Supplementary Information

Genomic perspective on the bacillus causing paratyphoid B fever

Jane Hawkey, Lise Frézal, Alicia Tran Dien, Anna Zhukova, Derek Brown, Marie Anne Chattaway, Sandra Simon, Hidemasa Izumiya, Patricia I. Fields, Niall de Lappe, Lidia Kaftyreva, Xuebin Xu, Junko Isobe, Dominique Clermont, Elisabeth Njamkepo, Yukihiro Akeda, Sylvie Issenhuth-Jeanjean, Mariia Makarova, Yanan Wang, Martin Hunt, Brent M. Jenkins, Magali Ravel, Véronique Guibert, Estelle Serre, Zoya Matveeva, Laëtitia Fabre, Martin Cormican, Min Yue, Baoli Zhu, Masatomo Morita, Zamin Iqbal, Carolina Silva Nodari, Maria Pardos de la Gandara, François-Xavier Weill

Supplementary Methods

Short-read sequencing

At NHSGGC, genomic DNA was extracted with the QIAsymphony system (Qiagen, Hilden, Germany), the libraries were prepared with the Illumina DNA Prep Kit (Illumina) and sequencing was performed with MiSeq (Illumina). At UKHSA, genomic DNA was extracted with the QIAsymphony system (Qiagen), the libraries were prepared with the Nextera XT kit (Illumina) and sequencing was performed with HiSeq 2500 (Illumina). At NIID, genomic DNA was extracted with the QIAseq FX DNA Library Kit (Qiagen), the libraries were prepared with the QIAseq FX DNA Library Kit (Qiagen), and sequencing was performed with MiSeq (Illumina). At UCD, genomic DNA was extracted with the EZ1® DNA Tissue kit

(Qiagen), the libraries were prepared with the Nextera DNA Flex library prep kit (Illumina), and sequencing was performed with MiSeq (Illumina). At RKI, genomic DNA was extracted with the GenElute™ Bacterial Genomic DNA Kit (Sigma-Aldrich, St. Louis, MO, USA) or by combined thermal and mechanical disruption with acid-washed glass beads (Sigma-Aldrich) in a TissueLyser II bead mill (Qiagen), the libraries were prepared with the Nextera XT kit (Illumina), and sequencing was performed with MiSeq (Illumina) or NextSeq 500 (Illumina).

Supplementary Text

Epidemiology of paratyphoid B fever during the first half of the 20th century

From its discovery in France in 1896 until 1903, 69 cases of paratyphoid B fever (PTB) — mostly sporadic or small-scale outbreaks — were reported in various European countries and the United States of America (US)[1,2]. In the United Kingdom (UK), in 1906, it was estimated that 3% of the 3,000 typhoid fever cases notified each year in London could be PTB cases; in Germany in 1901 and 1905, this percentage was around 7% (ref.[3]). Whereas, in the US, no SPB was isolated among 250 cases of enteric fever (200 of typhoid fever and 50 of paratyphoid A) in Philadelphia in 1908-1909 (ref.[2]). Between December 1914 and February 1915, 6.7 % of the ~4,500 enteric fever cases seen at the Military Hospital of Zuydcoote, France were PTB cases[4]. A new combined vaccine (TAB) — extending the initial vaccination against typhoid fever to paratyphoid A and B fevers — was therefore introduced in 1915-1916 for Allied forces[5]. At the end of the 1930s, PTB became much more prevalent than typhoid fever in England[6]. Between 1923 and 1941, the vast majority of the 40 outbreaks of PTB, involving a total of more than 4,200 cases (from 4 to 883 cases per outbreak), reported in the UK were due to food contaminated by transient or chronic carriers[7]. The foods

contaminated were natural or synthetic cream, unpasteurised milk, ice-cream, bakery products or confectionery[7]. One contaminated synthetic cream, in particular, was implicated in seven outbreaks between 1940 (when the sale of natural cream was forbidden**)** to 1941 and resulted in 1,462 cases and at least 10 deaths. After World War II, commercial bakeries in the UK were through via a new source: contaminated frozen whole egg imported from China[8,9]. Contaminated water, either alone or on edible plants, was less frequently implicated than food items[7,10]. Interestingly, whereas typhoid fever is restricted to humans, two reports from Scandinavia (Norway in 1937 and Sweden in 1938) mentioned dogs as sources of SPB$^-$ infections in humans[7].

Validation of the SPB$^-$ PG1 diversity dataset

Our serotype prediction approach identified one genome among the 568 genomes of the diversity dataset that did not correspond to SPB (the 116K strain, see below). Twenty-one genomes were identified as monophasic SPB (without the *fljB* gene encoding the H2 antigen "1,2"), and it was not possible to predict the O antigen for five genomes due to a low read coverage in the corresponding region (these five genomes had the correct *fliC* and *fljB* genes, encoding, the "b" and "1,2" antigens, respectively, and had been phenotypically serotyped as SPB at Institut Pasteur) (**Supplementary Data 1**). The rest of the genomes were inferred to be SPB. We then ensured that these 568 genomes belonged to PG1, the invasive lineage of SPB, described by Connor and coworkers[11]. We therefore used the EnteroBase core-genome MLST (cgMLST) scheme — based on 3,002 core genes — that had been successfully used to study the population structure of *Salmonella enterica*[12,13] to confirm that our 568 genomes belonged to invasive PG1, by establishing a link between cgMLST and PG data. After curation of the genomes and metadata described by Connor and coworkers[11] (see Methods section "Genomic typing methods") (**Supplementary Data 7**), a tree based on the cgMLST

allelic distance for these genomes made it possible to recognise the 10 known PGs of SPB (**Supplementary Fig. 1a**). Following the hierarchical clustering[14] of cgMLST data, also implemented in EnteroBase, all isolates assigned to PG1-PG5 clustered in HC2000_155 and HC900_155. However, the PG1 isolates could be distinguished from PG2 to PG5 isolates at the HC400 level (i.e., grouping together genomes with no more than 400 allelic differences). All the invasive PG1 genomes, and only these genomes, belonged to HC400_1620 (**Supplementary Fig. 1b**). Furthermore, only the HC400_1620 genomes contained the specific SNV within STM 3356 described in SPB$^-$ strains (ref.[15]). All 446 genomes from SPB$^-$ isolates and strains contributed by various reference laboratories across the world for this study, and the 109 previously published genomes belonged to HC400_1620 and contained the *d*-Tar$^-$ specific SNV (**Supplementary Data 1**).

We also used a combination of this HC400_1620 level and the presence of the *d*-Tar$^-$ specific SNV to search for additional unpublished SPB$^-$ genomes in EnteroBase, a very large genomic database containing >400,000 *Salmonella* genomes at the time of study. This search captured 12 additional genomes from reference strains (e.g., SARA collection)[16] or from isolates collected locally in regions of the world not well covered by our initial dataset (**Supplementary Data 1**). During this search, we unexpectedly found within HC400_1620, a reference strain (116K) of an extremely rare serotype, Onarimon (antigenic formula: 1,9,12:b:1,2), deposited independently by one of the participating laboratories (Institut Pasteur). Serotype prediction based on genomic sequence confirmed this serotype and the *d*-Tar$^-$ specific SNV was present. There were only 28 serotype Onarimon strains reported in 1965 (among the 547,386 strains from diverse sources across the world)[17] and this serotype has been reported to cause paratyphoid fevers[18]. As in *Salmonella* spp., serotype antigens can

100 be subject to horizontal gene transfer and homologous recombination[19], we therefore
101 considered 116K to be an O antigen-variant of SPB$^-$ and we included it in the study.
102
103 For one unpublished genome (ATCC 10719, original name 41-H-6) we were unable to
104 confirm membership of HC400_1620 because it was a draft genome prepared from 454
105 sequences and was not, therefore, accepted by EnteroBase (**Supplementary Data 1**).
106 However, this genome came from an old SPB strain used to prepare the TAB vaccine of the
107 US Army in 1940 (ref.[20]) and it contained the *d*-Tar$^-$ specific SNV. We therefore retained this
108 genome in the study.
109
110 Using MLST7, Achtman and coworkers[19] found that only SPB isolates containing the *d*-Tar$^-$-
111 specific SNV belonged to ST86 or five single-locus variants (SLVs) of ST86. We used this
112 same scheme implemented in EnteroBase to analyse the 568 genomes of the diversity dataset.
113 The vast majority of genomes (96.7%, 549/568) belonged to ST86 (**Supplementary Data 1**).
114 However, 17 of these genomes belonged to 13 different SLVs of ST86. These 17 isolates
115 included the three lineage L1 genomes, all three belonging to ST5113. One genome (13-80)
116 could not be typed by MLST due to an incomplete *purE* gene. Finally, one genome
117 (ERR129867) described by Connor and coworkers[11] was a triple-locus variant of ST86
118 (ST2134). We therefore confirm that even though ST86 is an excellent predictor of SPB$^-$ PG1
119 isolates, the existence in such isolates of other SLVs of ST86 (such as ST772, ST2340,
120 ST5113, ST6558, ST7999, ST8505, ST8506, ST8931, ST8932, ST9224, ST10005, ST10013,
121 and ST10014 and potentially others), a triple-locus variant of ST86 (ST2134), or even the
122 possibility of non-typability by MLST (**Supplementary Data 1**) might complicate the use of
123 MLST7 as a unique tool for identifying SPB$^-$ PG1. Furthermore, the SLVs of ST86 are also
124 observed in other SPB PGs, such as ST43 in PG3 and PG4, and ST149 in PG5 (ref.[11]). If

MLST7 is used, we recommend its use in combination with the detection of the specific *d*-Tar$^-$ SNV within STM 3356, this SNV being crucial for the formal assignment of non-ST86 genomes to SPB$^-$ PG1.

Genotypes found in non-human SPB$^-$ PG1 isolates

Twelve different genotypes were observed for the 14 animal isolates (from six molluscs, two insects, two dogs, one squirrel, one pig, one bird and one crustacean) and 14 genotypes were observed for the 42 environmental isolates (mostly from river water). These genotypes were identical to those found in human isolates collected in the same geographic region. The first dog isolate (#1190) was obtained from a dog reported sick (diarrhoea and spontaneous abortion) in the week before the onset of three human cases in a Swedish village in 1938 (refs.[21,22]). This isolate belonged to genotype 4. The second dog isolate was collected in Algeria (North Africa) in 1966 and belonged to genotype 7.3.1_NorthAfrica1. The only two food isolates studied were isolated in Iraq in 1976 from locally prepared food[23] and were of genotype 9.0, which was mostly isolated in Western Europe. The other four Iraqi isolates, obtained from humans between 1974 and 1980, belonged to three other genotypes (10.3.1_SouthAsia1, 10.3.2_MiddleEast1, and 10.3.8.4_MiddleEast4).

Age of SPB$^-$ PG1

Based on our dataset of 568 SPB$^-$ PG1 genomes, we estimated the age of this pathogen at ~750 years (1274 CE; 95% CI, 915 – 1583), which is very close to the previous median date of origin estimated by Connor and coworkers[11] (1188 CE; 95% CI, 469 BC – 1799 CE), who used only 25 SPB$^-$ PG1 genomes (i.e., those with a known year of isolation). SPB$^-$ is older than SPA, which is estimated to have originated 450 – 700 years ago[24]. SPA was discovered two years after SPB (in the USA in 1898)[1,2,25] but is currently the most frequent agent of

150 paratyphoid fever[26]. Due to lineage extinction, in particular, times to the MRCA are often
151 underestimated and the inclusion of ancient DNA in the analysis would increase precision and
152 make it possible to establish dates of origin further in the past[24]. The dating of a representative
153 collection of modern isolates of SPC estimated the origin of SPC to date back 456 – 664
154 years. When a draft SPC genome from an 800-year-old Norwegian skeleton was added to the
155 analysis, the time to the MRCA increased to 1162 – 1526 years[2]. Unfortunately, no ancient
156 DNA is currently available for SPB$^-$ strains.
157
158 **Prophages of SPB$^-$ PG1**
159 We facilitated the pan-genome analysis, including the assignment of accessory genes to
160 clearly delineated prophages (**Supplementary Data 8**), by also including in our analysis the
161 complete genomes of 14 isolates from the "diversity dataset" together with 12 (including the
162 reference genome, CIP 54.115) genomes generated for this study and two publicly available
163 genomes (P7704 and SARA41_FB_1) (**Supplementary Data 1**).
164
165 Of the 1,506 accessory genes present in < 95% of the genomes (**Supplementary Fig. 7,**
166 **Supplementary Data 4**), 696 (46.2%), 242 (16.1%), and 28 (1.9%) were found to belong to
167 prophages, plasmids, and transposases, respectively. The mapping of these accessory genes
168 onto the core genome phylogeny revealed that an absence of phylogenetic clustering for
169 plasmid genes, whereas some prophage genes were (sub)lineage-specific **(Supplementary**
170 **Fig. 8).** We then evaluated the occupancy of the 10 prophage insertion sites — identified in
171 the 14 complete genomes — with short-read assemblies from the entire diversity dataset
172 (**Supplementary Fig. 10**).
173

Three (sites #1 to 3) of the 10 sites were occupied by prophages containing the *sopE* virulence gene. Only two of the 568 SPB$^-$ PG1 genomes (173-73 and 73-67) did not contain *sopE*. Two types of *sopE* prophages, both belonging to the Caudoviricetes class, were identified; the first was 41,250 bp to 44,615 bp in size and displayed 89.8% to 95.9% nucleotide identity (38% to 55% coverage) to the *Salmonella* Brunovirus SEN34 (GenBank accession no. NC_028699); the second was 34,723 bp long and displayed ~98% nucleotide identity (83% coverage) to the *Enterobacterium* Xuanwuvirus P88 (GenBank accession no. NC_026014) (**Supplementary Table 2**).

The *sopE*-containing SEN34-like prophages occupied sites #1 and #2, whereas the P88-like prophage occupied site #3. This *sopE*-prophage content was correlated with SPB$^-$ phylogeny, with a *sopE*-containing SEN34-like prophage inserted at site #1 in lineages L1 to L4 and at site #2 in lineages L5 to L11 **(Fig. 6b)**. The P88-like prophage was not seen alone but always in addition to the SEN34-like prophage and its presence was strongly associated with the Dundee PT (Chi squared test, $p < 0.0001$) (**Supplementary Fig. 10**).

The presence of a particular *sopE* bacteriophage (ΦSopE309) was previously used by Prager and coworkers[28] to distinguish the systemic SPB$^-$ isolates from the enteric SPB$^+$ isolates. By contrast, Connor and coworkers[11] suggested that the *sopE* gene was not a suitable marker for identifying SPB$^-$ PG1 isolates because this gene was not present in all of their PG1 isolates and some of these isolates had a gene homologous to *sopE* also found in all other SPB PGs. As ΦSopE309 was initially isolated from SPB$^-$ strain B309 — also included in our study — we were able to determine that it was actually the Brunovirus SEN34-like prophage. Our in-depth analysis supports the recommendation of Prager and coworkers[28] to use *sopE* as a marker gene for SPB$^-$, because the *sopE*-carrying SEN34-like prophage was present in almost

all the 568 SPB$^-$ PG1 isolates from our diversity dataset (including 28 genomes from the study by Connor and coworkers[11]). Only two isolates (173-73 and 73-67) were devoid of *sopE* and each of the three insertion sites for *sopE* prophages were empty in these isolates. These two isolates were collected more than 50 years ago and we cannot therefore rule out the possibility that the *sopE* prophage was excised during storage or subculture. Furthermore, according to their nearest neighbours on the phylogenetic tree, these isolates would have contained only one *sopE* prophage. The gene homologous to *sopE* unexpectedly found in some SPB$^-$ PG1 genomes by Connor and coworkers[11] is almost certainly *sopE2*, a chromosomal gene present in all S*almonella* lineages encoding an effector protein, SopE2, 69% identical to SopE (ref.[29]). By contrast, the *sopE* gene, carried by different bacteriophages (lambda-like Gifsy-2, P2-like, mTmV), was previously found in only some serotypes of *Salmonella* (e.g., Gallinarum, Typhi, Dublin, Heidelberg) or some strains of certain serotypes (e.g., epidemic *S. enterica* serotype Typhimurium DT204 in the 1970s or the currently dominant monophasic *S. enterica* serotype Typhimurium ST34)[30,31,32]. It has been suggested that the acquisition of *sopE* by lysogenic conversion increases the fitness of *S . enterica* serotype Typhimurium, thereby contributing to the emergence of epidemic strains[33]. SopE and SopE2 are G-nucleotide exchange factors that are translocated into the host cell, where they activate host cellular Rho-GTPases — such as Cdc42 and Rac1 for SopE and Cdc42 alone for SopE2 — which act as key regulators of diverse activities, such as the expression of pro-inflammatory cytokines and the organisation of the actin cytoskeleton, ultimately promoting bacterial uptake and even intracellular replication[33,34]. In addition to their chromosomal *sopE2* gene and their SEN-34-like prophage *sopE* gene, 17.8% and 7.1% of SPB$^-$ PG1 isolates contain one and two additional copies of *sopE* (carried by either the SEN-34-like or P88-like prophages), respectively. Three copies of *sopE* were observed exclusively in the PT Dundee strain (genotype 9.1_France) that was epidemic in France after WWII and is still isolated, even now,

224 from long-term carriers. Similarly, two copies of *sopE* were previously described in
225 *S. enterica* serotype Heidelberg isolates from a large multistate outbreak linked to turkey meat
226 in the US[35] and three copies were found in a cluster of monophasic *S. enterica* serotype
227 Typhimurium ST34 isolates from humans and pigs in the UK[32]. An increase in *sopE* copy
228 number is observed in ~25% of SPB$^-$ PG1 isolates, but additional studies are required to
229 determine the consequences of this increase for the expression of *sopE* and interaction with
230 host cells.
231
232 At least three sorts of non-*sopE* prophages could occupy insertion site "A". This site is
233 located right next to the *hin-fljB-fljA* region, which is involved in the expression of the phase
234 2 flagellin ("1,2" antigen encoded by *fljB*) (**Supplementary Fig. 11**). Prophage rearrangement
235 at site "A" leading to the loss of the of *hin-fljB-fljA* region might account for the appearance
236 of monophasic isolates (i.e., lacking phase 2, with the antigenic formula 4:b:-) in genotypes 4
237 and 7.3 (**Supplementary Data 1**).
238
239 **Development of a new SNV-based genotyping tool and comparison to cgMLST**
240 Mykrobe checks (i) for the presence of *invA* to ensure that the genome belongs to the genus
241 *Salmonella*, (ii) and then for the presence of the STM 3356 *d*-Tar$^-$- specific SNV to confirm
242 that the genome belongs to SPB$^-$, and (iii) finally assigns genomes to genotypes based on
243 presence of the 38 genotype-specific SNVs. For validation of our scheme, we first analysed
244 the 568 genomes of our diversity dataset. A concordance of 100% was obtained between the
245 genotypes assigned by Mykrobe and those initially defined on the basis of both hierBAPS and
246 visual inspection (**Fig.2a,b**). The scheme was then used on the surveillance dataset,
247 containing 336 routinely obtained genomes (111 already present in the diversity dataset and
248 225 new genomes) originating from public health laboratories in four countries (UK, $n = 200$;

France, *n* = 84; USA, *n* = 39; Canada, *n* = 13), with isolation dates between 2014 and 2023 (see Methods section "*S. enterica* serotype Paratyphi B sequence data collection") (**Supplementary Data 6)**. The SNV-based genotyping scheme accurately captured the population structure as defined by a core-genome phylogeny on the 793 genomes from both datasets (**Supplementary Fig. 8).** Three new routinely sequenced genomes, without travel information, from the US and Canada were assigned to genotype 10.3.8. However, on the basis of phylogeny, one (PNCS011535) of these genomes was considered to be intermediate between genomes typed as 10.3.8 and those typed as 10.3.8.4_MiddleEast4, and the other two (PNUSAS023302 and PNUSAS023173) were grouped together and considered intermediate between genomes typed as 10.3.8 and those typed as 10.3.8.1_SouthAsia2. If similar isolates were to be identified in the future, we would perhaps have to refine the definition of genotype 10.3.8 slightly. Only one of the 793 genomes genotyped (44-66) was not called by Mykrobe due to a missing *invA* gene. Fourteen other genomes were called correctly by Mykrobe despite read coverage being too low for the *d*-Tar SNV region, precluding formal identification of the *d*-Tar$^-$ SNV. The genotyping tool yields the result "unknown" if the *Salmonella* specific *invA* gene is not detected (for non-*Salmonella* genomes) or if the *d*-Tar$^+$ SNV is detected (for non-PG1 *Salmonella* genomes). However, if there is no call for the *d*-Tar$^-$ SNV marker, this does not prevent a final genotype call being obtained. Therefore, to avoid the possibility of a non-PG1 *Salmonella* genome with a low read coverage for the *d*-Tar SNV region (shown in the Mykrobe output table under the column "species_depth") being erroneously assigned to a PG1 genotype, it is therefore recommended to ensure that all genomes genotyped with this scheme belong to cgMLST HC400_1620, a robust signature of SPB$^-$ PG1.

With a higher-resolution HC level, such as HC200, it was possible to identify lineages L1 to L4 (L1, HC200_137805; L2, HC200_17706; L3, HC200_301037; L4, HC200_12575)

(**Supplementary Fig. 12, Supplementary Data 1**). However, all other isolates belonging to Lineages L5 to L11 were assigned to the same HC200 cluster (HC200_1620), with the exception of three isolates previously found to be outliers in the root-to-tip analysis, which were each assigned to a unique HC200 cluster. The use of HC100 to HC50 did not permit the recognition of SPB$^-$ population structure as determined by our core-genome SNV-based phylogenetic analysis and our SNV-based genotyping scheme. For example, the isolates belonging to the emerging South American genotype 10.3.6 were found in three different HC50 clusters. The predominant cluster, HC50_1857, was even not specific to 10.3.6 (also found in isolates of five other genotypes) and the 10.3.6 isolates within HC50_1857, were further subdivided into 15 different HC20 clusters, making them hard to track by cgMLST (**Supplementary Fig. 12, Supplementary Fig. 13, Supplementary Data 1**). The superiority of SNV-based genotyping over cgMLST for tracking epidemiologically relevant strains has already been reported for *S. sonnei*[36].

Supplementary Table 1. GenBank accession numbers and co-ordinates of the genes studied to confirm that our isolates were SPB$^-$

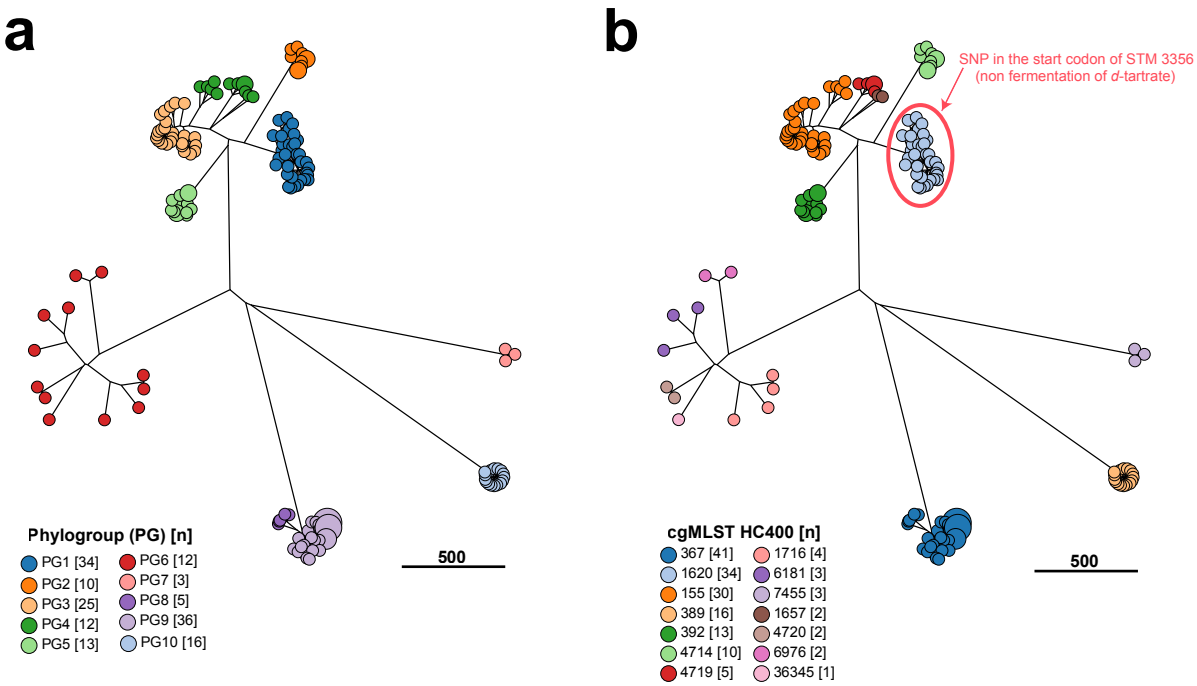
Target	Strain	Accession no.	Coordinates
rfb_O4	*S. enterica* serotype Typhimurium str. LT2	NC_000913.3	2160595-2182675
rfb_O9	*S. enterica* serotype Enteritidis str. P125109	AM933172.1	2162790-2184501
fliC_b	*S. enterica* serotype Paratyphi B str. B62	CP147902	1931292-1932779
fljB_1,2	*S. enterica* serotype Paratyphi B str. B62	CP147902	1120688-1122208
***d*-Tar$^+$**	*S. enterica* serotype Paratyphi B str. NCTC 5706	AY211490.1	1-291*
***d*-Tar$^-$**	*S. enterica* serotype Paratyphi B str. NCTC 3176	AY211491.1	1-291*

*single-nucleotide variant (SNV) at position 252 (gene STM 3356): G (*d*-Tar$^+$) or A (*d*-Tar$^-$)

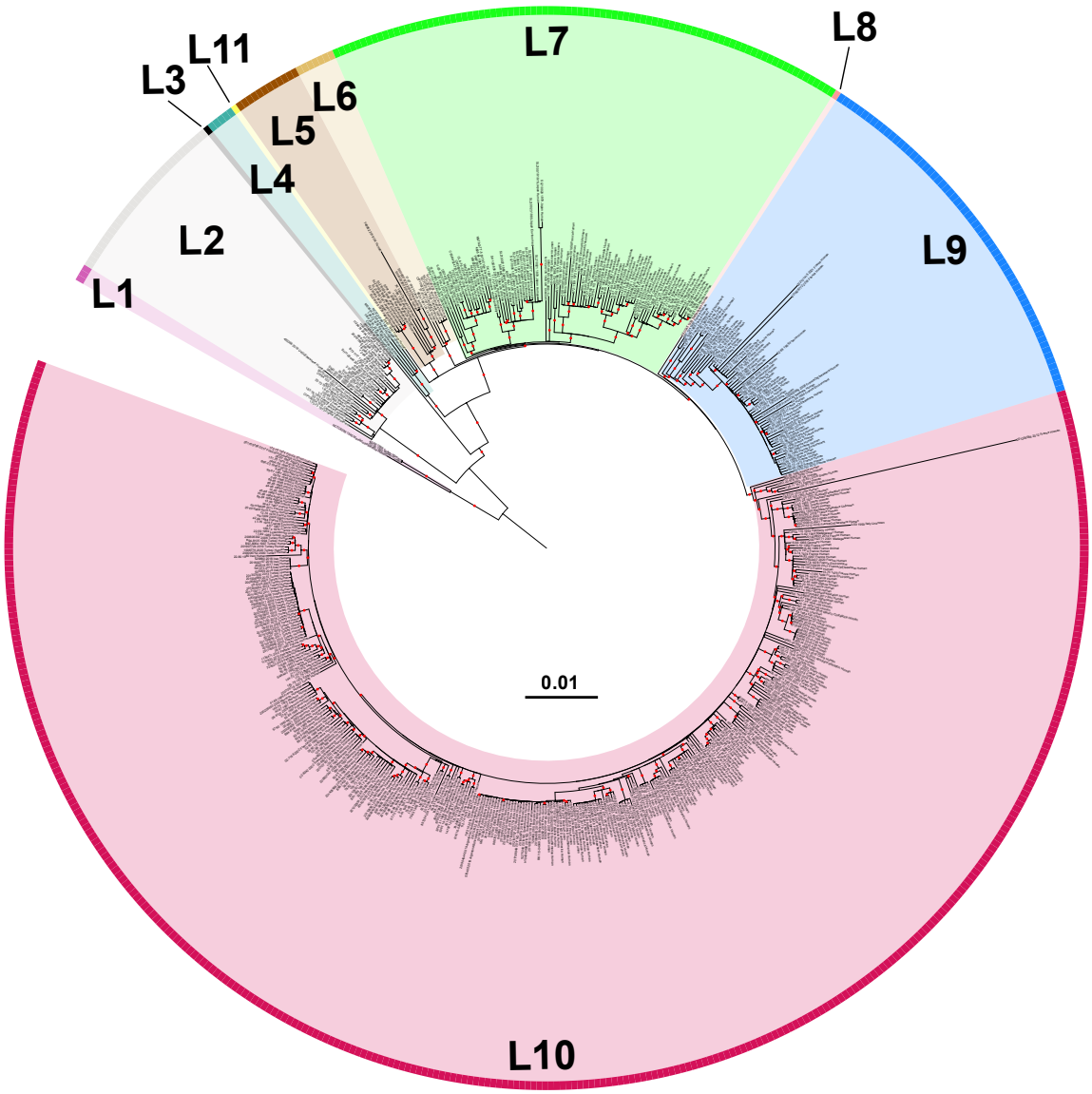
Supplementary Table 2. Comparison of *sopE* prophages found in the 14 complete SPB$^-$ PG1 genomes

Isolate	Genotype	Prophage insertion site	Prophage length	Prophage genus	Blastn query *	Query name	Query coverage (%)	Nucleotide identity (%)	#hit_genes from PHASTER
63-90	1	1 [SopE]	43665	*Brunovirus*	NC_028699.1	*Salmonella* phage SEN34	53	90.9	36
B76	2.1	1 [SopE]	43084	*Brunovirus*	NC_028699.1	*Salmonella* phage SEN34	45	89.8	30
CIP 54.100	2.1	1 [SopE]	43084	*Brunovirus*	NC_028699.1	*Salmonella* phage SEN34	45	89.8	30
CIP A214	4	1 [SopE]	43096	*Brunovirus*	NC_028699.1	*Salmonella* phage SEN34	45	89.8	30
B2590	9.1_France	1 [SopE]	43987	*Brunovirus*	NC_028699.1	*Salmonella* phage SEN34	44	95.9	33
B2590	9.1_France	2 [SopE]	41250	*Brunovirus*	NC_028699.1	*Salmonella* phage SEN34	50	93.7	29
SARA41	7.3.1_NorthAfrica1	2 [SopE]	43291	*Brunovirus*	NC_028699.1	*Salmonella* phage SEN34	55	91.9	37
B2227	7.3.2_BAOR	2 [SopE]	44615	*Brunovirus*	NC_028699.1	*Salmonella* phage SEN34	55	91.9	39
B624	6	2 [SopE]	41336	*Brunovirus*	NC_028699.1	*Salmonella* phage SEN34	36	95.9	26
P7704	10.3.6_SouthAmerica	2 [SopE]	43943	*Brunovirus*	NC_028699.1	*Salmonella* phage SEN34	38	95.9	31
B1655	10.3	2 [SopE]	43943	*Brunovirus*	NC_028699.1	*Salmonella* phage SEN34	38	95.9	31
B1727	7.3	2 [SopE]	43943	*Brunovirus*	NC_028699.1	*Salmonella* phage SEN34	38	95.9	31
B97	7.2_EuropeEasternAsia	2 [SopE]	43943	*Brunovirus*	NC_028699.1	*Salmonella* phage SEN34	38	95.9	31
B62	7.2_EuropeEasternAsia	2 [SopE]	43941	*Brunovirus*	NC_028699.1	*Salmonella* phage SEN34	38	95.9	31
CIP 54.115	7.3	2 [SopE]	43941	*Brunovirus*	NC_028699.1	*Salmonella* phage SEN34	38	95.9	30
B2227	7.3.2_BAOR	3 [SopE]	34723	*Xuanwuvirus*	NC_026014.1	Enterobacteria phage P88	83	97.9	44
B2590	9.1_France	3 [SopE]	34723	*Xuanwuvirus*	NC_026014.1	Enterobacteria phage P88	83	97.8	44
B624	6	3 [SopE]	34724	*Xuanwuvirus*	NC_026014.1	Enterobacteria phage P88	83	97.8	44

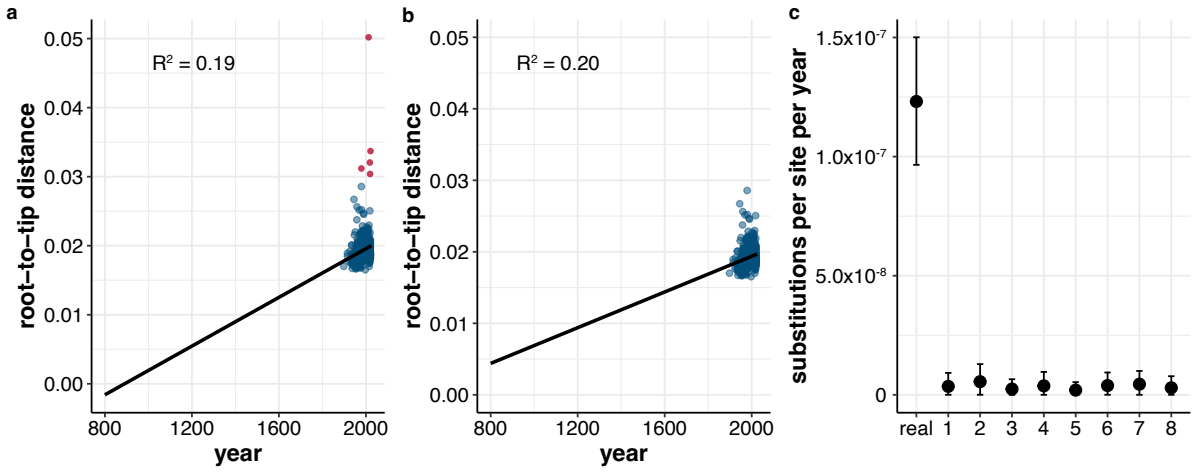
* best hit identified by PHASTER



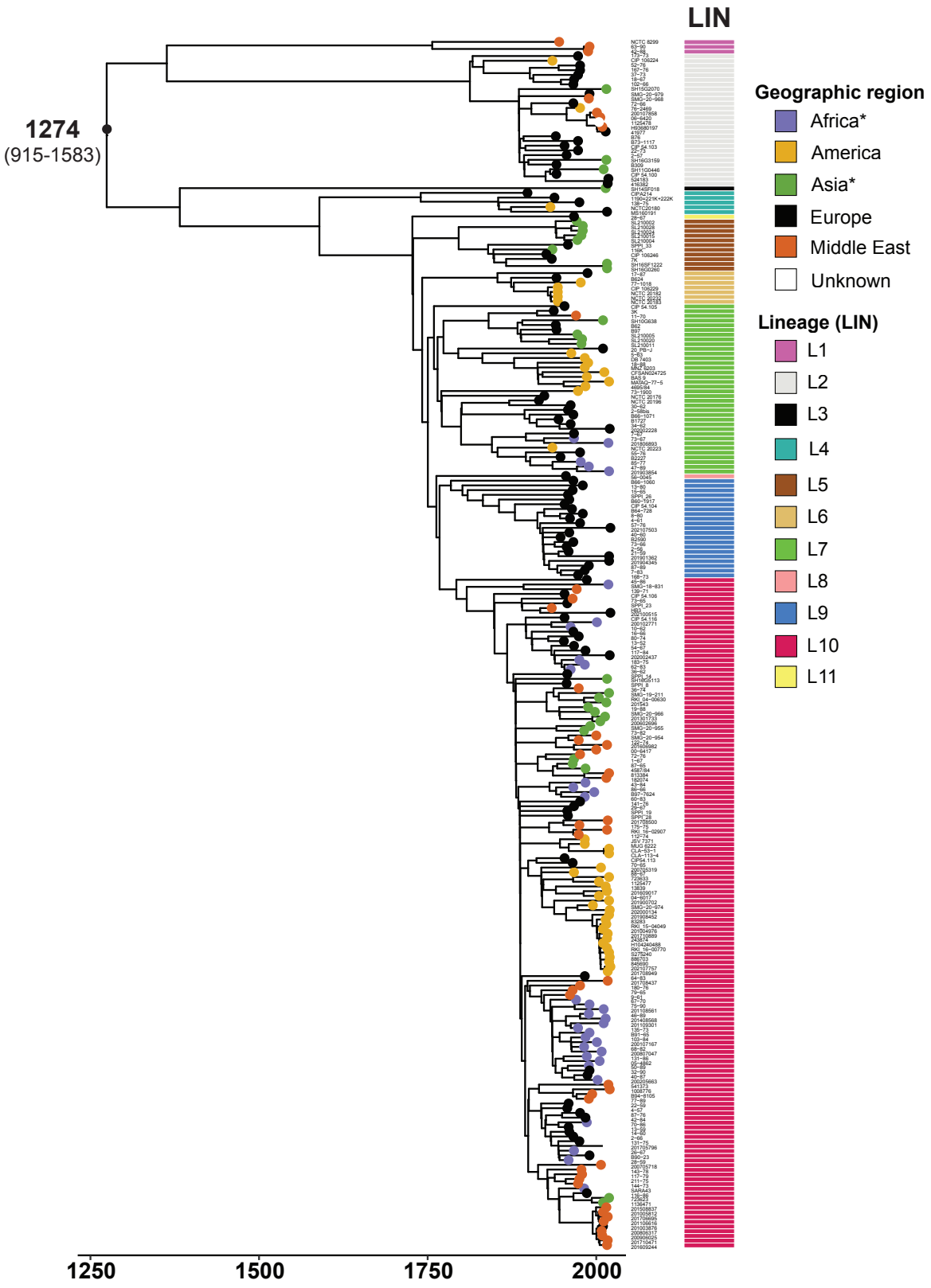
Supplementary Figure 1. A NINJA neighbour-joining GrapeTree of 166 SPB genomes described by Connor *et al.*[11]. a, The tree nodes are colour-coded by phylogroup (PG) (see the legend, inset). **b,** The tree nodes are colour-coded by cgMLST HC400 data (see the legend, inset). The presence of the specific SNV found in the STM 3356 gene of SPB$^-$ (ref.[15]) is indicated. The scale bars indicate the number of cgMLST allelic differences. The interactive version of the tree is publicly available from https://enterobase.warwick.ac.uk/ms_tree?tree_id=92077



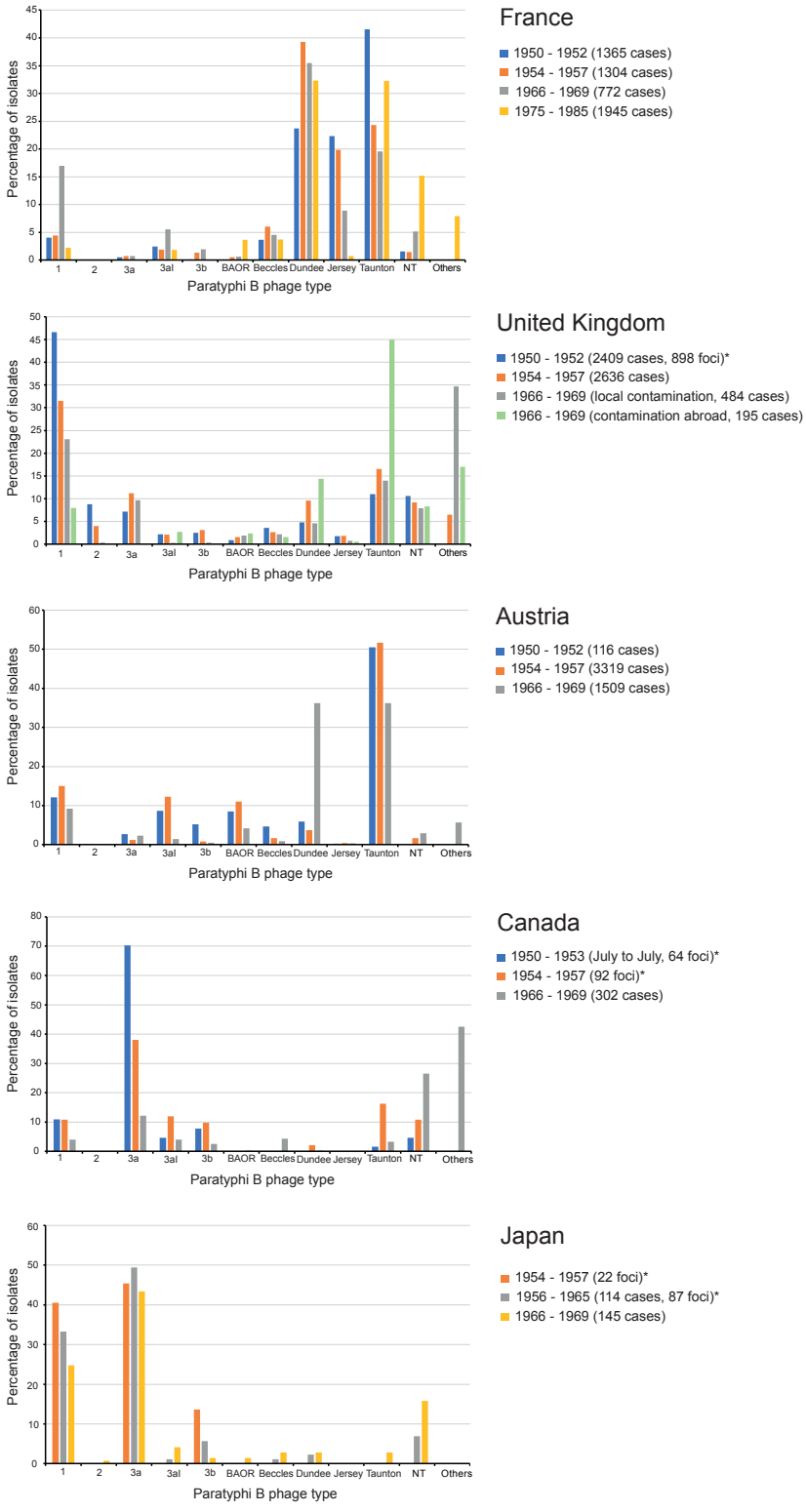
Supplementary Figure 2. Circular maximum likelihood phylogeny of the 568 SPB$^-$ PG1 genomes of the diversity dataset. Same phylogeny as in Fig. 1a, except that for each isolate, its name, year of collection, country of origin, and source, are shown at the tips of the tree. The lineages are also shown. Red dots indicate bootstrap values ≥95%.



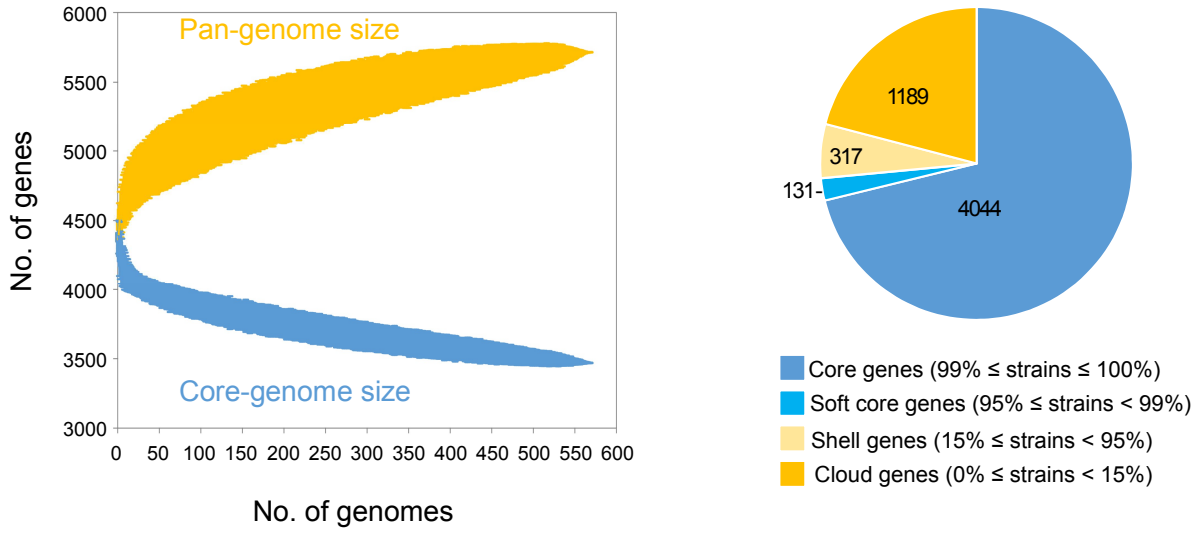
Supplementary Figure 3. Temporal structure of SPB⁻ PG1 genomes. a, Correlation of root-to-tip distances with year for all 568 genomes of the diversity dataset in the maximum likelihood phylogeny. Red points indicate outlier genomes, which were long branches with a distance of >0.03, and blue points represent all other genomes. R2 shows the Pearson correlation coefficient. **b,** Correlation of root-to-tip distances with year after excluding outlier genomes shown in panel "a". **c,** Date randomisation test for the dated BEAST2 phylogeny. The first point indicates the median rate (in substitutions site^{-1} year^{-1}) estimated by BEAST2 for the real dates, with bars showing the 95% height posterior density (HPD) interval. Subsequent points show the rate estimates from eight independent BEAST2 runs in which dates were randomised across the phylogeny.



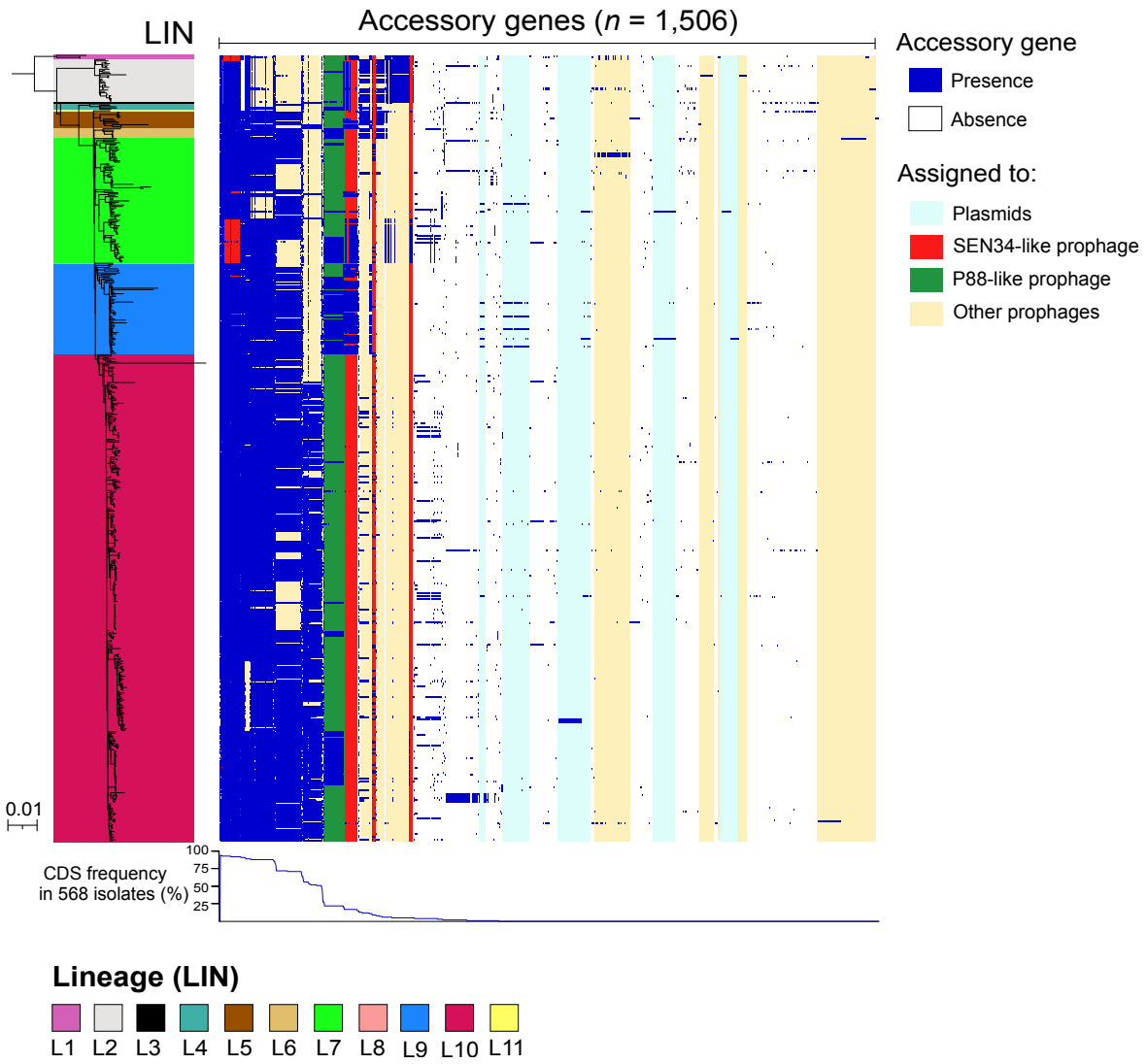
Supplementary Figure 4. Timed phylogeny of a representative subsample of 256 SPB^{-} PG1 genomes. Same phylogeny as in Fig. 2a, except that the name of the isolates is indicated.



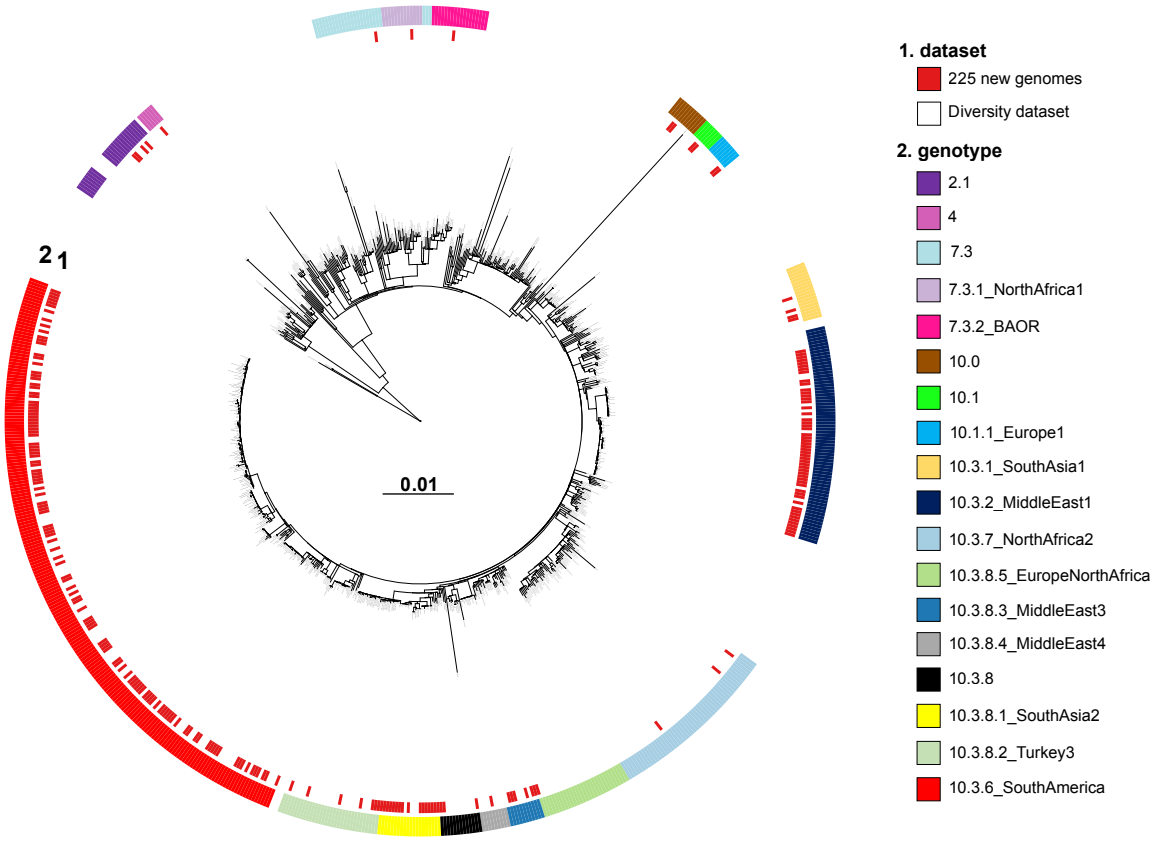
Supplementary Figure 5. Selected phage-typing data reported for SPB isolates from France, UK, Austria, Canada, and Japan. The original data can be found in refs[37-41].



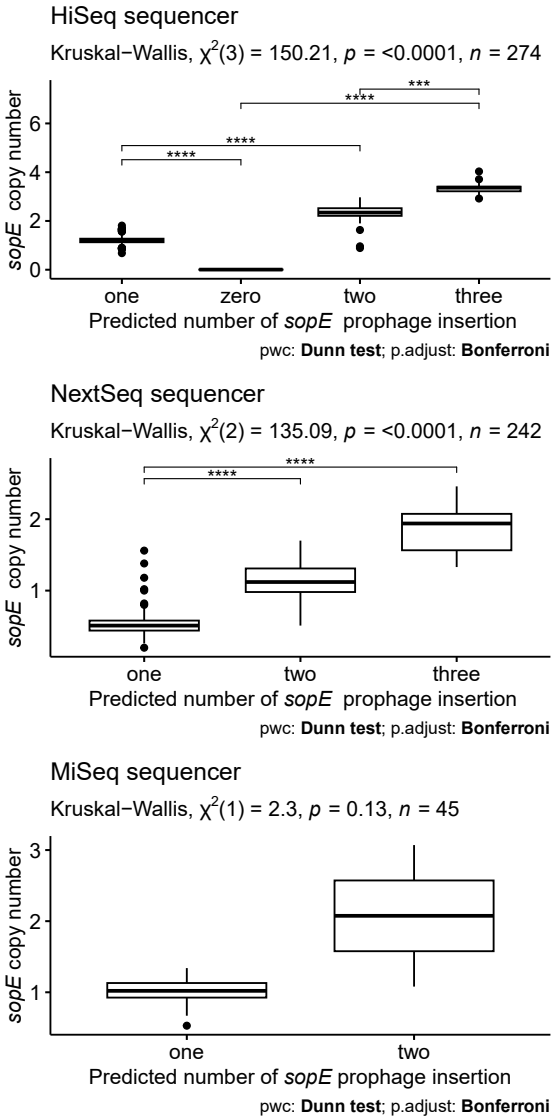
Supplementary Figure 6. Characteristics of the core- and pan-genomes of SPB$^-$ PG1. The pan-genome analysis was performed across the 11 SPB$^-$ PG1 lineages, with the assemblies of the 568 isolates of the diversity dataset. In the left panel, the pan-genome curve (dark yellow) shows the number of genes subsequently discovered as more genomes are added to the dataset. The rarefaction curve (blue) shows the decay in the number of core genes as more genomes are added to the dataset. Both pan-genome and core-genome curves were estimated from the panaroo binary matrix with PanGP[42] using a totally random sampling method and 10 sample repeats. The pie chart (right panel) shows the relative proportions of the core (dark blue), soft core (blue), shell (light yellow) and cloud (dark yellow) genes.



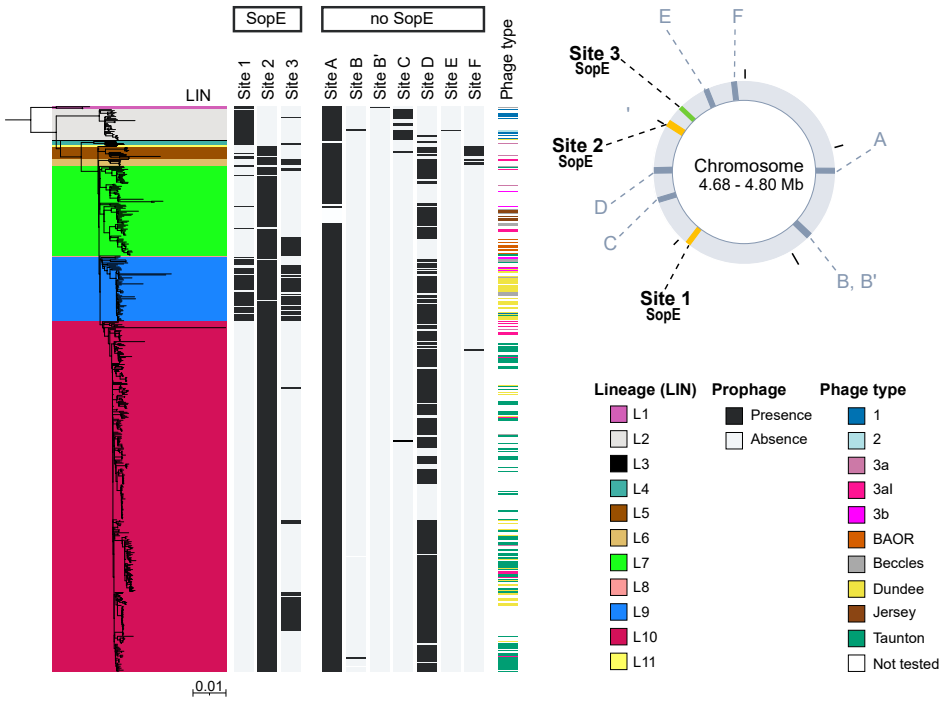
Supplementary Figure 7. Distribution of the 1,506 accessory genes across the phylogeny of SPB⁻ PG1 and their attribution to prophages or plasmids. The pan-genome analysis was performed with the assemblies of the 568 isolates from the diversity dataset. The phylogeny is similar to that shown in Fig. 2a.



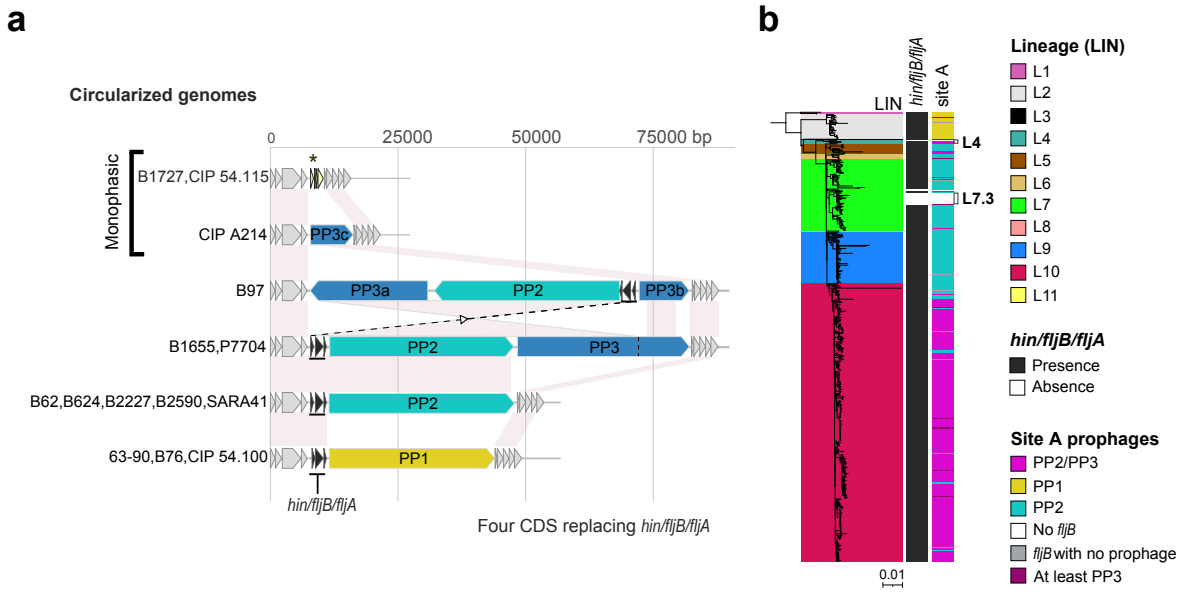
Supplementary Figure 8. Maximum likelihood phylogeny for 793 SPB$^-$ PG1 isolates. This phylogeny includes the 568 isolates of the diversity dataset (not coloured in ring 1) and 225 additional recent isolates (coloured in red in ring 1). For each isolate, its name, year of collection, country of origin, and source, are shown at the tips of the tree. The scale bar indicates the number of substitutions per variable site (SNV). The genotypes (see legend) of these additional 225 isolates are colour-coded in ring 2.



Supplementary Figure 9. Identification of the discrepancies between the predictions of *sopE* copy number and insertion site occupancy at sites #1, #2, and #3. The *sopE* copy number per genome was estimated from short-read mapping onto the B62 genome across the 11 SPB$^-$ PG1 lineages, for the 568 isolates of the diversity dataset. The distribution of *sopE* copy number values according to the number of *sopE* prophage insertion sites occupied can be used to detect potential discrepancies between site occupancy and *sopE* copy number predictions.

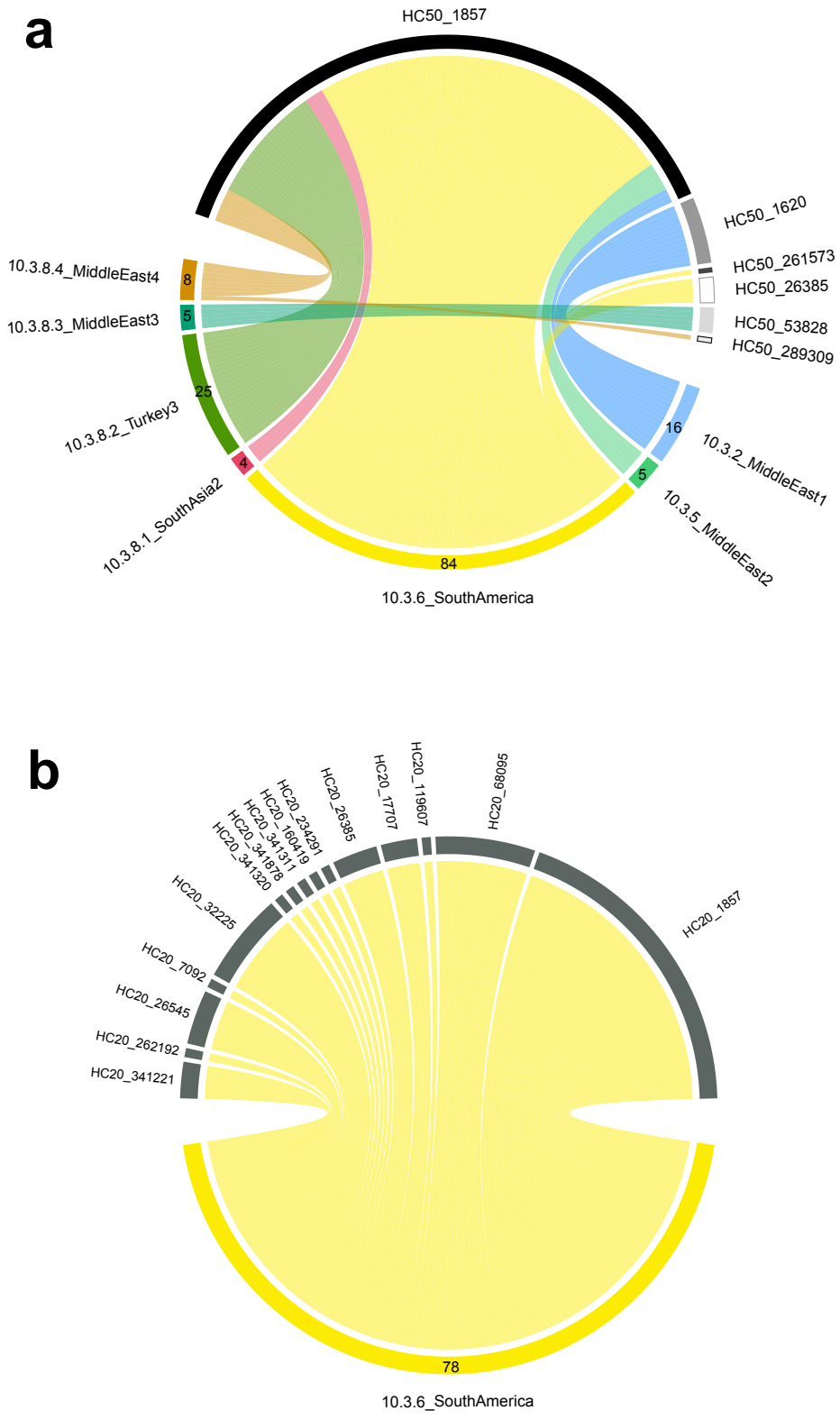


Supplementary Figure 10. Prophage content at 10 insertion sites (1-3, A, B, B', C-F) and phage types across the 11 lineages of SPB⁻ PG1. Ten prophage insertion sites were identified from the comparative analysis of the 14 complete genomes (**Supplementary Data 8**). The occupancy of the 10 insertion sites was assessed across the 11 lineages of SPB⁻ PG1, for the 568 isolates of the diversity dataset. Prophages at sites #1 to # 3 contain the *sopE* virulence gene. The schematic representation of the consensus chromosome starts at the *dnaA* locus (positions 1..1401).



Supplementary Figure 11. Prophage insertion polymorphism in the close vicinity of the phase 2 flagellin gene (*fljB)* of SPB⁻ PG1. a, The *hin-fljB-fljA* gene cluster encoding the phase 2 flagellin was examined in detail in the 14 complete genomes. *hin* is the flagellar phase variation DNA invertase gene; *fljB* is the phase 2 flagellin gene; *fljA* is the phase 1 flagellin gene (*fliC*) repressor gene. The *hin-fljB-fljA* gene cluster is located between the *iroB* (encoding a salmochelin biosynthesis C-glycosyltransferase) and tmRNA-*ssrA* loci. The cluster is deleted in three isolates (B1727, CIP 54.115, CIP A214). The gene arrow maps illustrate the six genomic environments detected between *iroB* and tmRNA-*ssrA*. **b,** The presence of the *hin-fljB-fljA* gene cluster and prophage content near the *fljB* locus were determined across the 11 lineages of SPB⁻ PG1, with the 568 isolates of the diversity set. Prophages were detected with short-read assemblies and the blastn algorithm. Further details are provided in **Supplementary Data 8.**

Supplementary Figure 12. A NINJA neighbour-joining GrapeTree for 567 SPB^{-} PG1 isolates from the diversity dataset. The tree nodes are colour-coded by lineage, genotype, cgMLST HC200, HC100, and HC50 clusters (see legends). The scale bars indicate the number of cgMLST allelic differences. The interactive version of the tree is publicly available from https://enterobase.warwick.ac.uk/ms_tree?tree_id=92095



Supplementary Figure 13. Correlation between cgMLST and genotyping data for the tracking of particular strains. Circular plots illustrating the difficulty of using cgMLST HC50 (**a**) or HC20 (**b**) clustering to identify the most frequently isolated genotype, 10.3.6_SouthAmerica. The flow bars are coloured according to the genotype. The number of isolates for each genotype is also indicated.

Supplementary References

1. Pratt, J.H. On paratyphoid fever and its complications. *Boston Med. Surg.* **148,** 137-142 (1903).
2. Proescher, F. & Roddy, J.A. Bacteriological studies on paratyphoid A and paratyphoid B. *Arch. Intern. Med.* **3,** 263-312 (1910).
3. Boycott, A.E. Observations on the bacteriology of paratyphoid fever and on the reactions of typhoid and paratyphoid sera. *J. Hyg.* **6,** 33-73 (1906).
4. F. Rathery, F., Ambard, L., Vansteenberghe, P. & Michel, R. Les fièvres paratyphoïdes B à l'hôpital mixte de Zuydcoote, de décembre 1914 à février 1916. 1st ed. F. Alcan, Paris, 248pp (1916).
5. Gradmann, C., Harrison M., & Rasmussen, A. Typhoid and the military in the early 20th century. *Clin. Infect. Dis.* **69,** S385–S387 (2019).
6. Felix, A. & Callow, B.R. Typing of paratyphoid B bacilli by Vi bacteriophage. *Br. Med. J.* **2,** 127-130 (1943).
7. Savage, W. Paratyphoid fever: an epidemiological study. *J. Hyg.* **42,** 393-410 (1942).
8. Sharp, J.C., Brown, P.P. & Sangster, G. Outbreak of paratyphoid in Edinburgh area. *Br. Med. J.* **1,** 1282-1285 (1964).
9. Newell, K.W., Hobbs, B.C. & Wallace, E.J. Paratyphoid fever associated with Chinese frozen whole egg; outbreaks in two bakeries. *Br. Med. J.* **2,** 1296-1298 (1955).
10. Sloan, R.S., Wilson, H.D. & Wright, H.A. The detection of a carrier of multiple phage-types of *Salmonella paratyphi B*. *J. Hyg.* **58,** 193-200 (1960).
11. Connor, T.R. *et al*. What's in a name? Species-wide whole-genome sequencing resolves invasive and noninvasive lineages of *Salmonella enterica* serotype Paratyphi B. *mBio* **7,** e00527-16 (2016).
12. Alikhan, N.F., Zhou, Z., Sergeant, M.J. & Achtman, M. A genomic overview of the population structure of *Salmonella*. *PLoS Genet.* **14,** e1007261 (2018).
13. Zhou, Z., Alikhan, N.F., Mohamed, K., Fan, Y.; Agama Study Group & Achtman, M. The EnteroBase user's guide, with case studies on *Salmonella* transmissions, *Yersinia pestis* phylogeny, and *Escherichia* core genomic diversity. *Genome Res.* **30,** 138-152 (2020).
14. Zhou, Z., Charlesworth, J. & Achtman, M. HierCC: a multi-level clustering scheme for population assignments based on core genome MLST. *Bioinformatics* **37,** 3645-3646 (2021).
15. Malorny, B., Bunge, C. & Helmuth, R. Discrimination of d-tartrate-fermenting and -nonfermenting *Salmonella enterica* subsp. *enterica* isolates by genotypic and phenotypic methods. *J. Clin. Microbiol.* **41,** 4292-4297 (2003).
16. Achtman, M., Hale, J., Murphy, R.A., Boyd, E.F. & Porwollik, S. Population structures in the SARA and SARB reference collections of *Salmonella enterica* according to MLST, MLEE and microarray hybridization. *Infect. Genet. Evol.* **16,** 314-325 (2013).
17. Kelterborn, E. *Salmonella*-species. Erstfunde, Namen und Vorkommen. Den Haag (Junk) 535 pp (1967).
18. Aoki, Y. Distribution of *Salmonella* Types in East Asia. *Endemic Diseases Bulletin of Nagasaki University* **7,** 192-220 (1965).
19. Achtman M. *et al*. Multilocus sequence typing as a replacement for serotyping in *Salmonella enterica*. *PLoS Pathog.* **8,** e1002776 (2012).
20. Longfellow, D. & Luippold, G.F. Typhoid Vaccine Studies VII: Typhoid-Paratyphoid Vaccine. *Am. J. Public Health Nations Health* **33,** 561-568 (1943).

21. Gard, S. Ein neuer Salmonella-Typ (S. abortus canis). *Zeitschr. f. Hygiene*. **121,** 139-141 (1938).
22. Magnusson, K.E. Ein Hund als Ansteckungsquelle von Paratyphusinfektionen. Zeitschr. f. Hygiene. **121,** 136–138 (1938).
23. Allos, G. Inventaire des sérotypes de *Salmonella* rencontrés en Irak. *Bull. Soc. Pathol. Exot. Filiales* **71,** 323-328 (1978).
24. Zhou, Z. *et al.* Transient Darwinian selection in *Salmonella enterica* serovar Paratyphi A during 450 years of global spread of enteric fever. *Proc. Natl Acad. Sci. USA* **111,** 12199–12204 (2014).
25. Gwyn, N.B. On infection with a para-colon bacillus in a case with all the clinical features of typhoid fever. *Bull. Johns Hopkins Hosp.* **9,** 54-56 (1898).
26. Chattaway, M.A. *et al*. Phylogenomics and antimicrobial resistance of *Salmonella* Typhi and Paratyphi A, B and C in England, 2016-2019. *Microb. Genom.* **7,** 000633 (2021).
27. Zhou, Z. *et al.* Pan-genome analysis of ancient and modern *Salmonella enterica* demonstrates genomic stability of the invasive Para C lineage for millennia. *Curr. Biol.* **28,** 2420-2428.e10 (2018).
28. Prager, R. *et al*. Molecular properties of *Salmonella enterica* serotype Paratyphi B distinguish between its systemic and its enteric pathovars. *J. Clin. Microbiol.* **41,** 4270-4278 (2003).
29. Friebel, A. *et al*. SopE and SopE2 from *Salmonella typhimurium* activate different sets of RhoGTPases of the host cell. *J. Biol. Chem.* **276,** 34035-34040 (2001).
30. Prager, R. *et al*. Prevalence and polymorphism of genes encoding translocated effector proteins among clinical isolates of *Salmonella enterica*. *Int. J. Med. Microbiol.* **290,** 605-617 (2000).
31. Mirold, S., Rabsch, W., Tschäpe, H. & Hardt,W.D. Transfer of the *Salmonella* type III effector sopE between unrelated phage families. *J. Mol. Biol.* **312,** 7-16 (2001).
32. Tassinari, E. *et al*. Whole-genome epidemiology links phage-mediated acquisition of a virulence gene to the clonal expansion of a pandemic *Salmonella enterica* serovar Typhimurium clone. *Microb. Genom.* **6,** mgen000456 (2020).
33. Ehrbar, K., Mirold, S., Friebel, A., Stender, S. & Hardt, W.D. Characterization of effector proteins translocated via the SPI1 type III secretion system of *Salmonella typhimurium*. *Int. J. Med. Microbiol.* **291,** 479-485 (2002).
34. Vonaesch, P. *et al*. The *Salmonella* Typhimurium effector protein SopE transiently localizes to the early SCV and contributes to intracellular replication. *Cell Microbiol.* **16,** 1723-1735 (2014).
35. Hoffmann, M. *et al*. Comparative genomic analysis and virulence differences in closely related *Salmonella enterica* serotype Heidelberg isolates from humans, retail meats, and animals. *Genome Biol. Evol.* **6,** 1046-1068 (2014).
36. Hawkey J. *et al*. Global population structure and genotyping framework for genomic surveillance of the major dysentery pathogen, *Shigella sonnei*. *Nat. Commun.* **12,** 2684 (2021).
37. Felix A. World survey of typhoid and paratyphoid-B phages types. *Bull. World Health Organ.* **13,** 109-170 (1955).
38. Nicolle P. Rapport sur la distribution des lysotypes de *Salmonella typhi* et de *S. paratyphi B* dans le monde, d'après les résultats fournis par les centres nationaux membres du comité international de la lysotypie entérique à l'occasion du congrès international de microbiologie, Stockholm, 1958. *Ann. Inst. Pasteur* **102,** 389-409 (1962).

39. International Committee for Enteric Phage-Typing (ICEPT). The geographical distribution of *Salmonella typhi* and *Salmonella paratyphi A* and *B* phage types during the period 1 January 1966 to 31 December 1969. *J. Hyg.* **71,** 59-84 (1973).
40. Vieu, J.F., Binette, H. & Leherissey, M. Salmonella paratyphi B d-tartrate positif (var. java): lysotypie de 1200 souches isolées en France (1975-1985). *Zentralbl. Bakteriol. Mikrobiol. Hyg. A*. **268,** 424-432 (1988).
41. Fukumi, H. *et al.* Epidemiological investigations of typhoid and paratyphoid fever in Japan with aid of the phage typing method. I. Distribution of phage-types of *Salmonella typhi* and *Salmonella paratyphi B* in Japan, 1956-1965. *Jpn J. Med. Sci. Biol.* **20,** 447-460 (1967).
42. Zhao, Y. *et al.* PanGP: A tool for quickly analyzing bacterial pan-genome profile. *Bioinformatics* **30**, 1297–1299 (2014).