65]. Our phylogenetic data indicated that the outbreak strain was derived from a multidrug-resistant but nalidixic acid-susceptible strain isolated in Bangladesh in 1987.

In the former USSR, Sd1 or Grigoriev Shiga's bacillus was highly prevalent from 1917 to 1922, after which its prevalence decreased steadily, reaching negligible levels in the 1950s[66]. It re-emerged during the 1980s in the Central Asian Soviet Republics, including the Uzbek Republic, and it was linked to the Afghanistan war and the flow of populations in Central Asia[67]. The Uzbek SSR served as a springboard for the subsequent spread of infection, in the form of sporadic cases, to cities located in the European part of the USSR (Riga, Moscow, Ulyanovsk, Kuibishev)[67]. Our data reveal that these two periods of activity were associated with two different lineages, European lineage II for the first and South Asian lineage IV (with different subclades) for the second.

The CDC 3036-94 isolate recovered from a child in Tennessee, USA, in 1994 was highly unusual as it belonged to lineage II and was susceptible to all the antimicrobial agents tested. The last such pan-susceptible lineage II isolates on record were recovered from North Africa during the 1970s. Whole-genome sequencing revealed that this case was probably associated with contamination from laboratory stocks, as the genome of CDC 3036-94 differed from CIP 57.28 by only five SNPs. CIP 57.28 was isolated in the UK in 1934 as the "Newcastle" strain or NCTC 4837, 60 years before the CDC 3036-94 sample was isolated. Following its isolation in 1934, this "Newcastle"/NCTC 4837 strain was deposited in various international collections, including CIP under accession no. 57.28 and ATCC under accession no. 13313. The presence in Tennessee of a biotechnology company using ATCC 13313 to prepare rabbit polyvalent antisera provides further support for the hypothesis of laboratory contamination, although no information is available to connect the patient from whom CDC 3036-94 was isolated with the biotechnology company. The possibility of tube switching has also been ruled out as we obtained and sequenced the CDC 3036-94 isolate two years after we sequenced CIP 57.28/ATCC 13313. Furthermore, the genome sequence we obtained for CDC 3036-94 was identical to the publicly available BS506 genome[2] obtained independently by another group from a different stock culture of CDC 3036-94.

Differences found compared to a previous study

In the main text, we show that there was no lack of consistency between phylogeny and geography, as claimed by Rohmer *et al.*[2]. Instead, there were strong phylogeographic patterns. Our study was based on a wide temporal and geographical sampling of more

than 300 isolates, resulting in 14,677 informative SNPs compared to 56 isolates and 989 SNPs for the other group. Their lineage A comprised a single isolate, BS506 (original name CDC 3036-94), found to belong to our old European lineage II. However, as this isolate was recovered in the USA in 1994, and the likely laboratory contamination with an old collection strain (see above) was not identified, this lineage was attributed to the USA and this confused the phylogeographic analysis. Their lineage B comprised only the Sd197 reference genome and corresponded to our Eastern Asian sublineage IIIa. Their lineage C corresponded to our American sublineage IIIb, with, however, two spurious sequences from African isolates (DH02 and BS504), also confusing the phylogeographic analysis (see below). Their lineage D corresponded to our lineage IV. Lineage I and sublineages IIIc and IIId were not found in their study.

The hypothesis of long-term human carriers proposed by Rohmer *et al.*[2] was essentially based on this lack of consistency between phylogeny and geography, as observed for *Salmonella enterica* serotype Typhi[68], the agent of typhoid fever, which may be carried for several decades in the gallbladder of some convalescent patients. The clear pattern of successive transmission waves following the importation of lineage III and IV strains into Africa does not support this hypothesis of long-term carriers for Sd1.

Rohmer *et al*.[2] also claimed that the massive outbreak that hit Central America (estimated 500,000 cases)[10,11] during the late 1960s might have been caused by an African strain that became established in the New World at the beginning of the 1960s. This hypothesis was based on the grouping of two of their African isolates, DH02 and BS504 (original name CDC ZB4), at the base of the C lineage, the tips of which correspond to Central American isolates. As our CDC ZB4 isolate, together with our

258-11E and PAMA isolates collected during an outbreak in Cameroon in 1998[69] (likely the same outbreak as for DH02) clustered within the ONPG- lineage IV T8 subclade, which contains almost all the African isolates obtained since 1991, we analysed the deposited BS504 and DH02 short reads for certain SNPs characteristic of either sublineage IIIb (North and Central American isolates) or the ONPG-negative lineage IV subclade. We observed a heterogeneous distribution of two nucleotides at these positions (Supplementary Fig. 13) indicative of a mixture of two isolates from different genetic backgrounds, explaining the spurious grouping.

Our results demonstrate that sublineage IIIb, containing only North and Central American isolates, had a common ancestor dating back to 1893 [95% credible interval (CI), 1885-1901]. One of the isolates was even isolated in Mexico in 1955, several years before the postulated establishment of the African strain in America. Our genomic data are also consistent with reliable old published reports of the isolation of Sd1 at medical institutions in New England during the early 1900s[70,71] or at a camp for Mexican workers in Michigan in 1938[72].

It was also claimed that the diagnostic tools might be jeopardized by genetic drift affecting metabolic activities as well as surface antigens, some of which were targeted by serotyping. We found, to the contrary, considerable phenotypic and genetic homogeneity in our dataset (see next section), for all the typing and subtyping tools used by clinical and public health microbiology laboratories. The only SNP (within *lacZ*) associated with the loss of a typing character (ONPG test) was a useful marker of the strain that spread across Africa during the 1990s and 2000s.

Correlation of **S. dysenteriae** *type 1 phylogenetic lineages with existing typing and subtyping schemes*

Sd1 is known to contain the *stxA* and *stxB* genes encoding the Shiga toxin STX1, on a defective lambdoid prophage[73]. In a context of the presence of hundreds of insertion sequences (ISs) within the Sd1 genome[17,18], the *stxA* and *stxB* genes have remained remarkably conserved over a period of almost a hundred years. Only two isolates have lost both *stxA* and *stxB*, and three isolates have a SNP within *stxA*. These findings do not reflect a sampling bias, as the identification of Sd1 is based on biochemical tests and serotyping. Searches for *stx* genes or STX production are not carried out routinely for *Shigella* spp. isolates in clinical microbiology or public health laboratories.

The genetically distinct lineages of Sd1 showed only low levels of uncorrelated diversity on assessment with existing subtyping methods: biotyping[74], multilocus sequence typing (MLST)[36], CRISPR typing[22,75], plasmid profiling[76], and pulsed-field gel electrophoresis (PFGE)[25,77] (Supplementary Figs 3 and 14).

The ONPG test was the only conventional phenotypic test to give variable results. This test assesses β-galactosidase activity, which is intense and rapid in Sd1 (generally taking less than 3 hours). The loss of β-galactosidase activity was observed in some isolates from across the tree but was a constant marker for the lineage IV African isolates of the T8 intercontinental transmission wave (Fig. 2). This ONPG- character was first reported in the DRC in 1994[78] but we detected this marker in older African isolates (Zambia,

1991) and in some Indian and Nepalese genomes isolated earlier and genetically ancestral to the T8 African genomes (Fig. 2). These South Asian genomes and the derived African T8 genomes had in common a non-synonymous SNP (C to T at position 363,921 of the reference genome Sd197, leading to a glycine-to-serine substitution) within the *lacZ* gene. The other sporadic ONPG- isolates in the other lineages do not have this non-synonymous SNP. This SNP thus constitutes a good candidate marker for the emerging African strain.

MLST, which has become the gold standard for bacterial population typing, revealed the presence of two main STs, ST260 (*n*=54, lineage II isolates) and ST146 (*n*=270, other lineages), differing by a single SNP in one of the seven 500 bp "housekeeping" genes targeted by this method. In addition, seven genomes belonged to four new STs that were single-locus variants of ST260 and ST146.

CRISPR types were also very stable across the lineages. We analysed the *de novo* assemblies for the different CRISPR spacer sequences and found that all but six genomes belonged to CRISPR type (CT) 3, with the following four spacers: A1-var1, 6, 24, and 21 (Supplementary Table 3). One genome belonged to CT32 (A1-var1, 6, 115, and 21; spacer 115 being a single SNP variant of spacer 24) and five genomes belonged to CT3var1, which differed from CT3 by a single SNP within one direct repeat (DR) within the CRISPR sequence.

Plasmid profiling based on the number and size of plasmids within a single isolate has been widely used for differentiating Sd1 isolates. The independent acquisition of

similarly sized multidrug-resistant plasmids (Supplementary Fig. 4) and the distribution of plasmids not containing ARGs, such as pSS04-spC, across all lineages (Supplementary Fig. 14), preclude the assessment of phylogenetic relationships between isolates by this method. Furthermore, the shift from plasmids to genomic islands as the support for antibiotic resistance in the last two to three decades has probably decreased the plasmid content of isolates.

Over the last two decades, PFGE has become the method of choice for subtyping enteric bacteria at strain level. In light of the genome sequences obtained, we re-evaluated two outbreaks that occurred in the Central African Republic in 2003-2004[25], and which we had investigated with PFGE as a molecular epidemiology tool (Supplementary Fig. 3). PFGE distinguished two groups of profiles, one (PFGE profiles X1 to X4) for the "Ouham-Pende" outbreak and the other (PFGE profiles X5 to X7) for the "Nana-Grebizi" outbreak. Both PFGE groups were tightly clustered, whereas the isolates used for comparison displayed PFGE profiles (X8 to X18) very different from those seen for the isolates of the two outbreaks. Genomic data showed that the strains that had caused the "Nana-Grebizi" and "Ouham-Pende" outbreaks belonged to different lineages, IIIc and IV, respectively, differing by ~700 SNPs. The intra-SNP variation observed among outbreak isolates and giving rise to slightly different PFGE profiles was 5-33 and 10-20 for the "Nana-Grebizi" and "Ouham-Pende" outbreaks, respectively. Most of the genomes of the comparison isolates were actually close to the "Ouham-Pende" genomes (differing from them by 37 to 61 SNPs), despite the lack of relationship suggested by PFGE. This lack of correlation between PFGE and WGS data confirms that PFGE should

not be used for assessing phylogenetic relationships in an organism with a very plastic genome containing hundreds of IS, such as Sd1.

Coevolution between the VP and the chromosome

A large virulence plasmid (VP) was present in 226 isolates. The ML phylogeny based on 290 informative SNPs was similar to the chromosome phylogeny (Supplementary Fig. 15), indicating a coevolution of the chromosome and the VP since at least 1853 (95% CI 1831-1871), the date of the MRCA of all the Sd1 isolates other than M115. The M115 isolate did not contain the VP, according to our search criteria (see Methods section).

Structure of the *cadBA* operon

In *Shigella* spp. and enteroinvasive *E. coli* (EIEC), an inability to synthesise lysine decarboxylase (LDC) and, thus, produce cadaverine has been identified as a convergent pathoadaptive mutation that enhances virulence[79-81]. Comparative analysis has shown that the ancestral LDC trait was lost through various rearrangements of the *cadBA* operon encoding LDC and its transporter. We analysed this operon from the PacBio sequences of nine Sd1 isolates from lineages I (M115), II (M116 and 17/89), IIIa (Sd197), IIIb (CDC 69-3818), IIIc (CAR10), and IV (40-81, CDC ZB4 and 99-9324). A similar structure was found in all these isolates. The *cadAB* operon was located between *ytfQ* (SDY_4463 of Sd197; GenBank accession no. CP000034) and *yjdL* (SDY_4467). The *cadA* gene (SDY_4466) displayed a five-nucleotide deletion leading to a frameshift with a premature stop codon, and the *cadB* gene (SDY_4465) was interrupted by an IS*1*.

The *cadC* gene, a regulator of the *cadAB* operon was absent. In addition, an IS*1* element was found inserted into the *cadA* gene of M115 and CDC 69-3818 and a second frameshift was found at the end of the *cadA* gene in the lineage IV isolates 40-81, CDC ZB4 and 99-9324.

Pan-genome and antibiotic resistance

The pan-genome analysis (Supplementary Fig. 16) identified a total of 11,830 genes for the 330 Sd1 genomes studied. A core genome of 2,194 genes was identified, comprising 1,132,109 bases. Of the 7,345 accessory genes, 5,630 were annotated with 22,135 Gene Ontology (GO) terms. The top GO terms corresponded to DNA/plasmid binding (GO:0003677) and transposition, DNA-mediated functions (GO:0006313) (Supplementary Table 2). Taking into account the various large multidrug-resistant plasmids (70 to 160 genes per plasmid) we have sequenced by 454 or PacBio, together with other plasmid fragment sequences obtained by Illumina short-read sequencing, we can conclude that the number of genes in the accessory genome supporting antibiotic resistance probably exceeds 1,000.

The accessory genome linked to antibiotic resistance was maintained in the descendants, whereas structures not involved in antibiotic resistance, such as prophages, were found in only one or a few isolates and were not studied further.

Cotrimoxazole, a combination of sulfamethoxazole and trimethoprim, has been widely used to treat *Shigella* infections since the late 1960s, when the first multidrug-resistant strains appeared. The first cotrimoxazole-resistant Sd1 strains were isolated on the Indian

subcontinent in the early 1980s[82]. The *dhfrI* (or *dfrA1*) gene was found in all the isolates, either on a 20-MDa plasmid for the ampicillin-resistant isolates, or chromosomally encoded for the ampicillin-susceptible isolates. Haider *et al.*[82] thought that there might be a transposition of this *dhfrI* gene between the 20-MDa plasmid and the chromosome. Our first cotrimoxazole-resistant Sd1 isolate was obtained in India in 1978. It contains the *dfrA1* gene on a Tn*7*-like transposon integrated into the chromosome. The class 2 integron did not contain the *aadA1* gene (encoding resistance to streptomycin and spectinomycin), as for the classical Tn*7*[83]. The *intI2-dfrA1-sat2-orfX-ybfA-ybfB-ybgA-tnsE-tnsD-tnsC-tnsB-tnsA* genes were found to be present and the transposon was named Tn*7*::In2-9, in accordance with the nomenclature of ref. 83. Tn*7*::In2-9 was also found in 114 other Sd1 isolates, all from lineage IV. A classical Tn*7* was found in three Egyptian isolates from 1999. A 30-kb IncX4 plasmid containing *dfrA1* was also found in 14 Sd1 isolates (one from lineage II and 13 from lineage IV), none of which contained the chromosomal Tn*7* or Tn*7*::In2-9. However, the plasmid *dfrA1* gene was in a class 2 integron (In2-9) with no trace of the Tn*7* transposition module, and could not, therefore, have transposed to the chromosome as suggested by Haider *et al.*[82]. In Africa, our first isolate resistant to cotrimoxazole was isolated in the DRC in 1983. Resistance to this antibiotic was observed in 1981, less than two years after the start of the so-called "Zairian" outbreak caused by a strain initially resistant to ampicillin, chloramphenicol, tetracycline, streptomycin and sulfonamides[84]. The initial multidrug resistance was due to a 50-kb IncX1 plasmid encoding resistance to ampicillin, chloramphenicol, and tetracycline (pA5468) and a 6-kb plasmid encoding resistance to streptomycin and sulfonamides (pETEC6). Resistance to cotrimoxazole was conferred by a *dfrA1* gene encoded on a 110-kb IncI1 pST186 plasmid (pBU53M1).

In our collection, the two oldest nalixidic acid-resistant Sd1 isolates were obtained in the DRC and Bangladesh, both in 1985. However, the first reported isolation of Sd1 isolates with this pattern of resistance was in April 1982 in the DRC[84]. This isolation occurred less than one year after the introduction of nalidixic acid as first-line therapy during the "Zairian" outbreak, in which isolates rapidly became resistant to cotrimoxazole.

The next step in the development of a multidrug resistance profile was the acquisition of resistance to ciprofloxacin, a fluoroquinolone, mediated by a double mutation in *gyrA* (S83L and a second mutation in codon 87) and a mutation in the topoisomerase IV *parC* gene (S80I). In our dataset, resistance to ciprofloxacin was acquired only once, in a group of 20 isolates from the Indian subcontinent collected between 1995 and 2010 (MIC ciprofloxacin 4-12 mg/L). This is consistent with published reports of an emergence of ciprofloxacin-resistant Sd1 in West Bengal in 2002, after a hiatus of 14 years in which Sd1 was not isolated[85,86]. A PFGE approach showed that the ciprofloxacin-resistant Sd1 isolates were clonal, a finding subsequently confirmed by whole-genome sequencing on a larger sample. However, we identified an internal branch corresponding to seven isolates with a mutation of codon 87 (D87G) other than the predominant mutation (D87N). Six of these seven isolates were collected in Bengal in 2002, during the Diamond Harbor and Siliguri outbreaks[86]. Similarly, two different mutations in codon 87 of *gyrA* were previously identified in genetically related ciprofloxacin-resistant enteric bacterial pathogens of the *S. enterica* serotype Kentucky ST198-X1, a bacterium subject to high levels of fluoroquinolone selection pressure in the poultry industry[87].

Supplementary References

64 Kruse, W. Ueber die Ruhr als Volkskrankheit und ihren Rrreger. *Deutsche Med. Woch*. **26***,* 637-639 (1900).

65 Hoge, C. W., Bodhidatta, L., Tungtaem, C. & Echeverria, P. Emergence of nalidixic acid-resistant *Shigella dysenteriae* type 1 in Thailand: an outbreak associated with consumption of a coconut milk dessert. *Int. J. Epidemiol.* **24,** 1228-1232 (1995).

66 Krasheninnikov, O. A. [Features of the geographic distribution of Shigellae. I. Changes in the etiologic structure of dysentery in Russia and the USSR (1900-1950)]. Russian. *Zh. Mikrobiol. Epidemiol. Immunobiol.* **45,** 21-31 (1968).

67 Solodovnikov, I. u. P. *et al*. [The epidemiological characteristics of the spread of Grigor'ev-Shiga dysentery in the territories of the former USSR in recent years]. Russian. *Zh. Mikrobiol. Epidemiol. Immunobiol.* **1,** 31-36 (1994).

68 Holt, K.E., *et al*. High-throughput sequencing provides insights into genome variation and evolution in *Salmonella* Typhi. *Nat. Genet.* **40,** 987-993 (2008).

69 Cunin, P. *et al*. An epidemic of bloody diarrhea: *Escherichia coli* O157 emerging in Cameroon? *Emerg. Infect Dis.* **5,** 285-290 (1999).

70 Vedder EB, Duval CW. The etiology of acute dysentery in the United States. *J. Exp. Med.* **6,** 181-205 (1902).

71 Hiss, P. H. On fermentative and agglutinative characters of bacilli of the "Dysentery Group". *J. Med. Res.* **13,** 1-51 (1904).

72 Block, N. B. & Ferguson, W. An outbreak of Shiga dysentery in Michigan, 1938. Am. *J. Public Health Nations Health* **30,** 43-52 (1940).

73 Greco, K. M., McDonough M. A. & Butterton J. R. Variation in the Shiga toxin region of 20th-century epidemic and endemic *Shigella dysenteriae 1* strains. *J. Infect. Dis.* **190,** 330-334 (2004).

74 Le Minor L, Richard C. Méthodes de laboratoire pour l'identification des entérobactéries. Paris, France: Institut Pasteur; 1993. pp. 72–78.

75 Touchon, M. & Rocha, E. P. The small, slow and specialized CRISPR and anti-CRISPR of *Escherichia* and *Salmonella*. *PLoS One* **5**, e11126 (2010).

76 Haider, K., Kay B. A., Talukder, K.A. & Huq, M. I. Plasmid analysis of *Shigella dysenteriae* type 1 isolates obtained from widely scattered geographical locations. *J. Clin. Microbiol.* **26,** 2083-2086 (1988).

77 Talukder, K. A, Dutta, D. K. & Albert, M.J. Evaluation of pulsed-field gel electrophoresis for typing of *Shigella dysenteriae* type 1. *J. Med. Microbiol.* **48,** 781-784 (1999).

78 Cavallo, J. D., Niel, L., Talarmin, A. & Dubrous, P. [Antibiotic sensitivity to epidemic strains of *Vibrio cholerae* and *Shigella dysenteriae* 1 isolated in Rwandan refugee camps in Zaire]. French. *Med. Trop.* **55,** 351-353 (1995).

79 Maurelli, A.T., Fernández, R.E., Bloch, C.A., Rode, C.K., & Fasano A. "Black holes" and bacterial pathogenicity: a large genomic deletion that enhances the virulence of *Shigella* spp. and enteroinvasive *Escherichia coli*. *Proc. Natl Acad. Sci. U S A*. **95,** 3943-3948 (1998).

80 Casalino, M., Latella, M.C., Prosseda, G., & Colonna, B. CadC is the preferential target of a convergent evolution driving enteroinvasive *Escherichia coli* toward a lysine decarboxylase-defective phenotype. *Infect. Immun.* **71,** 5472-5479 (2003).

81 Day, W.A. Jr, Fernández, R.E., & Maurelli, A.T. Pathoadaptive mutations that enhance virulence: genetic organization of the cadA regions of *Shigella* spp. *Infect. Immun.* **69,** 7471-7480 (2001).

82 Haider, K. *et al.* Trimethoprim resistance gene in *Shigella dysenteriae* 1 isolates obtained from widely scattered locations of Asia. *Epidemiol. Infect.* **104,** 219-228 (1990).

83 Ramírez, M. S., Piñeiro, S., Argentinian Integron Study Group & Centrón, D. Novel insights about class 2 integrons from experimental and genomic epidemiology. *Antimicrob. Agents Chemother.* **54,** 699-706 (2010).

84 Rogerie, F. *et al.* Comparison of norfloxacin and nalidixic acid for treatment of dysentery caused by *Shigella dysenteriae* type 1 in adults. *Antimicrob. Agents Chemother.* **29,** 883-886 (1986).

85 Dutta, S. *et al. Shigella dysenteriae* serotype 1, Kolkata, India. *Emerg. Infect. Dis*. **9,** 1471-1474 (2003).

86 Pazhani, G.P. *et al.* Clonal multidrug-resistant *Shigella dysenteriae* type 1 strains associated with epidemic and sporadic dysenteries in eastern India. *Antimicrob. Agents Chemother*. **48,** 681-684 (2004).

87 Le Hello, S. *et al*. International spread of an epidemic population of *Salmonella enterica* serotype Kentucky ST198 resistant to ciprofloxacin. *J. Infect. Dis.* **204,** 675-684. (2011).