

Supplementary Material

**Elucidation of global and local epidemiology of *Salmonella* Enteritidis
through multilevel genome typing**

Lijuan Luo¹, Michael Payne¹, Sandeep Kaur¹, Dalong Hu¹, Liam Cheney¹, Sophie Octavia¹, Qinning Wang², Mark M. Tanaka¹, Vitali Sintchenko^{2,3} and Ruiting Lan^{1,*}

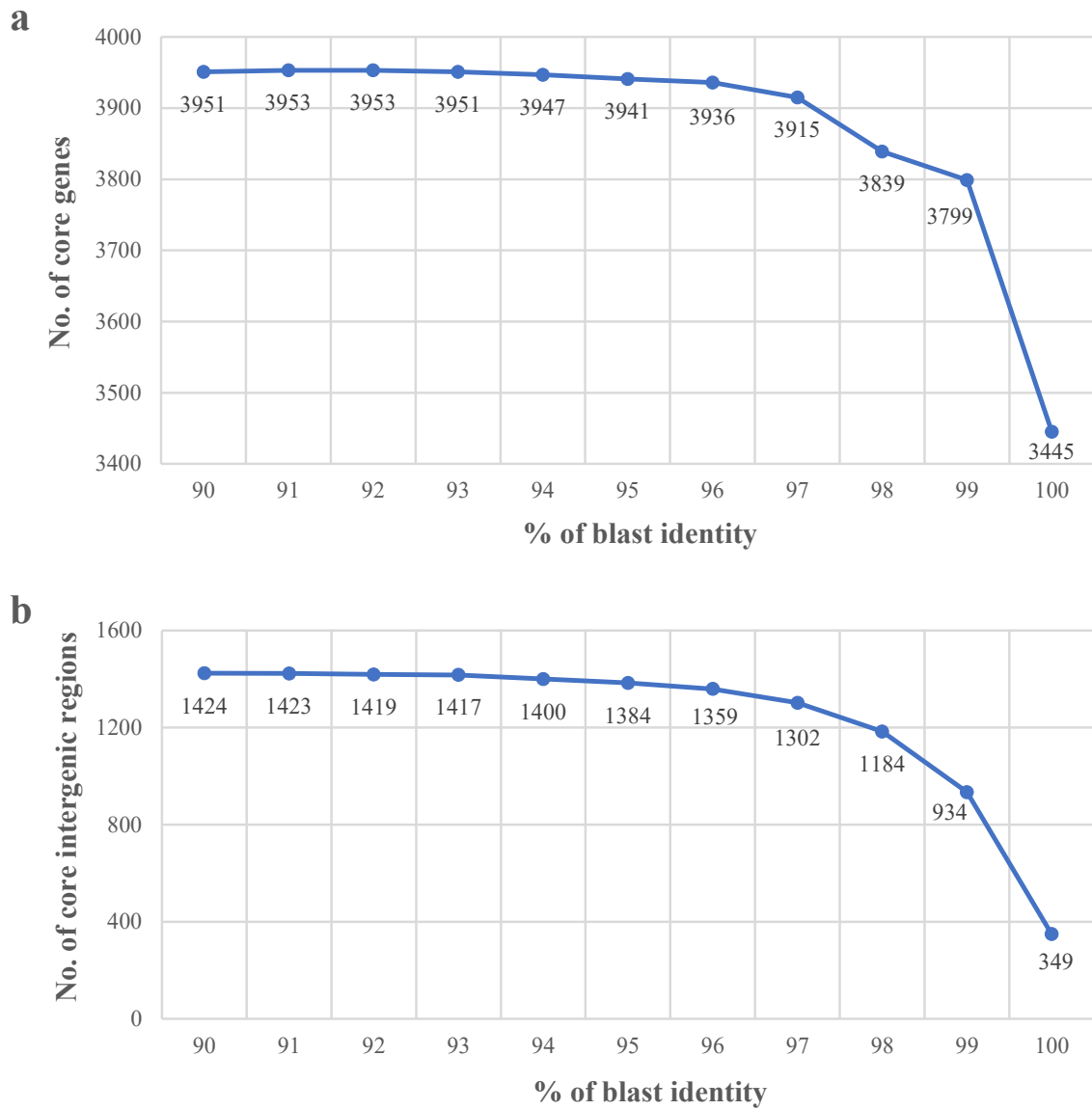


Fig. S1. Blast identity and the number of core genes and core intergenic regions defined for *S. Enteritidis*. **a.** Number of core genes defined by Roary v3.11.2 with varied blast identity from 90% to 100% and gene coverage of 99%. **b.** Number of core intergenic regions defined by Piggy v3.11.2 with varied blast identity from 90% to 100% and coverage of 99% using the Roary output.

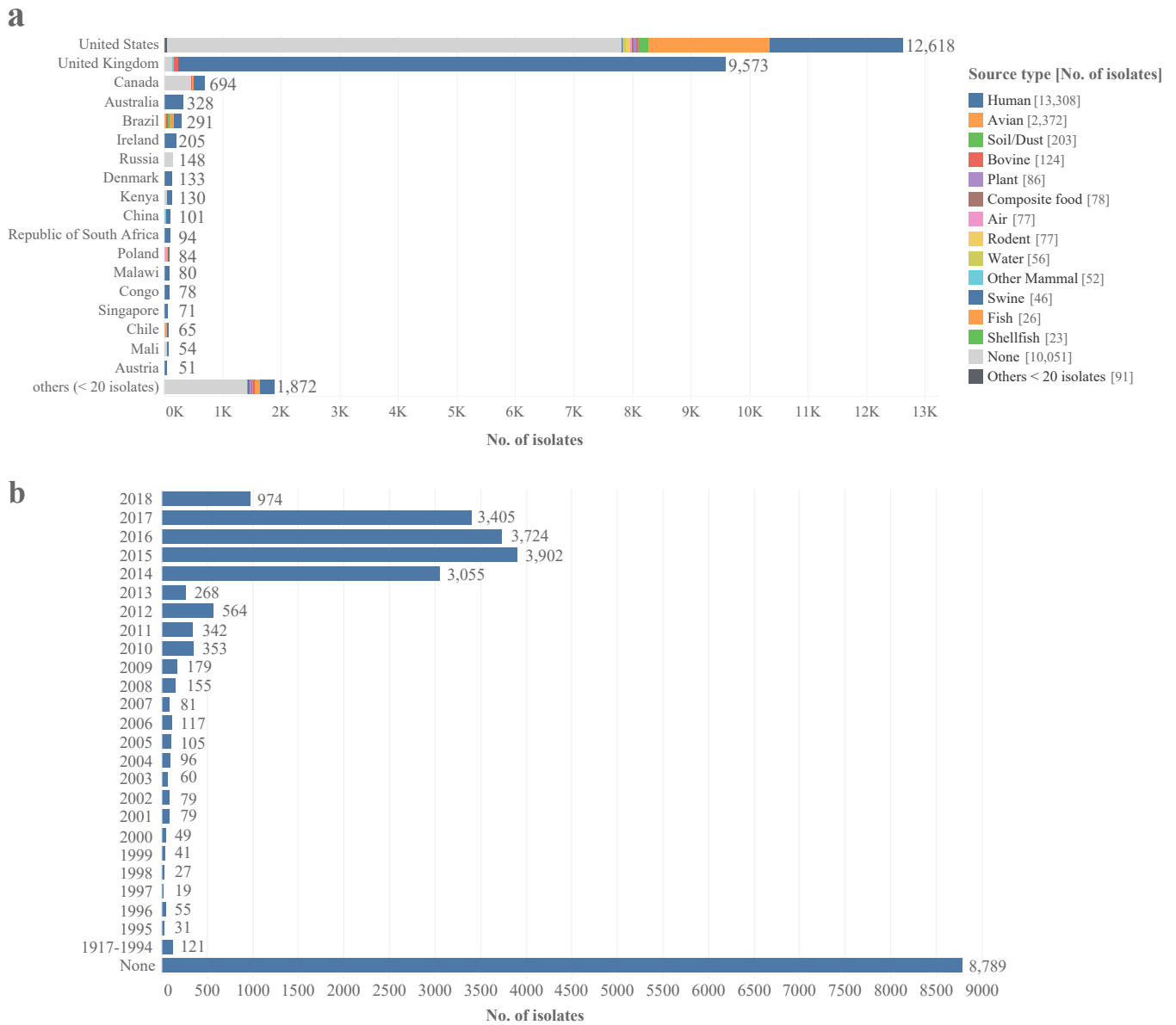


Fig. S2. General epidemiological information for the 26670 isolates analysed with the following general metadata information: a. isolated from 86 different countries [with the majority of the isolates isolated from the United States (47.3%) and United Kingdom (35.9%)]; isolated from 26 different source types (the source types of each country were shown in different colours); **b.** collected between 1917 and 2018 [with 2014 (11.5%), 2015 (14.6%), 2016 (14.0%), 2017 (12.8%) and 2018 (3.7%) having more than 500 isolates each].

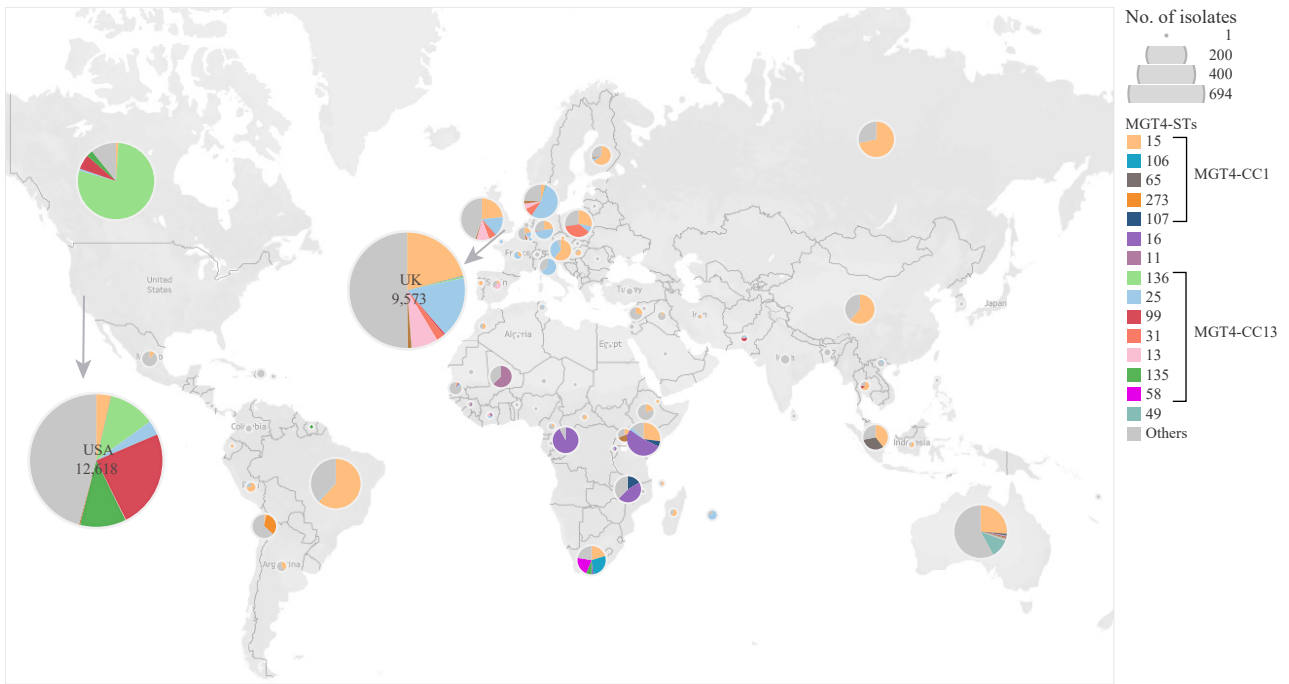


Fig. S3. MGT4-STs distribution in different countries. Although the majority of the *S. Enteritidis* genomes were from the USA and UK, there were 3017 isolates belonged to the other 86 countries. The top 15 MGT4-STs, representing 66.2% of the isolates, were represented with different colours. The size of each pie represented the number of isolates in each country except for USA (12,618) and UK (9,573). There were five top STs belonged MGT4-CC1 cluster, with MGT4-ST15 prevalent in most of the countries. Seven STs belonged to the MGT4-CC13 cluster, which were mainly observed in Canada and Europe countries. MGT4-ST16 was observed in Central/Eastern Africa, while MGT4-ST11 was prevalent in West Africa. MGT4-ST49 was unique to Australia and belonged to one of the two lineages outside the major global one. This map chart was created with Tableau v2019.2.

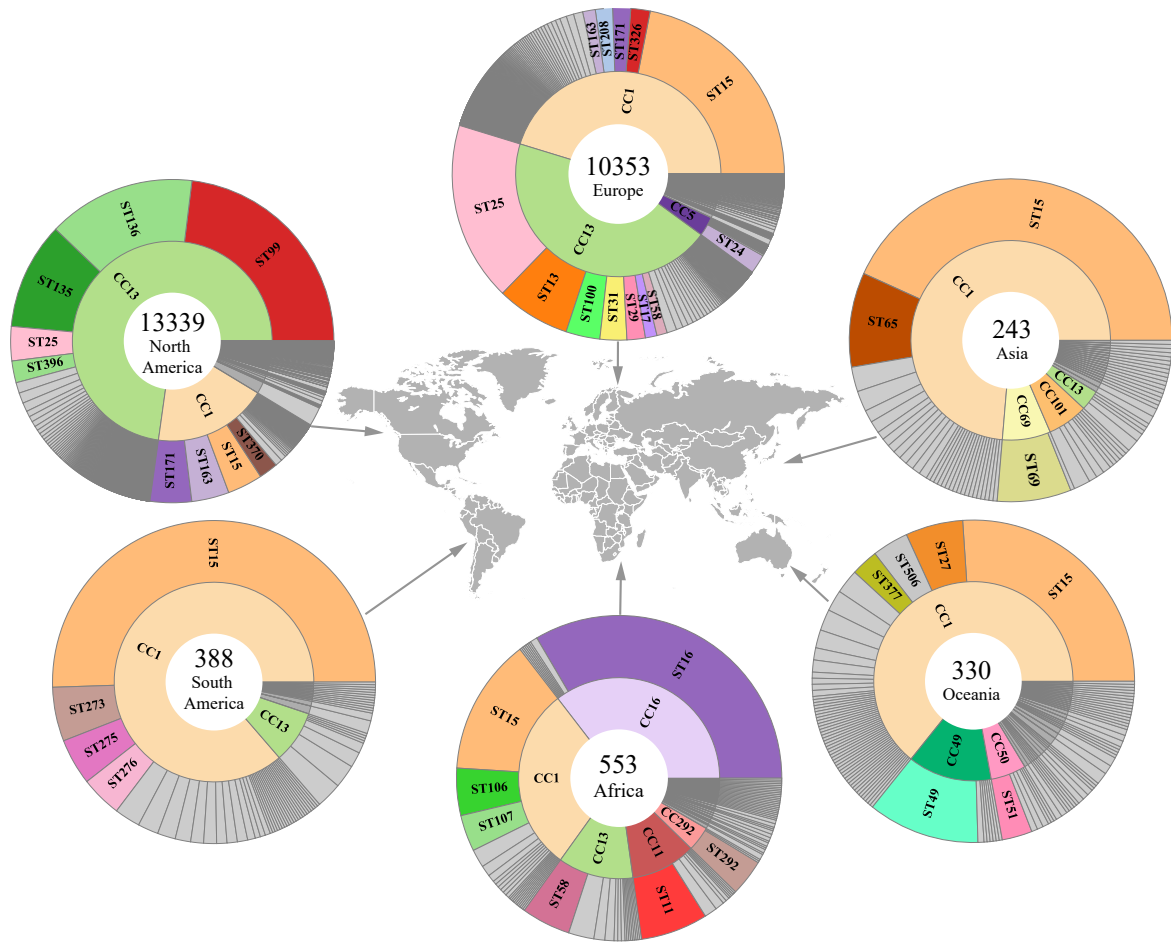


Fig. S4. Continental restricted STs of *S. Enteritidis* were identified within each CCs at MGT4. The numbers in each pie refer to the number of whole genome sequences analysed in each continent. The inner ring shows the main MGT4-CCs (with proportions of 3% or more) in each continent. MGT4-CC1 and CC13 were the two predominant MGT4-CCs with MGT4-CC1 being more widely distributed in all six continents. The outer ring shows the common MGT4-STs in each continent (with proportions of 2% or higher) within corresponding CCs. MGT4-ST15 was the dominant ST in MGT4-CC1, which were widely distributed all over the world. Continental restricted STs were identified, especially for the STs of MGT4-CC13, which was more prevalent in Europe and North America.



Fig. S5. Geographic and predicted plasmid type distribution of MDR isolates. a. Predicted antibiotic resistance of *S. Enteritidis* in different continents. The proportion of predicted antibiotic resistant isolates in different continents were shown with different colours representing different numbers of resistant drug classes. Isolates harbouring antimicrobial genes resistant to 3 or more drug classes (MDR) were represented with red colour, 2 different drug classes with green colour, 1 drug class with yellow colour, and no drug resistant gene with grey colour. **b.** Presence of AR genes in *S. Enteritidis* isolates with different predicted plasmid types. Different colors represent different types of AR. AR genes were of high proportion in *S. Enteritidis* isolates with IncQ1, IncN, IncI1, IncX1 plasmids. Particularly, 98.0% of the isolates with plasmid type IncQ1, were MDR associated.

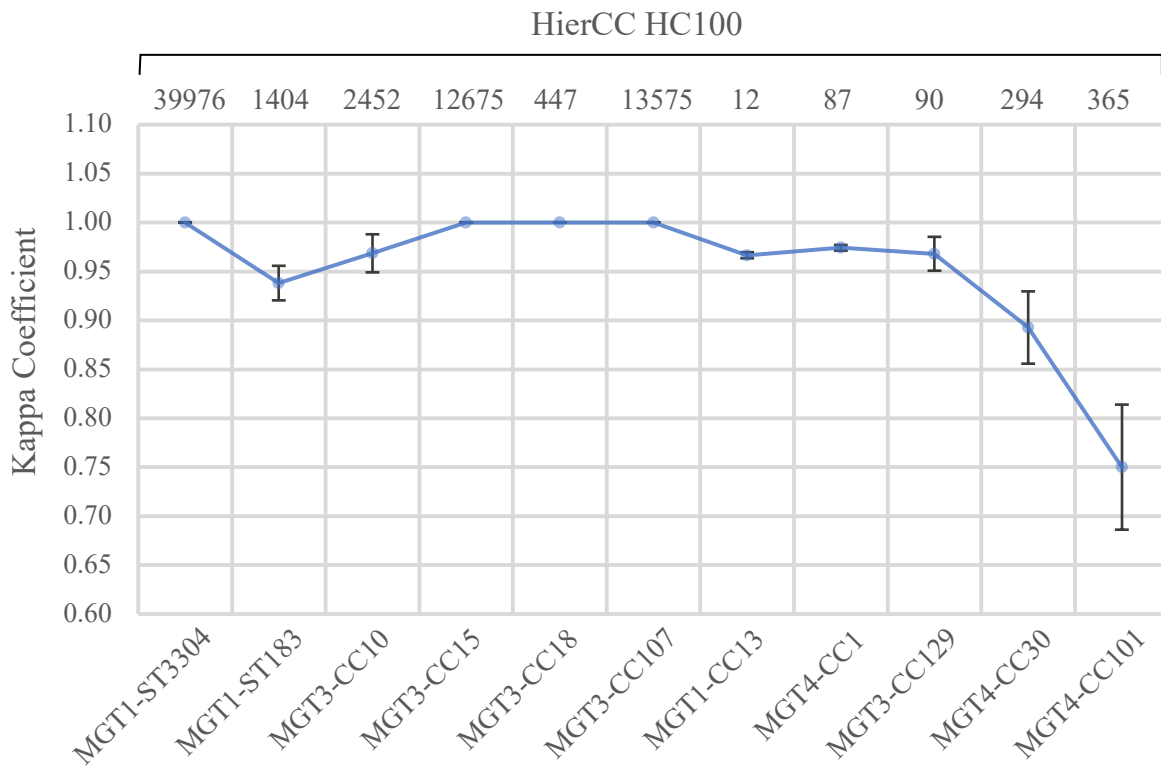


Fig. S6. Population structure markers of *S. Enteritidis* by MGT STs/CCs and HierCC HC100 cluster types. Three MGT1 (or 7 gene MLST) STs were representative of the Oceania prevalent clade A and clade C *S. Enteritidis* isolates, with MGT1-ST3304 as the main type. We identified 10 main lineages of clade B *S. Enteritidis*, which can be represented by 10 MGT-STs/CCs markers. By HierCC HC100, there were 10 different HC100 types corresponding to the 10 MGT lineage markers. The Cohen's kappa coefficient was calculated for each pair of the MGT-ST/CC lineage marker and the matched HC100 type using SAS v9.4. And the 95% confidential interval of the Cohen's kappa coefficient were shown. The closer the coefficient value were to one, the higher agreement between the two methods were.

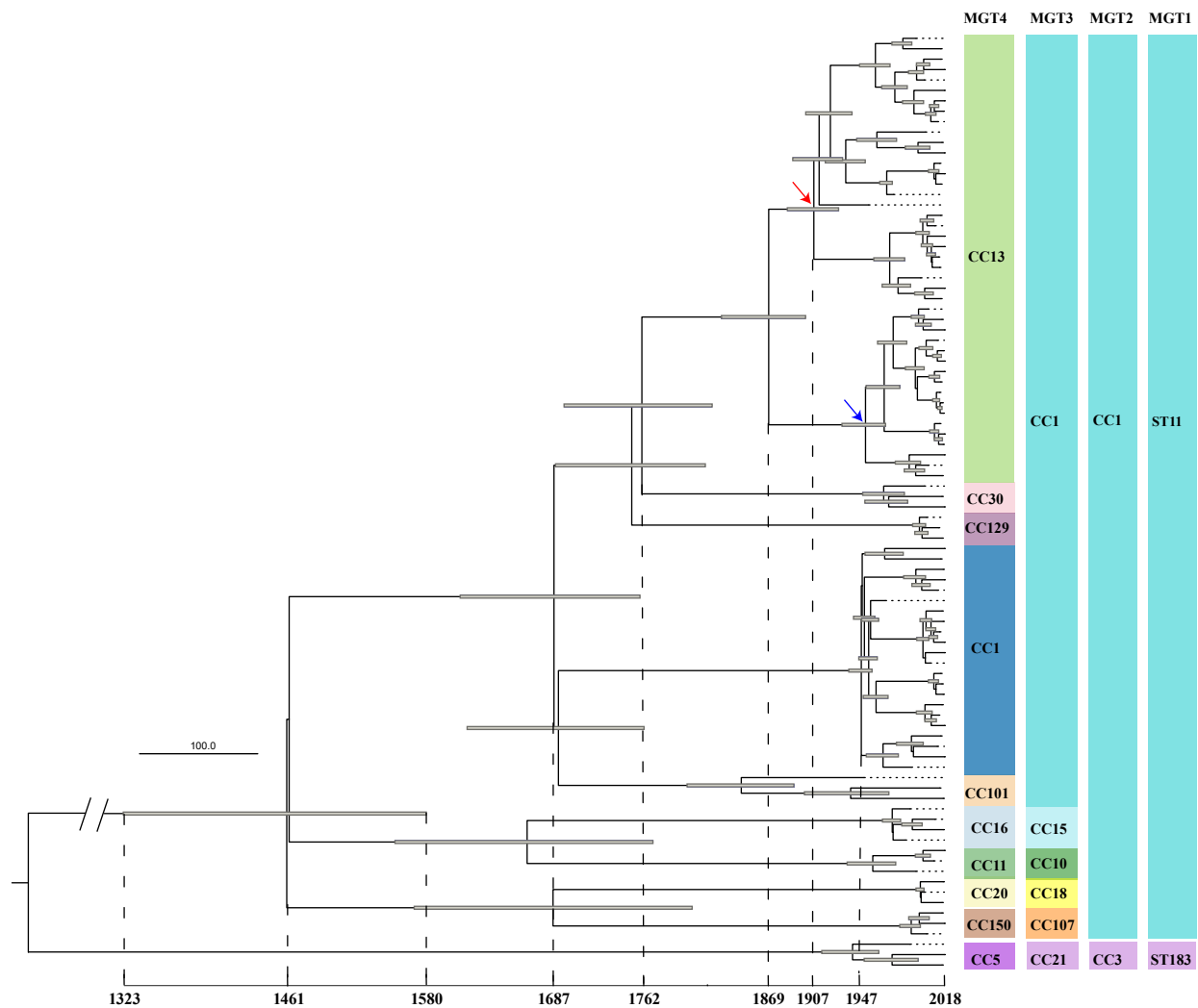


Fig. S7. Evolutionary phylogeny of the global clade estimated with BEAST. A total of 90 representative isolates were selected to generate a phylogeny with divergence date estimated using BEAST v1.10.4 with the uncorrelated relaxed log-normal clock and the Bayesian skyline model. The estimated divergence times were indicated for selected ancestral nodes with the 95% year confidence interval shown as error bars at each node. The CCs from MGT2 to MGT4 levels are shown next to the isolates, and the majority of the MGT2-CC1 isolates belonged to ST11 by 7 gene MLST (or MGT1). The most recent common ancestor (MRCA) of the 9 lineages belonging to MGT2-CC1, dated back to around 1460s (95% CI 1323-1580). Around 1687 (95% CI 1608-1760), MGT3-CC1 was separated into three lineages of MGT4-CC1, CC13 and CC101. Around 1869 (95% CI 1829-1900), the MGT4-CC13 lineage evolved into two sub-lineages of varied MGT4-STs, with one more prevalent in North America (labelled with red arrow) and the other one sub-lineage more prevalent in the Europe (labelled with blue arrow).

1 **Supplementary methods:**

2 **Core gene and core intergenic region definition of *S. Enteritidis***

3 The *S. Enteritidis* isolates from Enterobase (during May, 2018) were sub-sampled to identify
4 representative isolates in three different ways depending on the size of the ribosomal sequence type
5 (rST) [1]. For the isolates of the rSTs with fewer than 10 isolates each, all of the isolates were
6 sampled. For the isolates of the rST with more than 10 and fewer than 350 isolates assigned, 10
7 isolates for each type were randomly sampled. For the isolates of the rST with more than 350 isolates
8 (rST1425 and rST3888), 2% of the isolates for each type were sampled randomly while including
9 isolates from as many countries and collection years as possible. A total of 1801 genomes, which
10 were selected to represent 283 rSTs collected over 101 years (1917-2018) in 56 different countries.

11 The trimmed raw reads were then assembled using SPAdes v3.13.0 [2]. Prokka v1.12 was used to
12 predict and annotate genes [3]. Roary v3.11.2 was used for the determination of core genes, which
13 were defined as the genes present in 99% of the representative genomes. Core genes were identified
14 as present in each genome following the nucleotide identity threshold $\geq 96\%$ and alignment
15 coverage threshold $\geq 99\%$. Those thresholds were assessed as the optimal identity by comparing the
16 number of core genes generated with different identity from 90% to 100% (**Fig. S1**) [4]. Core genes
17 with duplicates in the Roary output (paralogues) were removed. Core intergenic regions were defined
18 using Piggy v3.11.2 based on the Roary outputs [5], and the nucleotide identity and coverage
19 thresholds were identical to those of core genes.

20 **Phylogenetic analysis**

21 To define the major clades of *S. Enteritidis* globally, 1508 isolates were sampled based on the STs of
22 MGT6 (random sampling of one isolate per ST with 3 or more isolates). A phylogenetic tree was

1 constructed using ParSNP v1.2, which called core-genome SNPs [6, 7]. The branches with more than
2 10 isolates on the phylogenetic tree were collapsed together with iTOL v4 [8].

3 Average mutation rates were estimated for the global epidemic clade B of *S. Enteritidis* and the
4 two dominant lineages, MGT4-CC1 and CC13. For the global epidemic clade B, a total of 90 isolates
5 were randomly sampled based on MGT4-CCs and STs (covering the main STs for each CC). For the
6 mutation rate of MGT4-CC1 or CC13, representative isolates were randomly sampled based on the
7 higher resolution level MGT5 and collections years. For each MGT5-CC type, one isolate from each
8 collection year were randomly sampled. SNPs derived from MGT9 alleles were then concatenated as
9 a multi-alignment file and potential recombinant SNPs were removed using both Gubbins v2.0.0 and
10 Recdetect v6.0 [9, 10]. Beast v1.10.4 was then used to estimate the mutation rate based on mutational
11 SNPs [11]. A total of 24 combinations of clock and population size models were evaluated with the
12 MCMC chain of 100 million states. Tracer v1.7.1 was used to identify the optimal model and to
13 estimate population expansion over time [12]. For the predominant clade of *S. Enteritidis* including
14 isolates of multiple lineages, uncorrelated relaxed lognormal clock combined with the Bayesian
15 skyline model was chosen by the effective sample size (ESS) of both mean rate and joint probability
16 of over 300. For the two dominant lineages of *S. Enteritidis*, MGT4-CC1 or CC13, the strict clock
17 and Bayesian skyline models were found to be the optimal combination model by the ESS of mean
18 rate and joint probability of over 100. The large size of each CC (more than 300 isolates for the Beast
19 analysis) made it hard to reach 200 of ESS even for the optimal models.

20 **Supporting Results:**

21 **Potential repeat sequencing bias evaluation**

22 To evaluate any bias that may be caused by resequencing of the same strain or isolate, we identified
23 all isolates of the same ST based on MGT9 and same metadata based on collection country,
24 collection year and month, and source type. Such isolates were conservatively treated as repeat

1 sequencing of the same isolate and the “duplicates” were removed from the dataset. A total of 4,026
2 out of the 26,670 isolates (17.8%,) were identified and “duplicates” removed from the dataset. The
3 prevalence of STs in different geographic regions, collection time and sources, were re-evaluated in
4 the reduced dataset of 22,644 isolates and compared with the original dataset. Based on the number
5 of isolates after resampling, the rank of STs and CCs in different geographic regions, collection time
6 and sources were compared against the original dataset.

7 1. For the continental distribution of *S. Enteritidis* (compared against **Fig. 2**):

8 The global distribution of *S. Enteritidis* were reevaluated based on MGT4 STs and CCs. The top
9 10 MGT4 CCs (with ≥ 10 isolates each) in each continent remain the same after resampling,
10 except for MGT4 CC378, which dropped from rank 9 to rank 12 in Europe (**Data Set S2, Tab 1**).

11 To compare the rank of the CCs before and after sampling, Kendall's tau value was calculated
12 [13]. If Kendall's tau value is close to 1, it indicates strong agreement. If Kendall's tau value is
13 close to -1, it indicates strong disagreement. The ranks of CCs in each continent before and after
14 sampling strongly agreed, with Kendall's tau value of 0.917 ($P < 0.001$).

15 The top 20 MGT4 STs (with ≥ 10 isolates each) in each continent were also the same, except for
16 MGT4-ST2060 dropped from rank 15 to 24 in North America (**Data Set S2, Tab 2**). The ranks of
17 STs agreed strongly with Kendall's tau of 0.944 ($P < 0.001$).

18 2. For the states and source distribution of *S. Enteritidis* in the USA (compared against **Fig. 4**):

19 The top 10 MGT4 STs in the USA remained the same before and after the resampling (**Data Set**
20 **S2, Tab 3**). The rank of the STs across all states has a Kendall's tau value as 0.987 ($P < 0.001$).

21 3. For the temporal analysis of *S. Enteritidis* in UK (compared against **Fig. 5**):

22 For the rank of UK MGT5 STs in different months before and after “duplicates” removal, the top
23 13 MGT5 STs remain the same (**Data Set S2, Tab 4**). The Kendall's tau value based on the rank
24 of STs in each month was 0.900 ($P \text{ value} < 0.001$).

1 4. For the outbreak propensity evaluation between MGT4-CC1 and CC13 using the clinical *S.*
2 Enteritidis isolates from UK during 2014 to 2018 (compared against **Data Set S1, Tab 4 and Tab**
3 **5**):

4 The top 12 STs for large outbreak clusters evaluation remain the same (**Data Set S2, Tab 5**).
5 Although the predicted OR value adjusted slightly, the P values remain significant. It must be
6 noted that the artificial removal of “duplicates” effectively removed many potential outbreak
7 isolates. Importantly, MGT4-CC13 was confirmed to be significantly associated with large
8 outbreak clusters than MGT4-CC1 (**Data Set S2, Tab 6**).

9 In summary, by removing potential “duplicates”, recalculation and comparison with the original
10 dataset, no significant difference was observed in either top STs and their ranks. Thus, any bias
11 brought by potential repeat sequencing of the same isolates or even same strain from an outbreak was
12 negligible.

13

14 **References**

- 15 1. Alikhan NF, Zhou Z, Sergeant MJ, Achtman M: **A genomic overview of the population**
16 **structure of *Salmonella***. *PLoS Genet* 2018, **14**(4):e1007261.
- 17 2. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM,
18 Nikolenko SI, Pham S, Pribelski AD *et al*: **SPAdes: a new genome assembly algorithm**
19 **and its applications to single-cell sequencing**. *J Comput Biol* 2012, **19**(5):455-477.
- 20 3. Seemann T: **Prokka: rapid prokaryotic genome annotation**. *Bioinformatics* 2014,
21 **30**(14):2068-2069.
- 22 4. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MT, Fookes M, Falush D,
23 Keane JA, Parkhill J: **Roary: rapid large-scale prokaryote pan genome analysis**.
24 *Bioinformatics* 2015, **31**(22):3691-3693.

- 1 5. Thorpe HA, Bayliss SC, Sheppard SK, Feil EJ: **Piggy: a rapid, large-scale pan-genome**
2 **analysis tool for intergenic regions in bacteria.** *GigaScience* 2018, **7(4):1-11.**
- 3 6. Treangen TJ, Ondov BD, Koren S, Phillippy AM: **The Harvest suite for rapid core-genome**
4 **alignment and visualization of thousands of intraspecific microbial genomes.** *Genome*
5 *Biol* 2014, **15(11):524.**
- 6 7. Price MN, Dehal PS, Arkin AP: **FastTree 2--approximately maximum-likelihood trees for**
7 **large alignments.** *PloS one* 2010, **5(3):e9490-e9490.**
- 8 8. Letunic I, Bork P: **Interactive Tree Of Life (iTOL) v4: recent updates and new**
9 **developments.** *Nucleic Acids Res* 2019, **47(W1):W256-W259.**
- 10 9. Croucher NJ, Page AJ, Connor TR, Delaney AJ, Keane JA, Bentley SD, Parkhill J, Harris
11 **SR: Rapid phylogenetic analysis of large samples of recombinant bacterial whole**
12 **genome sequences using Gubbins.** *Nucleic Acids Res* 2015, **43(3):e15.**
- 13 10. Hu D, Liu B, Wang L, Reeves PR: **Living Trees: High-Quality Reproducible and**
14 **Reusable Construction of Bacterial Phylogenetic Trees.** *Molecular biology and evolution*
15 2020, **37(2):563-575.**
- 16 11. Suchard MA, Lemey P, Baele G, Ayres DL, Drummond AJ, Rambaut A: **Bayesian**
17 **phylogenetic and phylodynamic data integration using BEAST 1.10.** *Virus Evol* 2018,
18 **4(1):vey016.**
- 19 12. Rambaut A, Drummond AJ, Xie D, Baele G, Suchard MA: **Posterior Summarization in**
20 **Bayesian Phylogenetics Using Tracer 1.7.** *Syst Biol* 2018, **67(5):901-904.**
- 21 13. Kendall MG: **A new measure of rank correlation.** *Biometrika* 1938, **30(1/2):81-93.**

22