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

Lijuan Luo[1], Michael Payne[1], Sandeep Kaur[1], Dalong Hu[1], Liam Cheney[1], Sophie Octavia[1], Qinning Wang[2], Mark M. Tanaka[1], Vitali Sintchenko[2,3] and Ruiting Lan[1,*]

[1]School of Biotechnology and Biomolecular Sciences, University of New South Wales, Sydney, New South Wales, Australia

[2]Centre for Infectious Diseases and Microbiology–Public Health, Institute of Clinical Pathology and Medical Research – NSW Health Pathology, Westmead Hospital, New South Wales, Australia

[3]Marie Bashir Institute for Infectious Diseases and Biosecurity, Sydney Medical School, University of Sydney, New South Wales, Australia

*Corresponding Author

Email: r.lan@unsw.edu.au

Phone: 61-2-9385 2095

Fax: 61-2-9385 1483

Keywords: *Salmonella* Enteritidis, MGT, global epidemiology, population structure, mutation rate, global database

Repositories: There were no newly sequenced data in this study.

Abstract

Salmonella Enteritidis is a major foodborne pathogen that causes both local and international outbreaks with complex transmission pathways. Invasive infections which are associated with multidrug resistance (MDR) are an increasing concern around the globe. However, the global epidemiological picture of *S.* Enteritidis remains unclear. Here, using the novel multilevel genome typing (MGT) approach, we have characterised the genomic epidemiology of *S.* Enteritidis in unprecedented detail. We examined 26,670 publicly available *S.* Enteritidis whole genome sequences from 86 countries over 101 years to reveal their spatial and temporal distributions. Using MGT4 and MGT5, we identified globally prevalent and regionally restricted sequence types (STs). Source associated STs were identified, such as poultry associated MGT4-STs, which were common in human cases in the USA. Temporal trends were observed in the UK with MGT5-STs from 2014 to 2018, revealing both long lived endemic STs and the rapid expansion of new STs. Using MGT3 to MGT6, we identified MDR associated STs to facilitate tracking MDR spread. The majority of the global *S.* Enteritidis population fell within two predominant lineages with significant differences in geographic distribution, outbreak frequency, antimicrobial resistance and mutation rate. An online open MGT database has been established for unified international surveillance of *S.* Enteritidis. We demonstrated that MGT provides a flexible and high-resolution genome typing tool for *S.* Enteritidis surveillance and outbreak detection.

Impact statement

Salmonella Enteritidis is a common foodborne pathogen, causing both diarrheal and invasive infections with distinct geographic and phylogenetic distributions. Large scale international outbreaks were frequently reported in certain subtypes of *S*. Enteritidis. Multilocus sequence typing is a well-established strain typing method based on exact comparisons of gene sequences. It is highly desirable to develop an MLST based stable and uniform genomic nomenclature system to track and examine these high-risk genomic subtypes. In this work, a novel MLST based genomic typing scheme, Multilevel Genome Typing (MGT) was established for *S*. Enteritidis. The global epidemiological characteristics of *S*. Enteritidis were elucidated. A total of 26,670 publicly available whole genome sequences from 86 countries over 101 years were analysed: globally prevalent sequence types (STs), regionally restricted STs, large scale outbreaks, invasive infections and multidrug resistance (MDR) associated STs were identified. These STs can facilitate the identification and control of outbreaks and help tracking the spread of MDR. Moreover, we developed an online MGT database for unified global public health surveillance for *S*. Enteritidis.

Abbreviations

MGT: Multilevel genome typing; ST: Sequence type; CC: Clonal complex; ODC: Outbreak detection cluster; SNP: Single nucleotide polymorphism; MLST: Multi-locus sequence typing; cgMLST: core genome multi-locus sequence typing; HierCC: Hierarchical clustering of cgMLST. AR: Antibiotic resistance; MDR: Multi-drug resistance.

Data Summary

1. The MGT database for *S.* Enteritidis is available at https://mgtdb.unsw.edu.au/enteritidis/.

2. All accession numbers of the public available genomes were available in the MGT database and Data Set S1, Tab 7. And there were no newly sequenced data in this study.

3. Supplementary material: Supplementary Fig. S1 to S7, supplementary methods and supporting results about the evaluation of potential repeat sequencing bias.

4. Data Set S1: Supporting tables of the main results.

5. Data Set S2. Supporting tables of the repeat sequencing bias evaluation by removing the potential repeat sequencing isolates. Note outbreak isolates may also be removed.

Introduction

Nontyphoidal *Salmonella* is ranked second in foodborne disease in Europe and North America [1, 2], with *Salmonella* Enteritidis a dominant serovar in many countries [3, 4]. *S.* Enteritidis mainly causes human gastrointestinal infections leading to diarrhoea. However, invasive infections manifesting as sepsis, meningitis and pneumonia, have recently been reported [5, 6]. As many *S.* Enteritidis strains causing invasive infections carry plasmids which confer multidrug resistance (MDR), the spread of these strains internationally is a major threat to global health [7, 8]. Additionally, *S.* Enteritidis has caused numerous large-scale national and international outbreaks with complex transmission pathways [3, 9, 10]. While there have been limited studies of the total global epidemiology of *S.* Enteritidis genomic subtypes [5, 11], the geographic distribution, outbreak propensity, MDR and evolutionary characteristics of different lineages of *S.* Enteritidis have not been systematically evaluated using the large numbers of newly available public genome sequences. A rapid, stable and standardized global genomic typing strategy for *S.* Enteritidis is required for the high-resolution and scalable surveillance of outbreaks, tracking of international spread and MDR profiles of *S.* Enteritidis.

Sequence types (STs) are based on exact matching of genes between isolates. The well-established 7 gene multi-locus sequence typing (MLST) of *Salmonella* is widely used [12]. However, as only 7 housekeeping genes are compared, the vast majority of *S.* Enteritidis are assigned to ST11. STs of higher resolution are required. With this in mind core genome MLST (cgMLST) and whole genome

MLST (wgMLST) schemes of *Salmonella* were developed including 3002 and 21,065 loci, respectively [13]. However, STs based on cgMLST and wgMLST are too diverse to offer useful relatedness information at anything but the finest scales. This issue was addressed by single linkage hierarchic typing systems, i.e., Single Nucleotide Polymorphism (SNP) address and Hierarchical Clustering of cgMLST (HierCC) [14, 15]. One major disadvantage of the pairwise comparison based single linkage clustering method is that the cluster types called may merge when additional data is added and differences in order of addition can result in different clusters [16]. For example, when an isolate meets the SNP/allele difference threshold to two clusters, this isolate would act as a bridge connecting the two clusters into one. This scenario will almost certainly occur in a large database with ever expanding numbers of sequences, especially at finer resolutions, and would be an obstacle for long term epidemiological investigation and comparison [16]. We recently have developed a novel core genome based method called multilevel genome typing (MGT), an exact matching method to assign multiple resolutions of relatedness without the need for clustering [16]. MGT is a genome sequence based typing system, including multiple MLST schemes with increasing resolution from the classic 7 gene MLST to cgMLST where each scheme is used to independently assign an ST. An MGT scheme and a public database were established for *Salmonella* Typhimurium where 7 to 5268 loci were included in MGT1 to 9 [16]. MGT1 corresponds to the classic 7-gene MLST scheme of *S. enterica* [12]. For *S.* Typhimurium, the stable STs from MGT levels of appropriate resolution were found to be useful identifying DT104, the MDR lineage, as well as African invasive lineages [16]. MGT of *S*. Typhimurium was also shown to perform well in outbreak identification and source tracing [16].

Here, we describe an MGT scheme and public database for *S.* Enteritidis. By using MGT, we systematically examine the global and local epidemiology of *S.* Enteritidis over time, and evaluate the application of this scheme to the outbreak evaluation and the monitoring of multidrug resistant STs.

Results

Establishing the MGT system for *S.* Enteritidis

The core genome of *S.* Enteritidis was defined with 99% of coverage and 96% of identity (**Fig. S1)**, which included 3932 genes with 977 not found in the *Salmonella* core as well as 1054 *S.* Enteritidis core intergenic regions (**Fig. 1a**) (**Data Set S1, Tab 1**).

The first 8 schemes of *S.* Enteritidis MGT used the same 2955 core genes of *Salmonella* as the MGT of *S.* Typhimurium defined by Payne *et al*. [16]. The rationale of locus selection has been fully described previously [16]. MGT1 corresponded to the classic 7 gene MLST of *Salmonella* [12]. A total of 4986 loci were incorporated in the MGT9 scheme, including the 3932 *S.* Enteritidis core genes and 1054 core intergenic regions defined (**Fig. 1a**).

A total of 26,670 *S.* Enteritidis genomes with publicly available raw reads, collected from 26 source types in 86 countries during 1917 to 2018 (**Fig. S2**), were analysed using the MGT scheme. At each MGT level, each isolate is assigned an ST and a CC. A CC is a group of STs with a genetic distance threshold of one allele, as defined in the classic 7 gene MLST [12, 17]. The number of STs and CCs at each MGT level is shown in **Fig. 1b.** As the resolution of typing increased from MGT1 to MGT9, an increasing number of STs and CCs were assigned. At MGT2, the 26,670 isolates were subtyped into 252 STs and 23 CCs. In contrast, MGT9 divided the isolates into 20,153 different STs and 14,441 CCs. The MGT scheme of *S.* Enteritidis is available through a global public online database (https://mgtdb.unsw.edu.au/enteritidis/).

The global epidemiology of *S.* Enteritidis by MGT

Of the 26,670 isolates typed by MGT, 25,207 have country metadata. By the 7-gene MLST (or MGT1) scheme, ST11 was the dominant type representing 94.8% of the isolates, followed by ST183 of 1.6%. For the predominant ST11 isolates (or MGT1-ST11), we found that the optimal MGT level

to describe their global epidemiology was MGT4, based on the distribution of each ST in different continents (**Fig. S3**). At the MGT4 level, the 26,670 isolates were subtyped into 2,236 STs and 423 CCs. Among the 2,236 MGT4-STs, 163 STs were predominantly found in only one continent (defined as ≥70% from one continent). These 163 STs contained 20,341 of the 26,670 genomes (76.3%).

The distribution of STs varied between continents. MGT4-ST99, ST136, ST135, ST171 and ST163 were the most prevalent STs in North America (**Fig. 2**), while in Europe, MGT4-ST25, ST13, ST100 and ST31 were common. However, the North America and Europe *S.* Enteritidis sequences analysed were mostly obtained from the USA and UK, which represented 94.6% and 92.5% of isolates from these two continents. In Africa, MGT4-ST11 was more prevalent in West Africa and MGT4-ST16 in Central and Eastern Africa (**Fig. S4**). These STs reflect the previously reported African invasive infection lineages (West African clade and Central/Eastern African clade) [5]. Finally, MGT4-ST15 was a global ST which was observed in all continents.

These dominant STs can be further grouped into CCs which offered a more inclusive picture. Among the 423 MGT4-CCs at MGT4, 10 represented 94.1% of all the isolates. The top two CCs, MGT4-CC1 and MGT4-CC13 accounted for 88.0% of the isolates. MGT4-CC1 was prevalent in all six continents (**Fig. 2)**. MGT4-CC13 was more common in North America and Europe.

National epidemiology of *S.* Enteritidis by MGT

As the majority of the *S.* Enteritidis sequences analysed in this study were from the USA and UK, we compared distribution of the STs between these two countries. We used MGT4 and MGT5 levels to describe the data. A total of 39 MGT4 STs with more than 50 isolates represented 75.1% of the USA and UK isolates and each country had its own specific types (**Fig. 3a)**. MGT4-ST99, ST135 and ST170 were prevalent in the USA, whereas MGT4-ST15, ST25 and ST13 were more prevalent in the UK. MGT4-ST15 and ST25 were further subtyped into 34 MGT5-STs (each with more than 20

isolates), of which the majority were prevalent in the UK, except for MGT5-ST412 and ST387 that were more prevalent in the USA (**Fig. 3b)**.

To examine the relationship between MGT-STs, isolation source and location, we examined MGT4-STs of 4383 genome sequences from the USA which contained source and state metadata (**Fig. 4**). Of the 4383 genomes, 46.7% were from avian source while 43.1% were from humans. Six MGT4 STs (with more than 50 isolates each) were isolated from both human and avian sources including MGT4-ST99, ST135, ST136, ST160, ST198 and MGT4-ST25, while 3 STs including MGT4-ST15, ST163 and MGT4-ST170 were mainly from human sources. All of the human-only STs belonged to MTG4-CC1 whereas the mixed source STs were mostly of MGT4-CC13 origin (**Data Set S1, Tab 2**). STs belonging to MGT4-CC13 were significantly more prevalent in avian sources than STs of MGT4-CC1 ($P < 0.001$, OR = 45.9) (**Data Set S1, Tab 2**). *S.* Enteritidis isolate data from 48 states of the USA were available and ST frequencies were similar across the country (**Fig. 4**). In almost all states, MGT4-ST99 was the dominant type, followed by MGT4-ST135 and MGT4-ST136, all of which belonged to MGT4-CC13.

UK isolates contained collection year and month metadata. There were 8,818 human *S.* Enteritidis isolates collected from March 2014 to July 2018 in the UK. We chose MGT5-STs to describe the monthly variation of *S.* Enteritidis in the UK. By MGT5, variation in the prevalence of MGT5-STs in different years and months was observed (**Fig. 5**). There were 13 MGT5-STs with more than 100 isolates, representing 46.3% of the 8,818 isolates. The top 5 STs over the entire period were MGT5-ST1, ST29, ST79, ST15 and MGT5-ST33, representing 30% of the total isolates. These STs showed differing temporal patterns. MGT5-ST1, ST29, ST79 and ST15 were consistently observed in each month across these four years while MGT5-ST15 included isolates previously reported as part of an outbreak [18]. MGT5-ST156 first appeared in April of 2012 in the UK, increased in frequency in June and July of 2014, and became rare after 2014. MGT5-ST156 contained all isolates reported as

involved in an international outbreak in Europe [19] which explains the relatively short duration of this large ST. MGT5-ST423 was a dominant type from May to September of 2015, then became rare and was dominant again during the September to December of 2016.

Detection of potential large outbreak clusters of *S.* Enteritidis using MGT

MGT9 offered highest resolution for spread tracking and outbreak detection. By MGT9, there were 124 MGT9-STs including isolates from two or more countries each, indicating international spread (**Data Set S1, Tab 3**). To facilitate outbreak detection, we identified single linkage clusters of isolates using MGT9 allele differences with a range of cut-offs (0, 1, 2, 5 and 10 allelic differences). These clusters were named outbreak detection clusters (ODC) and can be used to analyse frequencies of closely related isolates at different cut-off levels for population studies and may also be used as dynamic thresholds to detect outbreaks [20].

In this study, we used ODC2 - clusters of isolates with no more than 2 allele differences at MGT9 level- to detect potential outbreak clusters and to assess whether some STs were more likely to cause large outbreak clusters, based on the number of ODC2 clusters and the total number of isolates in these clusters in different STs. Since the global data may be biased towards outbreak isolates that were preferentially sequenced, we used UK data from 2014 to 2018 which included all human isolates referred to public health authorities [21]. There were 17 ODC2 clusters of more than 50 isolates representing 1855 isolates in total. The majority of these ODC2 clusters belonged to 12 different MGT4-STs and therefore we used MGT4 level to perform comparison (**Data Set S1, Tab 4**). MGT4-ST15 was the dominant ST in UK and contained only one ODC2 cluster of 62 isolates while MGT4-ST25, which was the second dominant ST in UK, contained 5 large ODC clusters containing 602 isolates. Statistically MGT4 ST15 is significantly less likely to contain large ODC2 clusters than the other STs (OR = 0.1, P value < 0.001), while MGT4-ST25 is significantly more likely to contain larger ODC2 clusters than other STs (OR = 3.1, P value < 0.001). It is noteworthy

that by clonal complexes, all of the 6 top STs belonged to MGT4-CC13 and were positively associated with large scale outbreaks, including 3 previously reported large scale outbreaks in Europe [18, 19, 22, 23]. Indeed, MGT4-CC13 was significantly more likely to cause larger scale outbreaks than MGT4-CC1 (OR = 6.2, P < 0.001). These associations of STs and CCs with large outbreak clusters were also significant when ODC5 clusters was used. Notably the cluster sizes were larger but the trend remains the same (**Data Set S1, Tab 4 and Tab 5**).

Antibiotic resistance (AR) gene profiling of *S.* Enteritidis

As nearly all of the isolates (99.98%, 26,665/26,670) harboured the aac(6')-Iaa_1 gene for aminoglycoside resistance, this gene was excluded from the antibiotic resistance analysis. A total of 2505 isolates (9.4% of the total 26,670 isolates) were found to harbor antibiotic resistant genes excluding aac(6')-Iaa_1. The most frequent predicted class was tetracyclines (5.1%, 1,350/26,670), followed by beta-lactams (4.5%, 1,213/26,670) and aminoglycosides (2.8%, 741/2270). Among the 2505 isolates with AR genes, 40.4% (1011/2505) were predicted to be MDR, harbouring genes conferring resistance to 3 or more different antibiotic classes. A total of 59.6% (1494/2505) of the AR isolates harboured genes conferring resistance to 1 to 2 different antibiotic classes. Although the selection of isolates for genome sequencing in some continents may be biased, African isolates were found to have highest proportion of AR/MDR isolates, followed by Asia and Oceania (**Fig. S5a)**.

STs containing more than 10 isolates at different MGT levels were screened to identify MDR or AR associated STs (defined as >= 80% MDR or AR isolates). Eleven STs from varied levels were identified as MDR STs, representing 49% (659/1011) of the MDR isolates. These STs were mutually exclusive at different MGT levels. The top two STs, MGT3-ST15 and ST10, were the two invasive types that were prevalent in Africa. For MGT3-ST15, 99.2% of the isolates harboured resistance genes corresponding to as many as six different drug classes (**Table 1**). For MGT3-ST10, 86.5% of the isolates were MDR, and the antibiotic resistant patterns were similar to those of MGT3-ST15.

MGT3-ST10 and ST15 represented 96.3% and 90.2% of the two Africa endemic lineages (MGT3-CC10 and CC15, respectively) as mentioned below. Among the other 9 MDR STs, 8 belonged to MGT4-CC1, and 93.5% of the isolates (243/260) harboured genes conferring resistance to as many as 8 classes of antibiotic drugs. Only one MDR ST (MGT6-ST2698) belonged to MGT4-CC13 with 89.7% (35/39) of the isolates harboured genes conferring resistance to 4 drug classes. Based on available population sampling, MGT4-CC1 had significantly more isolates with MDR genes (5.6%, 474/8539) than MGT4-CC13 (0.7%, 107/14972) (Fisher exact test, P value < 0.001, OR = 12.4).

A total of 707 isolates belonging to 47 different STs from MGT3 to MGT7, harboured genes conferring resistance to one or two different antibiotic classes including aminoglycosides, beta-lactams, tetracyclines and quinolones (**Data Set S1, Tab 6**). Among the 47 STs, 34 (72%, 34/47) belonged to MGT4-CC1, representing 80% (564/707) of isolates, and 12 (26%, 12/47) STs belonged to MGT4-CC13 representing 18% (125/707) of the isolates. Again MGT4-CC1 had significantly more isolates with AR genes than MGT4-CC13 (Fisher exact test, P value < 0.001, OR = 8.4).

Plasmid specific genes were identified from the PlasmidFinder database [24] and IncQ1, IncN, IncI1, IncX1 plasmids were common in S. Enteritidis isolates with AR genes (**Fig. S5b**). In particular, plasmid type IncQ1 was present in MGT3-ST15, ST10 and ST30, which harboured MDR genes up to 6 drug classes (**Table 1**). Plasmid type IncI1 was common in MGT3-ST10 and ST161 and plasmid type IncX1 was more common in the STs of MGT4-CC1.

The population structure and evolution of major *S.* Enteritidis STs/CCs

The majority of the STs from different MGT levels analysed belonged to the two MGT4 CCs, MGT4-CC1 and CC13. To describe the global phylogenetic structure of *S*. Enteritidis and explore the phylogenetic relationship of the two lineages, 1506 representative isolates were selected using representatives of MGT6-STs to encompass the diversity of the serovar. A previous study had suggested that *S*. Enteritidis has three clades, A, B and C. Clade A and C appeared to diverge earlier

and were phylogenetically more distant to the global clade B than *S*. Gallinarum [11] (**Fig. 6a**). *S.* Enteritidis clade B and *S.* Gallinarum were sister clades. The vast majority of the genomes analysed in this study belonged to clade B.

Within clade B, we identified 10 main lineages, which were concordant with MGT-CCs from MGT1 to MGT4 (**Fig. 6b**). These lineages can be described at different MGT levels as shown in **Fig. 6b**. The lineages were consistent with the progressive division from lower to higher MGT levels grouped by CCs. MGT2-CC3 represent 98.8% of the Europe endemic MGT1-ST183 lineage [25] while MGT2 CC1 included almost all other isolates. At MGT3, 4 lineages named MGT3-CC10, CC15, CC18 and CC107 defined separate lineages, while MGT3-CC1 included several lineages represented by MGT4-CCs. MGT3-CC10 and CC15 represented 100% of the two previously reported Africa lineages associated with invasive infection [5]. MGT3-CC1 included 5 main MGT4-CCs, with two, MGT4-CC1 and MGT4-CC13, representing 88.0% of all 26,670 *S.* Enteritidis isolates (**Fig. 6b**). MGT4-CC30 and CC129, both of which were phylogenetically closer to MGT3-CC13, were two endemic lineages in Europe and North America, respectively (**Fig. 6b**). The population structure of *S.* Enteritidis defined by MGT, were generally in accordance with cgMLST HierCC HC100 (**Fig. S6**).

We performed BEAST analysis using the *S.* Enteritidis core genome to determine the evolutionary rate of the clade B *S.* Enteritidis which was thus estimated to be 1.9×10^{-7} substitution/site/year (95% CI of 1.6×10^{-7} to 2.3×10^{-7}), corresponding to 0.8 SNPs per genome per year for the core genome (95% CI of 0.6 to 0.9 SNPs). The mutation rate of MGT4-CC13 lineage core genome was estimated to be 2.5×10^{-7} substitution/site/year (95% CI of 2.3×10^{-7} to 2.7×10^{-7}) or 1.0 SNP per genome per year (95% CI of 0.9 to 1.1 SNPs). The mutation rate of MGT4-CC1 lineage core genome was 1.7×10^{-7} substitution/site/year (95% CI of 1.6×10^{-7} to 2.0×10^{-7}), or 0.7 SNP per genome per year

(95% CI of 0.6 to 0.8 SNPs**)**. The mutation rate of MGT4-CC13 was significantly faster (1.5 times) than that of MGT4-CC1.

The most recent common ancestor (MRCA) of the 9 lineages belonging to MGT2-CC1, was estimated to have existed in the 1460s (95% CI 1323 to 1580) (**Fig. S7**). The two global epidemic lineages MGT4-CC1 and CC13 are estimated to have diverged at around 1687 (95% CI 1608 to 1760). In 1869 (95% CI=1829-1900), MGT4-CC13 diverged into two sub-lineages of various MGT4-STs, with one more prevalent in North America than in other continents (labelled with red arrow) and the other sub-lineage more prevalent in Europe (labelled with blue arrow). We further estimated the population expansion of the two global epidemic lineages MGT4-CC1 and CC13 (**Fig. 7**). For MGT4-CC1, there were two big expansions around 1950 and 1970. For MGT4-CC13, the population gradually increased until a more rapid expansion around 1970.

Discussion

S. Enteritidis is one of the most common foodborne pathogens causing large scale national and international outbreaks and food recalls [3, 26]. Understanding the population structure and genomic epidemiology of *S.* Enteritidis is essential for its effective control and prevention. In this study, we developed a new genomic typing tool, MGT, for *S.* Enteritidis and developed a database for international applications. We used MGT to describe its local and global epidemiology, and its population structure using 26,670 publicly available genomes and associated metadata.

MGT for *S.* Enteritidis uncovers geographic, source and temporal epidemiological trends of *S.* Enteritidis within and between countries

A total of 26,670 isolates were successfully typed using the MGT scheme which allowed the examination of the national and global epidemiology of *S.* Enteritidis. The salient feature of the flexibility of the different levels of MGT has also been illustrated through this analysis. The lower

resolution levels of MGT were found to be able to effectively describe the geographic variation in different continents, countries or regions. Of these levels we showed that MGT4-STs best described the global epidemiology of *S.* Enteritidis at continent level with some STs distributed globally while others were more restricted. MGT4-ST15 was a global epidemic type which was prevalent in almost all continents. In contrast, MGT4-ST16 (also described by MGT3-ST15) was prevalent in Central/Eastern Africa and MGT4-ST11 (also described by MGT3-ST10) was prevalent in West Africa, which agreed with a previous study [5].

In USA, MGT4-CC13 and MGT-CC1 showed remarkable difference in their epidemiology. MGT4-CC13 STs including MGT4-ST99, ST136 and ST135 were the main cause of clinical infections and were common in poultry. The prevalent STs were generally similar in different states. In particular, MGT4-ST99, ST136 and ST135 were the dominant types in almost all states of the USA. The distribution and poultry association of these STs suggests that they may be responsible for several multistate outbreaks caused by *S.* Enteritidis contaminated eggs [1, 27]. Since poultry related products (i.e. eggs, chicken and turkey, especially eggs) are known to be the main source of *S.* Enteritidis infections or outbreaks in the USA [4], this isn't surprising, but demonstrates the utility of the MGT.

On the other hand, MGT4-CC1 including MGT4-ST171, ST15 and ST163, were less prevalent in the USA. Nevertheless, these STs were isolated from human infections but were very rare in poultry, although sampling bias may affect this conclusion as the poultry isolates in the dataset was not from systematic sampling of poultry sources. On the other hand, beef, sprouts, pork, nuts and seeds can also be contaminated by *S.* Enteritidis [4]. And these non-poultry related foods may be the source of the MGT4-CC1 caused infections. MGT4-CC13 was found to be significantly associated with poultry source, which is concordant with a previous study [10]. Furthers studies are required to explain the association of MGT4-CC13 but not MGT4-CC1 with poultry products. The lack of

potential source isolates within MGT4-CC1 STs highlights the need for comprehensive sequencing and epidemiological efforts across the food production chain.

Temporal variation of UK *S.* Enteritidis was depicted clearly by MGT. From 2014 onwards, all clinical *S.* Enteritidis were routinely sequenced in the UK [21]. Some MGT5-STs were found to occur for a few months and then disappeared. Such patterns of birth and death of STs may be related to outbreaks. For example, MGT5-ST156, which increased substantially during June and July in 2014 and became rare after 2015, was associated with a reported large scale outbreak [19]. Other STs persisted for long periods of time. For example, MGT5-ST1 and ST29, appeared to be endemic isolates which may be associated with local reservoirs. Therefore, the variation of MGT-STs across different seasons offered additional epidemiological signals for suspected outbreaks and endemic infections of *S.* Enteritidis. The stability of STs avoids the potential cluster merging issues of single linkage clustering based methods and ensures continuity for long-term surveillance [16].

MGT for *S.* Enteritidis facilitates outbreak detection, source tracing and evaluation of outbreak propensity

Whole genome sequencing offered the highest resolution for outbreak detection and source tracing. However, resolution of cgMLST heavily depends on the diversity of the species. Our previous study showed that for *S.* Typhimurium, serovar core genome offered higher resolution than species level core genome for outbreak investigation [16, 28]. We designed MGT9 for *S.* Enteritidis using 4986 loci, which is around 2000 more loci than *Salmonella* core genes [13, 16], thus increasing the resolution of subtyping.

To facilitate outbreak detection MGT9 STs were further grouped to ODCs. There is no agreed single cut-off for outbreak detection and dynamic thresholds have been suggested [20]. Here, we used ODC2, which has a two-allele difference cut-off, to identify potential outbreak clusters. Although the data does not allow us to confirm whether any of these ODC2 clusters were actual outbreaks, a

number of known outbreaks from other studies fell into ODC2 clusters [20, 22, 29]. A two-allele cut-off was selected because it is at the lower end of cut-offs in reported Enteritidis outbreaks [22, 29], which should limit the number of false positive outbreak calls. However, due to the variability in diversity of isolates from different outbreaks, a dynamic threshold for cut-offs would be more sensitive and specific to detect outbreaks [20]. Further work for *S.* Enteritidis is required to address this issue fully. Using the UK clinical isolates from 2014 to 2018, we found that the European and North American prevalent lineage MGT4-CC13 was significantly correlated with larger ODC2 clusters (>= 50 isolates each) than the global epidemic lineage of MGT4-CC1 (Chi-Square test, $P < 0.05$). Thus, MGT4-CC13 is more likely to cause large scale outbreaks than MGT4-CC1. In the past few years, several large scale outbreaks in Europe were due to the contaminated eggs [22, 23, 26]; these isolates all belonged to MGT4-CC13. In the USA, MGT4-CC13 was the dominant lineage in both human infections and poultry product contamination, while MGT4-CC1 was relatively rare in poultry. The industrialised poultry/eggs production and marketing would have facilitated the spread of MGT4-CC13 causing large scale outbreaks. Further studies are required to understand the underlying biological and environmental mechanisms facilitating these large MGT4-CC13 outbreaks.

ODCs offer a means for accurate detection of outbreaks at the highest genome resolution MGT9 [16]. Other studies used SNP address or cgMLST [18, 19, 22, 23] with SNP cut-offs or HierCC clustering for outbreak detection. All three methods use the same single linkage clustering method for outbreak detection. MGT offers further advantage that potential outbreak clusters can be precisely defined by a stable MGT-ST identifier rather than a cluster number [16].

MGT for *S.* Enteritidis improves precision of detection and global tracking of MDR clones

The rise of AMR in *Salmonella* is a serious public health concern [5]. The global spread of AMR can be mediated by lateral transfer of resistance genes as well as clonal spread of resistant strains. This

study systematically evaluated the presence of antibiotic resistant genes in the 26,670 genomes of *S.* Enteritidis, 9.4% of which harboured antibiotic resistant genes. Among the two global epidemic lineages, MGT4-CC1 was found to contain significantly more antibiotic resistant isolates than MGT4-CC13.

We further identified STs that were associated with MDR. Eleven STs (>= 10 isolates each, 659 isolates in total) of different MGT levels were associated with MDR. These STs from different MGT levels were mutually exclusive, emphasising the flexibility of MGT for precise identification of MDR clones. MGT3-ST15 and MGT3-ST10 were representative of the previously reported Africa invasive infection related lineages [5], harboured resistant genes of up to 6 different antibiotic classes. The other STs, which were mainly from Europe and North America, harbor AR genes to up to 8 different classes of drugs. MGT3-ST30 and MGT4-ST718, which belonged to MGT4-CC1 and were observed in Europe, had been reported to be MDR phenotypically [30]. Significantly, isolates from these two STs were mainly isolated from blood samples [30]. These MDR STs correlated with blood stream infections should be monitored closely as they can potentially be more invasive [5].

Because plasmids are known to play a key role in the acquisition of drug resistance genes [31, 32], we identified Inc plasmid groups that are likely present in the *S.* Enteritidis population. Several of these putative plasmids were MDR associated. MDR of African isolates in MGT3-ST10 and ST15 has been reported to be mediated by plasmids [5]. The West Africa lineage MGT3-ST10 isolates were reported to have IncI1 plasmids [33]. In this study, both IncI1 and IncQ1 plasmid were present in the Africa MGT3-ST10 and ST15 isolates as well as STs belonged to global epidemic lineage MGT4-CC1. IncQ1 plasmids, which are highly mobile and widely transferred among different genus of bacteria [34], are correlated with MDR of *Salmonella* [35, 36]. IncI1 plasmids are responsible for the cephalosporin resistance of *Salmonella* and *Escherichia* in several countries [37-39]. Moreover,

IncN and IncX1 plasmid were common in the MDR associated STs of MGT4-CC1 and CC13. The use of MGT could enhance the surveillance of drug resistance plasmids.

Defining the population structure of *S*. Enteritidis with increasing resolution and precision at different MGT levels

In this study, we defined the global population structure of *S*. Enteritidis using MGT-CCs. Three clades of *S*. Enteritidis were previously defined with clade A and C as outgroups to the predominant clade B [11], a classification supported by this study. We further identified 10 lineages within clade B, which can be represented by different CCs from MGT2 to MGT4 corresponding with time of divergence. Lower resolution level MGT CCs described old lineages well while higher resolution level MGT CCs described more recently derived lineages (**Fig. S7**). Thus, the epidemiological characteristics of a lineage can be identified using STs and/or CCs at appropriate MGT level. MGT2-CC3 represent the Europe endemic MGT1-ST183 lineage, within which MGT2-ST3 and MGT2-ST82 were able to identify previously defined phage types PT11 and PT66 respectively [25]. MGT3-CC10 included all of the reported West Africa lineage isolates, and MGT3-CC15 included all of the reported Central/Eastern Africa lineage isolates as well as Kenyan invasive *S*. Enteritidis isolates [5, 8]. MGT3-ST10 and MGT3-ST15 (represented 96.3% of MGT3-CC10 and 90.2% of MGT3-CC15 respectively), were MDR with similar resistance patterns. Most importantly, the two major global epidemic lineages are clearly described by MGT4-CC1 and MGT4-CC13. MGT4-CC1 corresponds the global epidemic clade in the study of Feasey *et al*., and MGT4-CC13 to the outlier clade of that study [5]. MGT4-CC1 and CC13 correspond to lineage III and lineage V in Deng *et al*'s study [10]. Thus, by using STs and/or CCs of different MGT levels the population structure of *S.* Enteritidis can be described clearly.

Interestingly, the two dominant lineages, MGT4-CC1 and CC13, showed different population expansion time and evolutionary dynamics. Bayesian evolutionary analysis revealed a core genome

mutation rate for *S.* Enteritidis of 1.9×10^{-7} substitution/site/year which is similar to the genome mutation rate of 2.2×10^{-7} substitution/site/year estimated by Deng *et al* [10]. In contrast, the mutation rate of MGT4-CC13 was 1.5 times faster than MGT4-CC1. Population expansion of MGT4-CC1 occurred with two peaks in the 1950s and 1970s, while MGT4-CC13 population increased steadily until the 1970s. The expansion around the 1970s may have been driven by the development of the modern industrialised poultry industry, including poultry farms or processing plants, that *S.* Enteritidis colonized. This is concordant with Deng *et al*'s speculation that the expansion of *S.* Enteritidis population was correlated with the poultry farm industry [10, 40].

The *S.* Enteritidis MGT scheme enables scalable nomenclature at different typing resolutions

The design of the *S.* Enteritidis MGT is based on the *S.* Typhimurium MGT scheme published previously [16]. The first eight levels (MGT1 to MGT8) of *S.* Enteritidis MGT scheme used the same loci as for *S.* Typhimurium. We defined the core genome of *S.* Enteritidis which had 3932 core genes and 1054 core intergenic regions and was used as MGT9. The subtyping resolution of MGT for *S.* Enteritidis substantially [13]. Our study suggests that the MGT levels 1 to 8 could be applied to all *Salmonella* serovars as a common scheme for the species, and only an additional serovar-specific MGT9 scheme needs to be designed for individual serovars that require the highest resolution for outbreak investigations.

The online database of the *S.* Enteritidis MGT scheme offers an open platform for global communication of genomic data and facilitates detection of international transmission and outbreaks of *S.* Enteritidis. The STs, especially at the middle resolution MGT level, were associated with different geographic regions, sources and MDR. The extension of MGT to *S.* Enteritidis was based on our previous study on *S.* Typhimurium [16]. Importantly, the stable characteristics of STs, which are based exact comparison, makes up for the main drawbacks of single-linkage clustering methods that the cluster types may change. Those STs enables long term epidemiological communication

between different laboratories. And the multi-level characteristics of MGT offer flexibility for temporal and spatial epidemiological analysis using an underlying stable nomenclature of STs from the finest resolution level. Additionally, CCs at each MGT level were able to cluster the STs with one allele difference and are concordant with phylogenetic lineages as shown in this study (**Fig. 6**).

Limitations of this study and challenges for public database

The epidemiological results generated in this study were based on the publicly submitted metadata and the accuracy of the results therefore depend on the accuracy of this data. In as many cases as possible metadata were verified from published sources. However, the possibility exists that incorrect metadata or repeated sequencing of the same isolates could influence the results. We showed the effect of the latter was minimal (**Supplementary Material**). However, good metadata is essential, and it would be useful for an internationally agreed minimum metadata for better epidemiological surveillance of *S.* Enteritidis. The online MGT system offered unprecedented power for monitoring *S.* Enteritidis across the globe. Good metadata further enhances that power.

Conclusions

In this study, we defined the core genome of *S.* Enteritidis and developed an MGT scheme with 9 levels. MGT9 offers the highest resolution and is suitable for outbreak detection whereas the lower levels (MGT1 to MGT8) are suitable for longer term epidemiology. At the MGT4 level, globally prevalent and regionally restricted STs were identified, which would facilitate the identification of the potential geographic origin of *S.* Enteritidis isolates. Specific source associated STs were identified, such as poultry associated MGT4-STs, which were common in human cases in the USA. At the MGT5 level, temporal variation of STs was observed in *S.* Enteritidis infections from the UK, which reveal both long lived endemic STs and the rapid expansion of new STs. Using MGT3 to MGT6, we identified MDR-associated STs to facilitate tracking of MDR spread. Additionally, certain plasmid types were found to be highly associated with the same MDR STs, suggesting

plasmid-based resistance acquisition. Furthermore, we evaluated the phylogenetic relationship of the STs by defining the population structure of *S*. Enteritidis. A total of 10 main lineages were defined in the globally distributed clade B *S*. Enteritidis. Of these, MGT4-CC1 and CC13 were the two dominant lineages, with significant differences in large outbreak frequency, antimicrobial resistance profiles and mutation rates. The online open database for *S*. Enteritidis MGT offers a unified platform for the international surveillance of *S*. Enteritidis. In summary, MGT for *S*. Enteritidis provides a flexible, high resolution and stable genome typing tool for long term and short long surveillance of *S*. Enteritidis infections.

Methods

S. Enteritidis sequences used in this study

Raw reads of 26,670 *S*. Enteritidis whole genome sequences (WGS) downloaded from the European Nucleotide Archive (ENA) passed the quality filtering criteria by Quast [41, 42] (**Data Set S1, Tab 7)**. The epidemiological metadata of those isolates were collated from NCBI and EnteroBase [13]. Trimmomatic v0.39 was used with default settings to trim raw reads [43]. The trimmed raw reads were then assembled using either SPAdes v3.13.0 for core genome definition or SKESA v2.3 for MGT typing [44, 45]. Quality assessment for the assemblies were performed using Quast v5.0.2 [41]. Thresholds of the assembly quality were in accordance with Robertson *et al* 's criteria [42]. Kraken v0.10.5 was used to double check any possible contamination in the assembled genomic sequences or problematic taxonomic identification [46]. Seqsero v1.0 and SISTR v1.0.2 were used to verify the serotype [47, 48].

Core gene definition of *S*. Enteritidis

To define the core genome of *S*. Enteritidis, a total of 1801 genomes were selected based on the ribosomal sequence type [13] and strain metadata. The trimmed raw reads were then assembled

using SPAdes v3.13.0 [44]. The assemblies were annoted with Prokka v1.12[49]. The core genes and core intergenic regions were defined with Roary v3.11.2 and Piggy v3.11.2, respectively (**Fig. S1**). The detailed methods for isolates selection and core genome identification were described in **Supplementary material**.

Establishment of MGT scheme for *S.* Enteritidis and MGT allele calling

The initial allele sequence of each locus used for MGT were extracted from the complete genome sequence of *S.* Enteritidis strain P125109 (Genbank accession NC_011294.1). The first 8 levels of MGT were identical to the MGT scheme for *S.* Typhimurium [16], in which MGT1 refers to the classical 7 gene MLST for the *Salmonella* genus [12]. Loci used for levels MGT2 to MGT8 were orthologous to those in the *S.* Typhimurium MGT with the exception of one gene which was removed at MGT8 because of a paralogue in *S.* Enteritidis. MGT9 contains all loci in MGT8 as well as the core genes and core intergenic regions of *S.* Enteritidis.

The raw read processing and genome assembly procedures were the same as above, except that SKESA v2.3 was used to assemble the raw reads and SISTR v1.0.2 was used for the molecular serotype identification [45, 48]. SKESA v2.3 offers better per base accuracy and speed than SPADES and was therefore used in the allele calling pipeline [45]. For the assemblies that passed quality control by the criteria of Robertson *et al.* [42], a custom script was used to assign alleles, STs and CCs to those isolates at levels from MGT2 to MGT9 [16]. Allele, ST, CC and outbreak detection cluster (ODC) assignments were performed as previously described [16]. Briefly, if one or more allele difference was observed, a new ST was assigned to the isolate. The same strategy was also used for assigning a new CC, if at least two allele differences were observed comparing to any of STs in existing CCs in the database [16]. Outbreak detection cluster (ODC) is an extension of the CC clustering method, but it was performed only at MGT9 level [16]. For example, any two MGT9 STs

with no more than 2, 5, or 10 allele differences were assigned the same ODC2, ODC5 or ODC10 type [16].

Geographic distribution and temporal variation analysis

We first determined which MGT level would be the most useful in describing the global epidemiology of *S.* Enteritidis by summarizing STs that contained at least 10 isolates for their distribution across continents. STs where more than 70% of isolates were from a single continent were identified at each level of MGT (**Fig. S3a**). MGT4 STs matching this criterion had the highest percentage of all the levels at 76.3% of the 26,670 isolates, thus MGT4 was chosen to describe the continental distribution of *S*. Enteritidis. The USA states variation map and the UK monthly variation column chart were produced using Tableau v2019.2 [50].

Antibiotic resistance gene and plasmid specific gene identification

Abricate (https://github.com/tseemann/abricate) was used for identification of antibiotic resistance genes with ResFinder v3.0 database [51], and plasmid specific genes with PlasmidFinder v2.1 database [24]. The cut-off of the presence of a gene was set as both identity and coverage >= 90% [51]. At each MGT level, STs (>=10 isolates each) with more than 80% of isolates harbouring antibiotic resistance genes, were collected from MGT2 to MGT9 and were defined as high-antibiotic-resistant STs. The STs at each level of MDR were non-redundant meaning that no isolate was counted more than once.

Phylogenetic analysis

A phylogenetic tree was constructed using Parsnp v1.2, which called core-genome SNPs [52, 53]. And potential recombinant SNPs were removed using both Gubbins v2.0.0 and Recdetect v6.0 [54, 55]. Beast v1.10.4 was then used to estimate the mutation rate based on mutational SNPs [56]. A total of 24 combinations of clock and population size models were evaluated with the MCMC chain

of 100 million states. Tracer v1.7.1 was used to identify the optimal model and to estimate population expansion over time [57]. The detailed methods for phylogenetic analysis were described in **Supplementary material.**

Evaluation of the effect of repeat sequencing on the dataset for epidemiological analysis

To evaluate any bias that may be caused by resequencing of the same isolate, we identified all isolates of the same ST based on MGT9 and same metadata based on collection country, collection year and month, and source type. Such isolates were conservatively treated as repeat sequencing of the same isolate and the "duplicates" were removed from the dataset. We re-analysed the reduced dataset and compared against the original dataset with the Kendall's tau value calculated [58]. The detailed results were shown in **Supplementary material** and **Data Set S2, Tab 1 to Tab 6**.

Authors' contributions

L.L., M.P., and R.L. designed the study. L.L. and M.P. set up the database. S.K and M.P. set up the website, M.P., R.L., D.H. L.C. S.O., Q.W., M.T., V.S. and S.K. provided critical analysis and discussions, L.L. wrote the first draft and all authors contributed to the final manuscript.

Funding information

This work was funded by a project grant from the National Health and Medical Research Council of Australia (grant number 1129713). Lijuan Luo was supported by a UNSW scholarship (University International Postgraduate Award). The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

Acknowledgements

The authors thank Duncan Smith and Robin Heron from UNSW Research Technology Services for technical assistance.

Conflicts of interest

The authors declare that there are no conflicts of interest.

References

1. **Dewey-Mattia D, Manikonda K, Hall AJ, Wise ME, Crowe SJ**. Surveillance for Foodborne Disease Outbreaks - United States, 2009-2015. *MMWR Surveill Summ* 2018;67(10):1-11.
2. Salmonellosis - Annual Epidemiological Report for 2017 [database on the Internet]2020. Available from: https://www.ecdc.europa.eu/en/publications-data.
3. **Pijnacker R, Dallman TJ, Tijsma ASL, Hawkins G, Larkin L et al.** An international outbreak of *Salmonella enterica* serotype Enteritidis linked to eggs from Poland: a microbiological and epidemiological study. *Lancet Infect Dis* 2019;19(7):778-786.
4. **Snyder TR, Boktor SW, M'Ikanatha N M**. Salmonellosis Outbreaks by Food Vehicle, Serotype, Season, and Geographical Location, United States, 1998 to 2015. *J Food Prot* 2019;82(7):1191-1199.
5. **Feasey NA, Hadfield J, Keddy KH, Dallman TJ, Jacobs J et al.** Distinct *Salmonella* Enteritidis lineages associated with enterocolitis in high-income settings and invasive disease in low-income settings. *Nat Genet*, Article 2016;48(10):1211-1217.
6. **Mohan A, Munusamy C, Tan YC, Muthuvelu S, Hashim R et al.** Invasive *Salmonella* infections among children in Bintulu, Sarawak, Malaysian Borneo: a 6-year retrospective review. *BMC Infect Dis* 2019;19(1):330.
7. **Thung TY, Mahyudin NA, Basri DF, Wan Mohamed Radzi CW, Nakaguchi Y et al.** Prevalence and antibiotic resistance of *Salmonella* Enteritidis and *Salmonella* Typhimurium in raw chicken meat at retail markets in Malaysia. *Poult Sci* 2016;95(8):1888-1893.

8. **Akullian A, Montgomery JM, John-Stewart G, Miller SI, Hayden HS et al.** Multi-drug resistant non-typhoidal *Salmonella* associated with invasive disease in western Kenya. *PLoS Negl Trop Dis* 2018;12(1):e0006156.

9. **Parn T, Dahl V, Lienemann T, Perevoscikovs J, De Jong B**. Multi-country outbreak of *Salmonella* enteritidis infection linked to the international ice hockey tournament. *Epidemiol Infect* 2017;145(11):2221-2230.

10. **Deng X, Desai PT, den Bakker HC, Mikoleit M, Tolar B et al.** Genomic epidemiology of *Salmonella enterica* serotype Enteritidis based on population structure of prevalent lineages. *Emerg Infect Dis* 2014;20(9):1481-1489.

11. **Graham RMA, Hiley L, Rathnayake IU, Jennison AV**. Comparative genomics identifies distinct lineages of *S*. Enteritidis from Queensland, Australia. *PLoS One* 2018;13(1):e0191042.

12. **Achtman M, Wain J, Weill FX, Nair S, Zhou Z et al.** Multilocus sequence typing as a replacement for serotyping in *Salmonella enterica*. *PLoS Pathog* 2012;8(6):e1002776.

13. **Alikhan NF, Zhou Z, Sergeant MJ, Achtman M**. A genomic overview of the population structure of *Salmonella*. *PLoS Genet* 2018;14(4):e1007261.

14. **Zhou Z, Alikhan N-F, Mohamed K, Fan Y, Agama Study G et al.** The EnteroBase user's guide, with case studies on *Salmonella* transmissions, *Yersinia* pestis phylogeny, and *Escherichia* core genomic diversity. *Genome Res* 2020;30(1):138-152.

15. **Ashton P, Nair S, Peters T, Tewolde R, Day M et al.** Revolutionising Public Health Reference Microbiology using Whole Genome Sequencing: *Salmonella* as an exemplar. *bioRxiv* 2015:033225.

16. **Payne M, Kaur S, Wang Q, Hennessy D, Luo L et al.** Multilevel genome typing: genomics-guided scalable resolution typing of microbial pathogens. *Euro Surveill* 2020;25(20):1900519.

17. **Feil EJ**. Small change: keeping pace with microevolution. *Nat Rev Microbiol* 2004;2(6):483-495.

18. **Kanagarajah S, Waldram A, Dolan G, Jenkins C, Ashton PM et al.** Whole genome sequencing reveals an outbreak of *Salmonella* Enteritidis associated with reptile feeder mice in the United Kingdom, 2012-2015. *Food Microbiol* 2018;71:32-38.

19. **Hormansdorfer S, Messelhausser U, Rampp A, Schonberger K, Dallman T et al.** Re-evaluation of a 2014 multi-country European outbreak of *Salmonella* Enteritidis phage type 14b using recent epidemiological and molecular data. *Euro Surveill* 2017;22(50):17-00196.

20. **Payne M, Octavia S, Luu LDW, Sotomayor-Castillo C, Wang Q et al.** Enhancing genomics-based outbreak detection of endemic *Salmonella enterica* serovar Typhimurium using dynamic thresholds. *Microb Genom* 2019.

21. **Chattaway MA, Dallman TJ, Larkin L, Nair S, McCormick J et al.** The Transformation of Reference Microbiology Methods and Surveillance for *Salmonella* With the Use of Whole Genome Sequencing in England and Wales. *Frontiers in public health*, Original Research 2019;7(317):317.

22. **Inns T, Lane C, Peters T, Dallman T, Chatt C et al.** A multi-country *Salmonella* Enteritidis phage type 14b outbreak associated with eggs from a German producer: 'near real-time' application of whole genome sequencing and food chain investigations, United Kingdom, May to September 2014. *Euro Surveill* 2015;20(16):21098.

23. **Inns T, Ashton P, Herrera-Leon S, Lighthill J, Foulkes S et al.** Prospective use of whole genome sequencing (WGS) detected a multi-country outbreak of *Salmonella* Enteritidis. *Epidemiology & Infection* 2017;145(2):289-298.

24. **Carattoli A, Zankari E, Garcia-Fernandez A, Voldby Larsen M, Lund O et al.** In silico detection and typing of plasmids using PlasmidFinder and plasmid multilocus sequence typing. *Antimicrob Agents Chemother* 2014;58(7):3895-3903.

25. **Lawson B, Franklinos LHV, Rodriguez-Ramos Fernandez J, Wend-Hansen C, Nair S et al.** *Salmonella* Enteritidis ST183: emerging and endemic biotypes affecting western European hedgehogs (Erinaceus europaeus) and people in Great Britain. *Sci Rep* 2018;8(1):2449.

26. **Dallman T, Inns T, Jombart T, Ashton P, Loman N et al.** Phylogenetic structure of European *Salmonella* Enteritidis outbreak correlates with national and international egg distribution network. *Microb Genom* 2016;2(8):e000070.

27. **Nichols M, Stevenson L, Whitlock L, Pabilonia K, Robyn M et al.** Preventing Human *Salmonella* Infections Resulting from Live Poultry Contact through Interventions at Retail Stores. *J Agric Saf Health* 2018;24(3):155-166.

28. **Fu S, Octavia S, Tanaka MM, Sintchenko V, Lan R**. Defining the Core Genome of *Salmonella enterica* Serovar Typhimurium for Genomic Surveillance and Epidemiological Typing. *J Clin Microbiol* 2015;53(8):2530-2538.

29. **Taylor AJ, Lappi V, Wolfgang WJ, Lapierre P, Palumbo MJ et al.** Characterization of Foodborne Outbreaks of Salmonella enterica Serovar Enteritidis with Whole-Genome Sequencing Single Nucleotide Polymorphism-Based Analysis for Surveillance and Outbreak Detection. *Journal of clinical microbiology* 2015;53(10):3334-3340.

30. **Vidovic S, An R, Rendahl A**. Molecular and Physiological Characterization of Fluoroquinolone-Highly Resistant *Salmonella* Enteritidis Strains. *Front Microbiol* 2019;10:729.

31. **Kaldhone PR, Carlton A, Aljahdali N, Khajanchi BK, Sanad YM et al.** Evaluation of Incompatibility Group I1 (IncI1) Plasmid-Containing *Salmonella enterica* and Assessment of the Plasmids in Bacteriocin Production and Biofilm Development. *Front Vet Sci* 2019;6(298):298.

32. **Zhou X, Li M, Xu L, Shi C, Shi X**. Characterization of Antibiotic Resistance Genes, Plasmids, Biofilm Formation, and In Vitro Invasion Capacity of *Salmonella* Enteritidis Isolates from Children with Gastroenteritis. *Microb Drug Resist* 2019;25(8):1191-1198.

33. **Aldrich C, Hartman H, Feasey N, Chattaway MA, Dekker D et al.** Emergence of phylogenetically diverse and fluoroquinolone resistant *Salmonella* Enteritidis as a cause of invasive nontyphoidal *Salmonella* disease in Ghana. *PLoS Negl Trop Dis*, Research Support, Non-U.S. Gov't 2019;13(6):e0007485.

34. **Smalla K, Heuer H, Gotz A, Niemeyer D, Krogerrecklenfort E et al.** Exogenous isolation of antibiotic resistance plasmids from piggery manure slurries reveals a high prevalence and diversity of IncQ-like plasmids. *Appl Environ Microbiol* 2000;66(11):4854-4862.

35. **Castellanos LR, van der Graaf-van Bloois L, Donado-Godoy P, Leon M, Clavijo V et al.** Genomic Characterization of Extended-Spectrum Cephalosporin-Resistant *Salmonella enterica* in the Colombian Poultry Chain. *Front Microbiol* 2018;9:2431.

36. **Mastrorilli E, Pietrucci D, Barco L, Ammendola S, Petrin S et al.** A Comparative Genomic Analysis Provides Novel Insights Into the Ecological Success of the Monophasic *Salmonella* Serovar 4,[5],12:i. *Front Microbiol* 2018;9:715.

37. **Wong MH, Kan B, Chan EW, Yan M, Chen S**. IncI1 Plasmids Carrying Various blaCTX-M Genes Contribute to Ceftriaxone Resistance in *Salmonella enterica* Serovar Enteritidis in China. *Antimicrob Agents Chemother* 2016;60(2):982-989.

38. **Garcia-Fernandez A, Chiaretto G, Bertini A, Villa L, Fortini D et al.** Multilocus sequence typing of IncI1 plasmids carrying extended-spectrum beta-lactamases in *Escherichia coli* and *Salmonella* of human and animal origin. *J Antimicrob Chemother* 2008;61(6):1229-1233.

39. **Kameyama M, Chuma T, Yokoi T, Yabata J, Tominaga K et al.** Emergence of *Salmonella enterica* serovar infantis harboring IncI1 plasmid with bla(CTX-M-14) in a broiler farm in Japan. *J Vet Med Sci* 2012;74(9):1213-1216.

40. **Deng X, Shariat N, Driebe EM, Roe CC, Tolar B et al.** Comparative analysis of subtyping methods against a whole-genome-sequencing standard for *Salmonella enterica* serotype Enteritidis. *J Clin Microbiol* 2015;53(1):212-218.

41. **Gurevich A, Saveliev V, Vyahhi N, Tesler G**. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* 2013;29(8):1072-1075.

42. **Robertson J, Yoshida C, Kruczkiewicz P, Nadon C, Nichani A et al.** Comprehensive assessment of the quality of *Salmonella* whole genome sequence data available in public sequence databases using the *Salmonella* in silico Typing Resource (SISTR). *Microbial genomics* 2018;4(2).

43. **Bolger AM, Lohse M, Usadel B**. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 2014;30(15):2114-2120.

44. **Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M et al.** SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* 2012;19(5):455-477.

45. **Souvorov A, Agarwala R, Lipman DJ**. SKESA: strategic k-mer extension for scrupulous assemblies. *Genome Biol* 2018;19(1):153.

46. **Wood DE, Salzberg SL**. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol* 2014;15(3):R46.

47. **Zhang S, Yin Y, Jones MB, Zhang Z, Deatherage Kaiser BL et al.** *Salmonella* serotype determination utilizing high-throughput genome sequencing data. *J Clin Microbiol*, 10.1128/JCM.00323-15 2015;53(5):1685-1692.

48. **Yoshida CE, Kruczkiewicz P, Laing CR, Lingohr EJ, Gannon VP et al.** The *Salmonella* In Silico Typing Resource (SISTR): An Open Web-Accessible Tool for Rapidly Typing and Subtyping Draft *Salmonella* Genome Assemblies. *PLoS One* 2016;11(1):e0147101.

49. **Seemann T**. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 2014;30(14):2068-2069.

50. Tableau (version. 9.1). *J Med Libr Assoc* 2016;104(2):182-183.

51. **Zankari E, Hasman H, Cosentino S, Vestergaard M, Rasmussen S et al.** Identification of acquired antimicrobial resistance genes. *J Antimicrob Chemother* 2012;67(11):2640-2644.

52. **Treangen TJ, Ondov BD, Koren S, Phillippy AM**. The Harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes. *Genome Biol* 2014;15(11):524.

53. **Price MN, Dehal PS, Arkin AP**. FastTree 2--approximately maximum-likelihood trees for large alignments. *PloS one* 2010;5(3):e9490-e9490.

54. **Croucher NJ, Page AJ, Connor TR, Delaney AJ, Keane JA et al.** Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Res* 2015;43(3):e15.

55. **Hu D, Liu B, Wang L, Reeves PR**. Living Trees: High-Quality Reproducible and Reusable Construction of Bacterial Phylogenetic Trees. *Molecular biology and evolution* 2020;37(2):563-575.

56. **Suchard MA, Lemey P, Baele G, Ayres DL, Drummond AJ et al.** Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evol* 2018;4(1):vey016.

57. **Rambaut A, Drummond AJ, Xie D, Baele G, Suchard MA**. Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7. *Syst Biol* 2018;67(5):901-904.

58. **Kendall MG**. A new measure of rank correlation. *Biometrika* 1938;30(1/2):81-93.

Table 1. Antibimicrobial drug class and plasmid class of the MDR associated STs from differnet MGT levels.

				% of the MDR isolates							% of isolates with plasmid									
MGT-ST[a]	**MGT4-CCs**	**Total No. of isolates**	**No. of isolates with MDR (%)**	**Aminoglycoside**	**Beta-lactam**	**Sulphonamide**	**Tetracycline**	**Phenicol**	**Trimethoprim**	**Macrolide**	**Col156_1**	**ColRNAI_1**	**ColpVC_1**	**IncI1_1_Alpha**	**IncI2_1_Delta**	**IncN_1**	**IncQ1_1**	**IncX1_1**	**IncX1_4**	**Continent[d]**
MGT3-ST15	CC16[c]	261	259 (99%)	98%	74%	97%	89%	95%	96%			1%		14%	1%		95%			Africa
MGT3-ST10	CC11[c]	141	122 (87%)	74%	84%	86%	82%	77%	82%			1%		95%	1%	4%	71%			Africa
MGT3-ST30	CC1	10	10 (100%)	100%		100%	100%	100%	100%								80%			EU
MGT3-ST161[b]	CC1	79	29 (37%)	30%	84%	35%	84%	32%	3%	3%	1%		100%	86%				1%		EU
MGT3-ST9	CC1	70	66 (94%)	93%	80%	91%	80%	1%	1%	1%	1%			3%					99%	EU/NA
MGT3-ST301	CC1	11	11 (100%)	100%	100%				100%									100%		NA
MGT4-ST354	CC1	39	39 (100%)	100%	100%	100%	62%				3%			5%					100%	NA/EU
MGT4-ST54	CC1	29	28 (97%)	90%	41%	90%	86%												97%	EU/NA/OC
MGT5-ST86	CC1	12	12 (100%)	100%	100%	100%								8%					100%	NA/EU
MGT5-ST682	CC1	10	10 (100%)	100%		100%	100%	90%						30%		100%		100%		NA
MGT6-ST2698	CC13	39	35 (90%)	90%	90%	90%	90%									90%		100%		NA
Total		701	621 (89%)																	

[a]MGT-STs with more than 80% of the isolated harboring AR genes to 3 or more drug types were identified. Isolates belonging to the MDR associated MGT-STs didn't overlap. Note aac(6')-Iaa_1 gene for aminoglycoside resistance, which was present in almost all isolates, was excluded for the AR analysis.

[b]MGT3-CC161 include 29 isolates with MDR genes (to 5 drug types) and 38 isolates with AR genes to 2 different drug types.
[c]MGT4-CC16 and CC11 belonged to MGT3-CC15 and CC10, which were representative of the two African lineages.
[d]The main continents that the MDR related STs were observed: EU refers to Europe, NA to North America and OC to Oceania.

Figure legends:

Fig. 1. Makeup and assignments of each *S.* Enteritidis MGT level. a. Number of loci included in each MGT level. The first eight levels are composed of *Salmonella* species core genes, which are orthologous to those of the *S.* Typhimurium MGT scheme, except for one gene at MGT8 which was excluded due to duplication in *S.* Enteritidis. The MGT9 scheme includes core genes of S. Enteritidis (those not belonging to the *Salmonella* core coloured in yellow) and core intergenic regions (coloured in green). The number behind or within each bar refers to the number of loci included. **b.** The number of sequence types (ST) and clonal complexes (CC) assigned at each MGT level among the 26,670 genomes analysed. CC includes STs of no more than 1 allele difference. The numbers refer to the number of STs or CCs types assigned at each level, which were also indicated by colour. As MGT9 included 4986 loci with highest subtyping resolution, the 26,670 isolates were subtyped into 20,153 different ST types and 14,441 CC clustering types at MGT9.

Fig. 2. 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. 3. USA and UK *S.* Enteritidis isolates can be distinguished by STs at MGT4 and MGT5. a. At MGT4, there were 39 STs (50 or more isolates each) representing 75.1% of the UK and USA

isolates (22191 in total). MGT4-ST99, ST136 and ST135 were prevalent in the USA, MGT4-ST15, ST25 and ST13 were prevalent in the UK, while MGT4-ST15 and ST25 were mixed. **b.** MGT4-ST15 and ST25 can be subtyped into multiple STs at MGT5, the majority of which were prevalent in the UK, except for MGT5-ST412 and ST387 in the USA.

Fig. 4. MGT4 ST proportions across states in the USA is similar but only a subset of STs are associated with avian hosts. For the USA *S.* Enteritidis isolates with metadata (4,383 isolates), a total of 3935 genomes were from either avian or human source, 95.0% (3738/3935) of which belonged to MGT4-CC1 and CC13. The MGT4-STs belonged MGT4-CC1 and CC13 of either avian or human source were shown in different states of the USA. Each pie chart illustrated the MGT-ST types and the size of the main STs represented with different colours. The size of the pie indicated the total number of isolates in each state. While CC13 appears to originate in birds before moving to humans, the reservoir of CC1 is not in birds. Origins for CC1 STs were also not observed in any other source type. The maps were created with Tableau v2019.2.

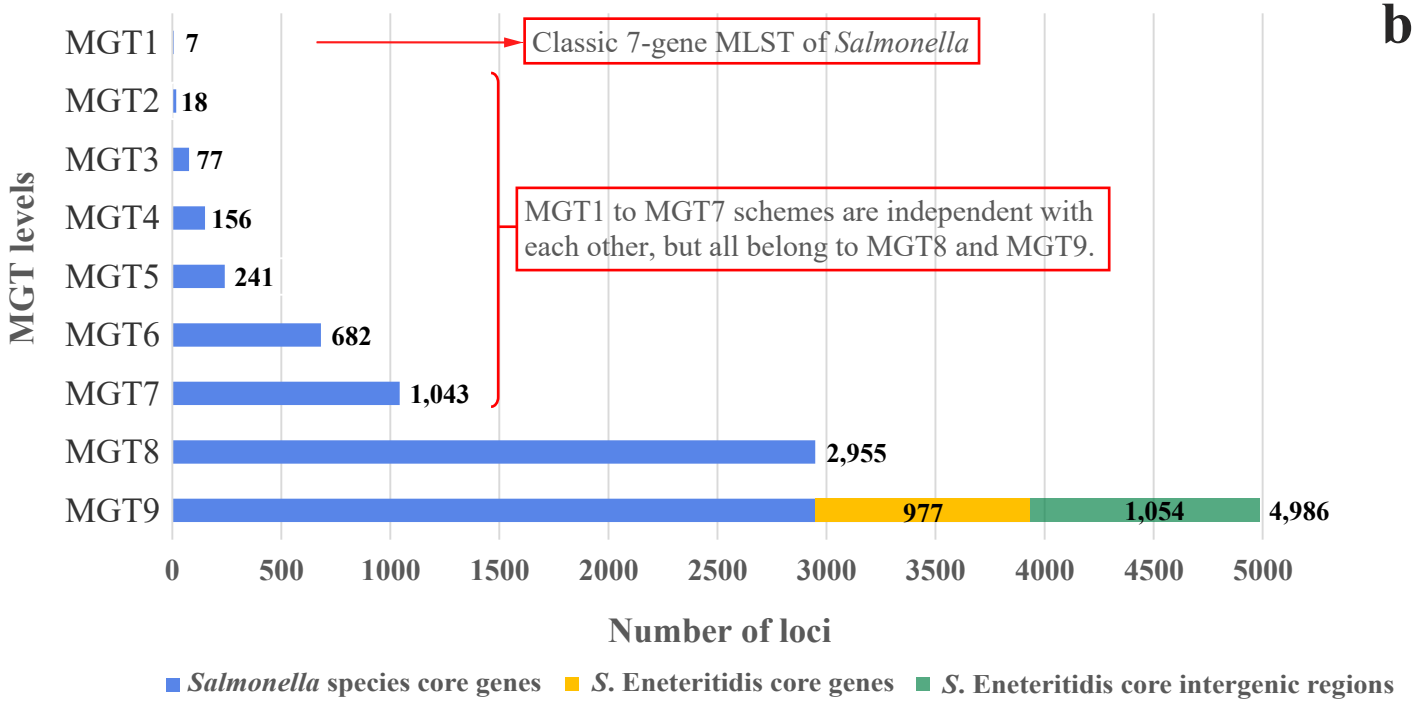
Fig. 5. MGT5 STs allow identification of temporal patterns in strain diversity in the UK. There were 13 MGT5-STs of more than 100 isolates shown with different colours. MGT5-ST156 in red was observed predominantly from June to August during 2014, and disappeared after 2014. MGT5-ST423 was a dominant type from May to September 2015, and became dominant again during the September to December of 2016.

Fig. 6. Global phylogenetic structure of *S.* Enteritidis. A total of 1508 genomes were randomly sampled from each ST in MGT6. A phylogenetic tree was constructed by ParSNP v1.2, which called core-genome SNPs extracted from alleles. The tree scales indicated the 0.01 substitutions per locus. The number on internal branches represented the percentage of bootstrap support. **a.** Three clades

were identified which were concordant with previous findings: The predominant clade, highlighted in yellow, is the global epidemic clade which was phylogenetically closer to *S.* Gallinarum than the other two clades in purple and green background which are limited in Oceania 9. **b**. The main, yellow highlighted clade was expanded into a more detailed phylogeny with *S.* Gallinarium as an outgroup. 10 main lineages were defined where each lineage had more than 10 isolates. STs of the 7 gene MLST (or MGT1) and CCs from MGT2 to MGT4 were shown in different colours for each of the 10 lineages. The pie charts for each lineage represent the proportion of the isolates belonged to the same MGT4-CC shown on the right side.

Fig. 7. Skyline population size estimation of MGT4-CC1 and CC13. Bayesian skyline model and strict clock were estimated to be the optimal combination. The vertical axis column indicates the predicted log scale effective population size and the horizontal axis shows the predicted time. The blue line represents the median posterior estimate of the effective population size, and the blue area shows the upper and lower bounds of the 95% highest posterior density interval. **a.** Skyline population expansion of MGT4-CC1. Around 1950 and 1970, there were two big population expansions within MGT4-CC1. **b.** Skyline population expansion of MGT4-CC13. The population of MGT4-CC13 gradually expanding until around 1970, when there was also a slight expansion.

a

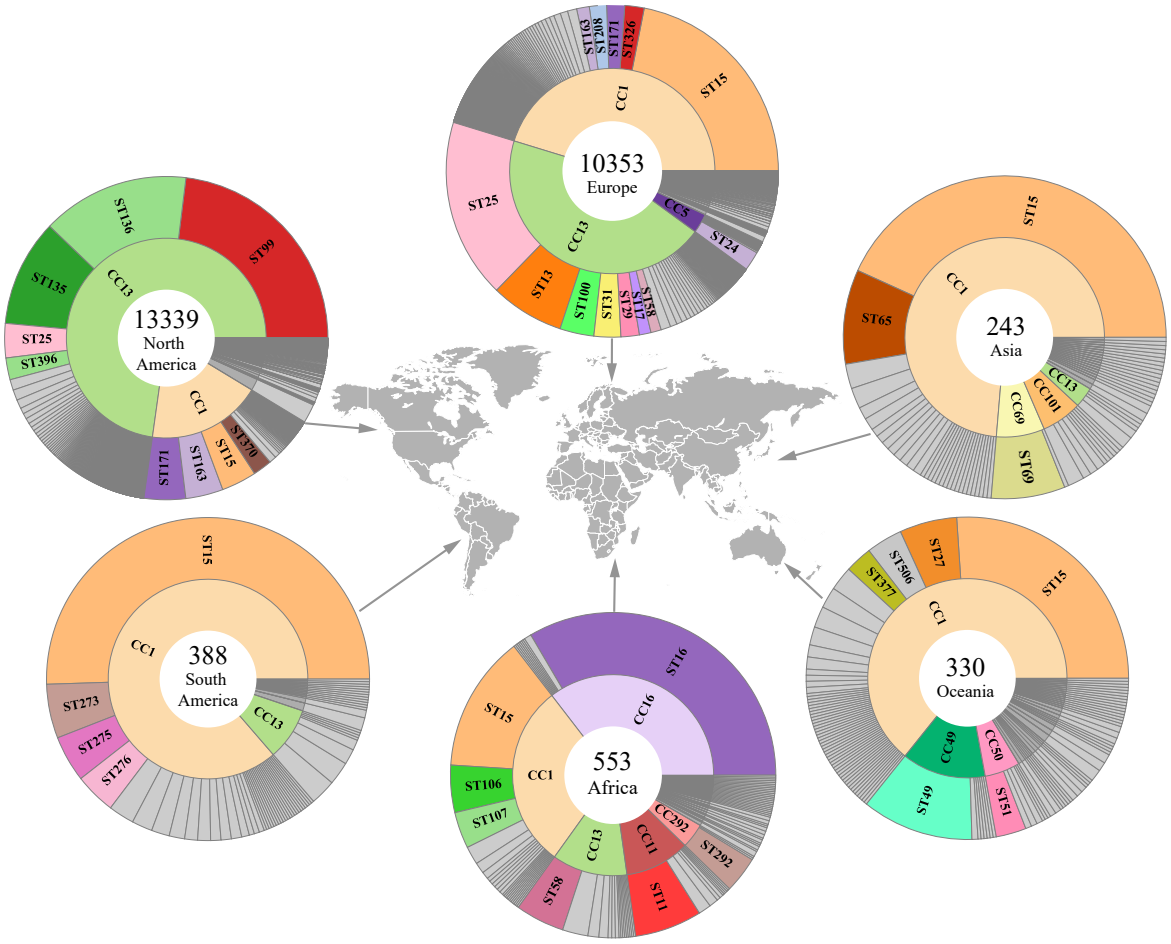


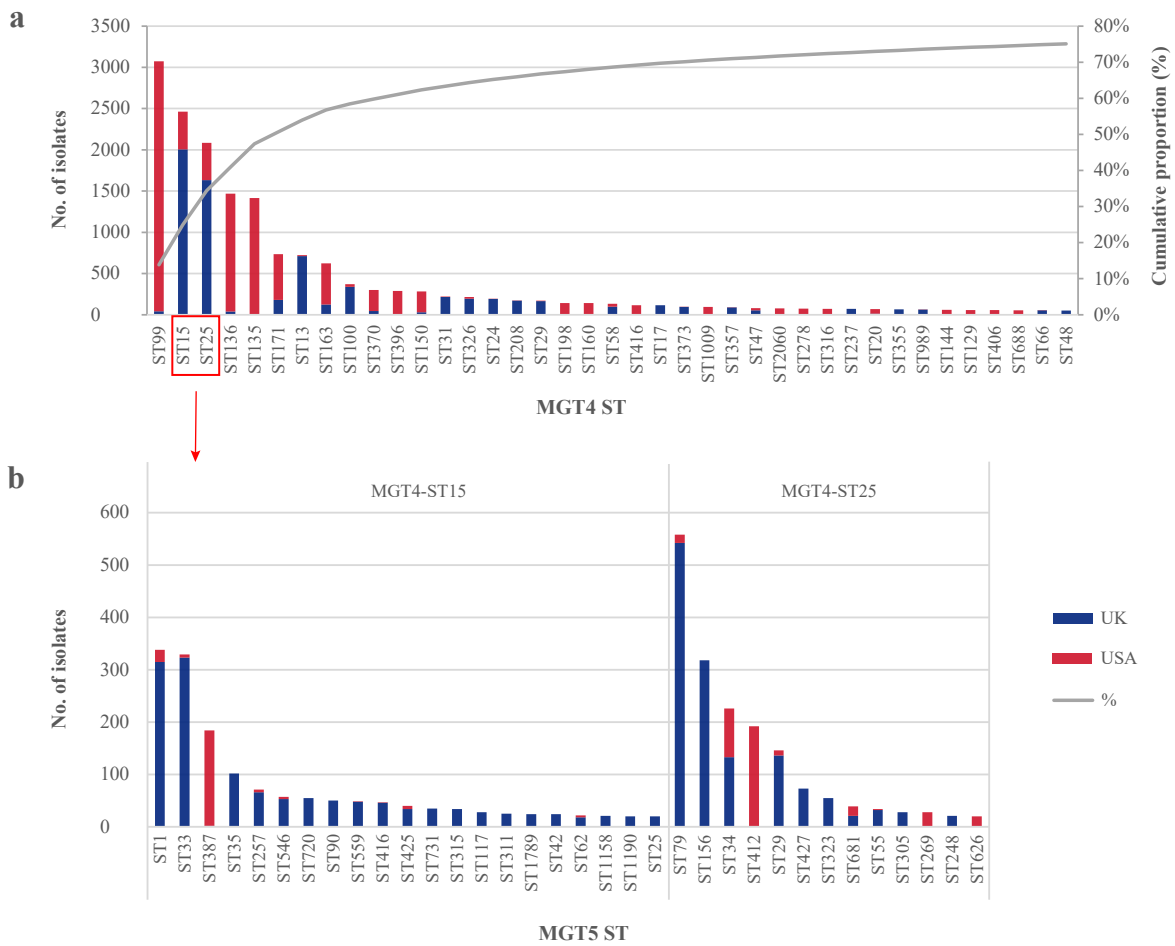
b

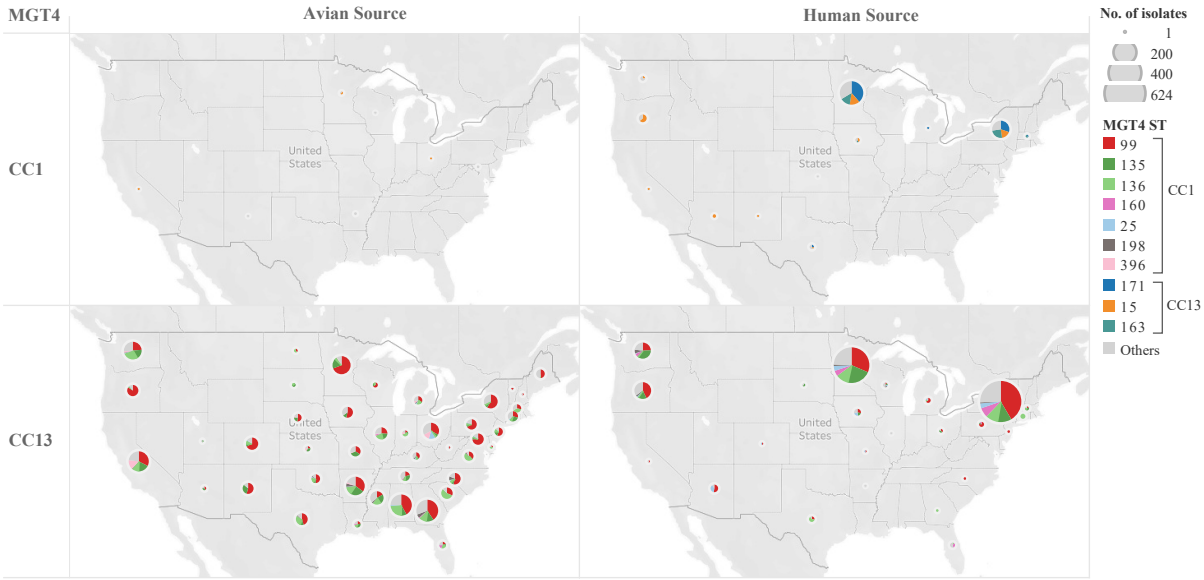
MGT	No. of STs	No. of CCs
MGT2	253	23
MGT3	1,222	179
MGT4	2,236	423
MGT5	4,128	1,017
MGT6	7,262	2,838
MGT7	11,107	5,680
MGT8	17,036	11,223
MGT9	20,153	14,441

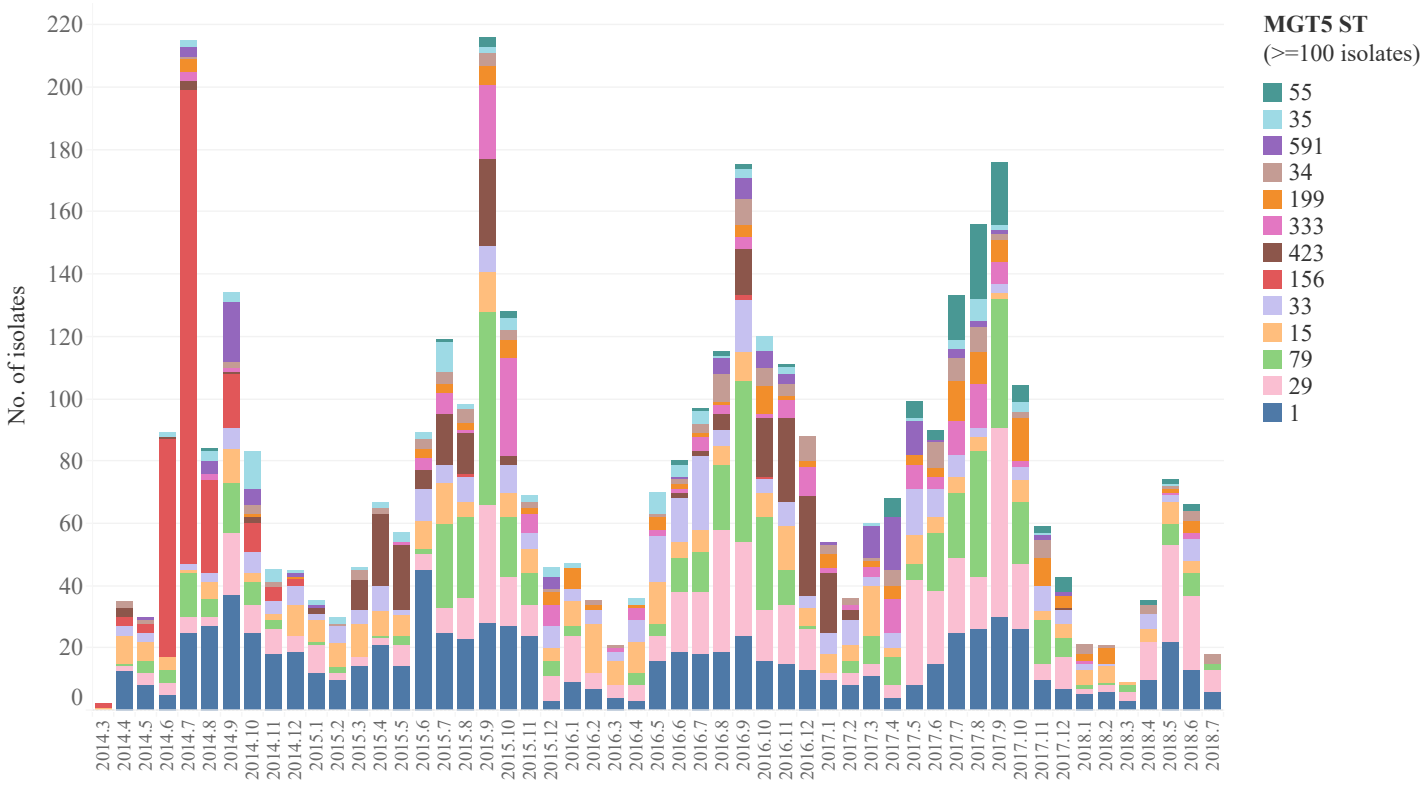
No. of types

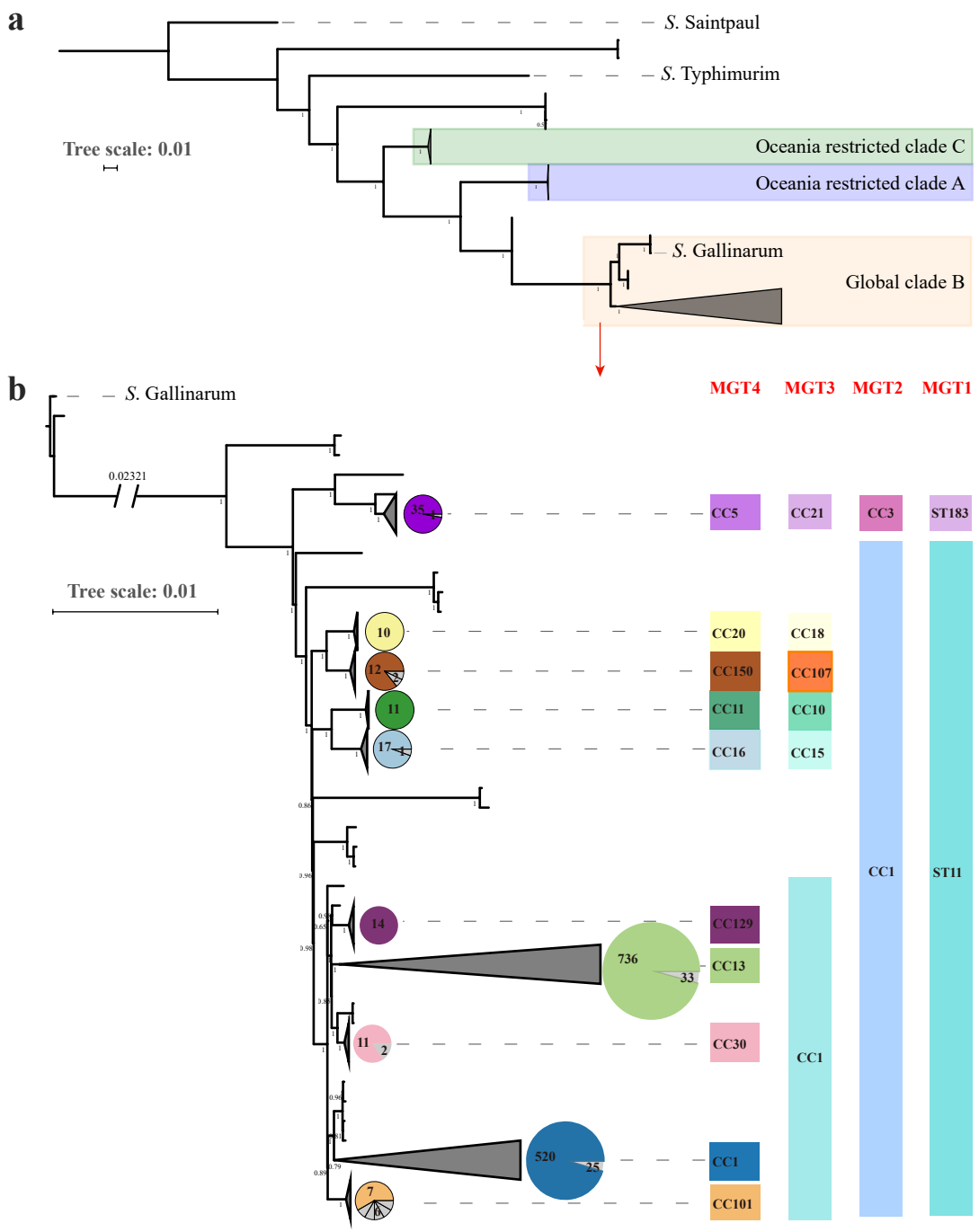
23 20,153

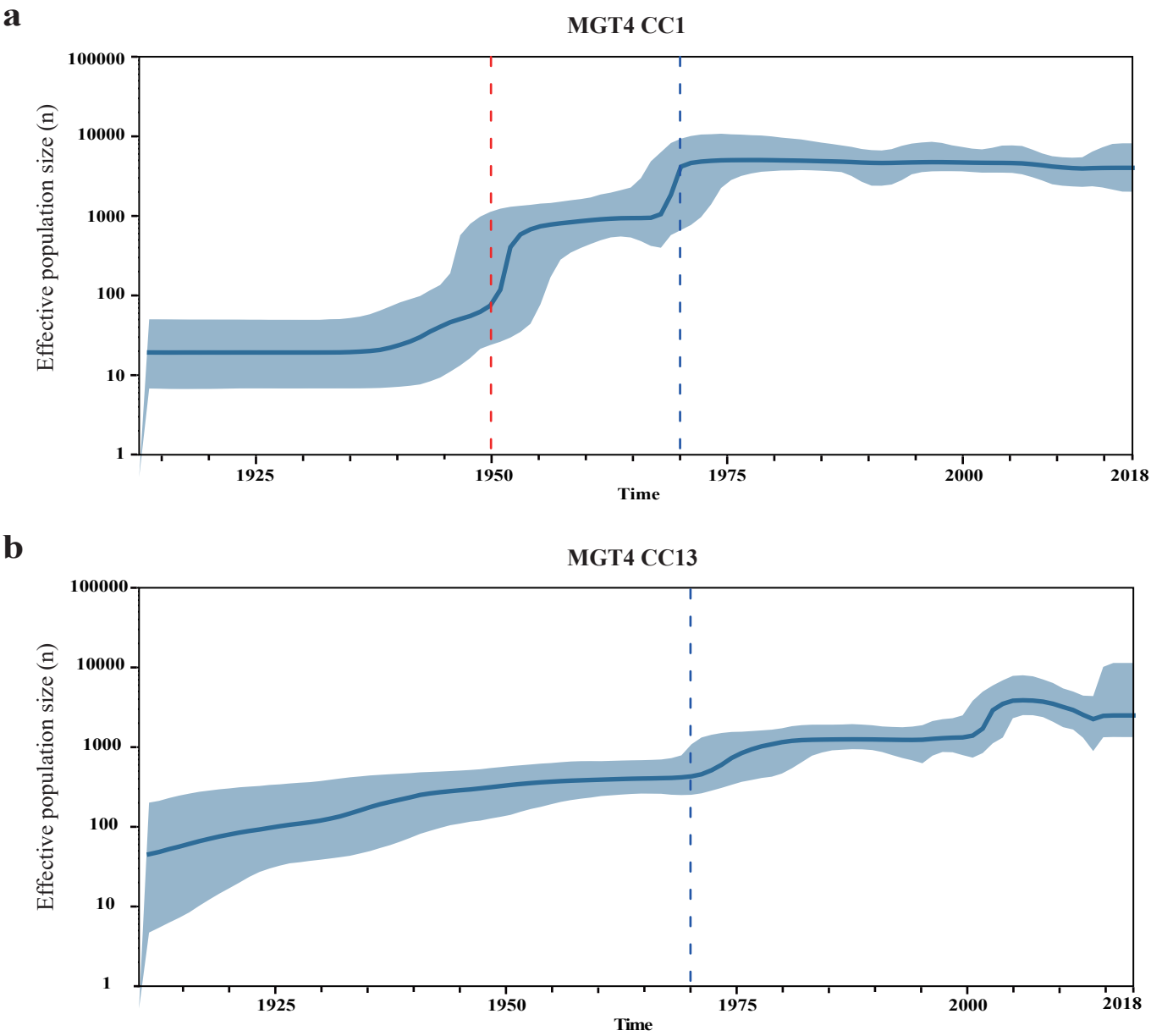

10353
Europe
CC1
CC13
ST15
ST326
ST171
ST298
ST163
ST25
CC5
ST24
ST13
ST100
ST31
ST129
ST117
ST158
13339
North America
CC13
CC1
ST136
ST99
ST135
ST25
ST396
ST171
ST163
ST15
ST370
243
Asia
CC1
ST15
ST65
CC13
CC101
CC69
ST69
388
South America
CC1
ST15
CC13
ST273
ST275
ST276
553
Africa
CC16
CC1
CC13
ST16
ST15
ST106
ST107
ST58
CC292
ST292
CC11
ST11
330
Oceania
CC1
ST15
ST27
ST506
ST377
CC49
CC50
ST49
ST51


a
No. of isolates
Cumulative proportion (%)
MGT4 ST
ST99 ST15 ST25 ST136 ST135 ST171 ST13 ST163 ST100 ST370 ST396 ST150 ST31 ST326 ST24 ST208 ST29 ST198 ST160 ST58 ST416 ST17 ST373 ST1009 ST357 ST47 ST2060 ST278 ST316 ST237 ST20 ST355 ST989 ST144 ST129 ST406 ST688 ST66 ST48
b
MGT4-ST15
MGT4-ST25
No. of isolates
MGT5 ST
ST1 ST33 ST387 ST35 ST257 ST546 ST720 ST90 ST559 ST416 ST425 ST731 ST315 ST117 ST311 ST1789 ST42 ST62 ST1158 ST1190 ST25 ST79 ST156 ST34 ST412 ST29 ST427 ST323 ST681 ST55 ST305 ST269 ST248 ST626
UK
USA
%


MGT4
Avian Source
Human Source
CC1
CC13
United States
No. of isolates
1
200
400
624
MGT4 ST
99
135
136
160
25
198
396
171
15
163
Others
CC1
CC13


No. of isolates
0
20
40
60
80
100
120
140
160
180
200
220
2014.3
2014.4
2014.5
2014.6
2014.7
2014.8
2014.9
2014.10
2014.11
2014.12
2015.1
2015.2
2015.3
2015.4
2015.5
2015.6
2015.7
2015.8
2015.9
2015.10
2015.11
2015.12
2016.1
2016.2
2016.3
2016.4
2016.5
2016.6
2016.7
2016.8
2016.9
2016.10
2016.11
2016.12
2017.1
2017.2
2017.3
2017.4
2017.5
2017.6
2017.7
2017.8
2017.9
2017.10
2017.11
2017.12
2018.1
2018.2
2018.3
2018.4
2018.5
2018.6
2018.7
MGT5 ST
(>=100 isolates)
55
35
591
34
199
333
423
156
33
15
79
29
1


a
S. Saintpaul
S. Typhimurim
Tree scale: 0.01
Oceania restricted clade C
Oceania restricted clade A
S. Gallinarum
Global clade B
b
S. Gallinarum
0.02321
Tree scale: 0.01
MGT4
MGT3
MGT2
MGT1
35
CC5
CC21
CC3
ST183
10
CC20
CC18
12
CC150
CC107
11
CC11
CC10
17
CC16
CC15
CC1
ST11
14
CC129
736
33
CC13
11
2
CC30
CC1
520
25
CC1
7
CC101


a
MGT4 CC1
Effective population size (n)
100000
10000
1000
100
10
1
1925
1950
1975
2000
2018
Time
b
MGT4 CC13
Effective population size (n)
100000
10000
1000
100
10
1
1925
1950
1975
2000
2018
Time