Distinct *Salmonella* Enteritidis lineages associated with enterocolitis in high-income settings and invasive disease in low-income settings

Authors

Nicholas A. Feasey 1,2,3,4, James Hadfield 2, Karen H. Keddy 5, Timothy J Dallman 21, Jan Jacobs 6,7, Xiangyu Deng 8,9, Paul Wigley 4, Lars Barquist 10, Gemma C. Langridge 2, Theresa Feltwell 2, Simon R. Harris 2, Alison E. Mather 2*, Maria Fookes 2, Martin Aslett 2, Chisomo Msefula 3,11, Samuel Kariuki 12, Calman Maclennan 2,13, Robert S. Onsare 12, François X. Weill 14, Simon Le Hello 14, Anthony M. Smith 5, Michael McClelland 15, Prereak Desai 15, Christopher M. Parry 16,17, John Cheesbrough 18, Neil French 4, Josefina Campos 19, Jose A. Chabalgoity 20, Laura Betancor 20, Katie L. Hopkins 21, Satheesh Nair 21, Tom J. Humphrey 22, Octavie Lunguya 23,24, Tristan A. Cogan 25, Milagritos Tapia 26, Samba Sow 27, Sharon M. Tennant 26, Kristin Bornstein 26, Myron M. Levine 26, Lizeth Lacharme-Lora 4, Dean B. Everett 4, Robert A. Kingsley 2,28, Julian Parkhill 2, Robert S. Heyderman 3,29, Gordon Dougan 2, Melita A Gordon† 3,4 & Nicholas R. Thomson† 2, 30

†These authors contributed equally

Affiliations

1. Liverpool School of Tropical Medicine, Liverpool, UK
2. Wellcome Trust Sanger Institute, Cambridge, UK
3. Malawi Liverpool Wellcome Trust Clinical Research Programme, University of Malawi College of Medicine, Blantyre, Malawi
4. Institute of Infection and Global Health, University of Liverpool, Liverpool, UK
5. National Institute for Communicable Diseases and Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa
6. Institute of Tropical Medicine, Antwerp, Belgium
7. Department of Microbiology and Immunology, University of Leuven, Belgium
8. Center for Disease Control, Atlanta, USA
9. Center for Food Safety, Department of Food Science and Technology, University of Georgia

10. Institute for Molecular Infection Biology, University of Würzburg, Würzburg, Germany
11. University of Malawi, The College of Medicine, Blantyre, Malawi
12. Centre for Microbiology Research, Kenya Medical Research Institute, Nairobi, Kenya
13. Jenner Institute, Nuffield Department of Medicine, University of Oxford, Oxford, UK
14. Institut Pasteur, Paris, France
15. Department of Microbiology and Molecular Genetics, University of California, Irvine, CA, USA
16. London School of Hygiene & Tropical Medicine, London, UK
17. School of Tropical Medicine and Global Health, Nagasaki University, Japan
18. Department of Epidemiology and Population Health, University of Liverpool, Liverpool, UK
19. Enteropathogen Division, Administración Nacional de Laboratorios e Institutos de Salud (ANLIS) Carlos G. Malbran Institute, Buenos Aires, Argentina
20. Departamento Desarrollo Biotecnologico, Instituto de Higiene, Facultad de Medicina, Universidad de la Republica, Uruguay
21. Gastrointestinal Bacteria Reference Unit, Public Health England, Colindale, UK
22. Swansea Medical School, Swansea University, Swansea, UK
23. National Institute of Biomedical Research, Kinshasa, the Democratic Republic of the Congo
24. University Hospital of Kinshasa, Kinshasa, the Democratic Republic of the Congo
25. School of Veterinary Sciences, University of Bristol, Bristol, UK
26. Center for Vaccine Development, University of Maryland School of Medicine, Baltimore, Maryland, USA
27. Centre pour le Développement des Vaccins, Bamako, Mali
28. Institute of Food Research, Colney, Norwich, UK
29. Division of Infection and Immunity, University College London, London, UK
30. The London School of Hygiene and Tropical Medicine, London, UK

* Current address: Department of Veterinary Medicine, University of Cambridge, Cambridge, UK

Corresponding Author
Nicholas Feasey
Liverpool School of Tropical Medicine
5 Pembroke Place
Liverpool
L3 5QA
Tel: 0151 705 3214

Running Title: Emergence of distinct lineages of *S*. Enteritidis

Key words: *Salmonella* Enteritidis, global, multidrug resistant, bacteraemia, gastroenteritis

Abstract

An epidemiological paradox surrounds *Salmonella enterica* serovar Enteritidis. In high-income settings, it has been responsible for an epidemic of poultry-associated, self-limiting enterocolitis, whilst in sub-Saharan Africa it is a major cause of invasive nontyphoidal *Salmonella* disease, associated with high case-fatality. Whole-genome sequence analysis of 675 isolates of *S*. Enteritidis from 45 countries reveals the existence of a global epidemic clade and two novel clades of *S*. Enteritidis that are each geographically restricted to distinct regions of Africa. The African isolates display genomic degradation, a novel prophage repertoire and have an expanded, multidrug resistance plasmid. *S*. Enteritidis is a further example of a *Salmonella* serotype that displays niche plasticity, with distinct clades that enable it to become a prominent cause of gastroenteritis in association with the industrial production of eggs, and of multidrug resistant, bloodstream invasive infection in Africa.

Introduction

Salmonella enterica serovar Enteritidis (hereafter referred to as *S*. Enteritidis) has been a global cause of major epidemics of enterocolitis, which have been strongly associated with intensive poultry farming and egg production [1]. The serovar is usually considered to be a generalist in terms of host range and has a low human invasiveness index, typically causing self-limiting enterocolitis [2]. Following a number of interventions in the farming industry involving both improved hygiene and poultry vaccination, epidemic *S*. Enteritidis has been in decline in many countries including the United Kingdom and USA [3,4]. *S.* Enteritidis has also been used extensively since the early 1900s as a rodenticide (named the "Danysz virus"), following development at Institut Pasteur, France. Although by the 1960s, *Salmonella*-based rodenticides had been banned in the US, Germany and the UK, *S*. Enteritidis is still produced as a rodenticide in Cuba, under the name Biorat®[5].

Serovars of *Salmonella* that cause enterocolitis in industrialised settings are strongly associated with life-threatening invasive nontyphoidal *Salmonella* (iNTS) disease in sub-Saharan Africa (SSA). *S*. Enteritidis and *Salmonella enterica* serovar Typhimurium (*S*. Typhimurium) are the two leading causes of iNTS disease in SSA [6] and both are associated with multidrug resistance (MDR)[7]. The clinical syndrome iNTS disease is associated with immunosuppression in the human host, particularly malnutrition, severe malaria and advanced HIV in young children and advanced HIV in adults [8]. It has been estimated to cause 681,000 deaths per year [9].

Salmonella is a key example of a bacterial genus in which there is a recognizable genomic signature that distinguishes between a gastrointestinal and an extra-intestinal/invasive lifestyle [10], whereby functions required for escalating growth in an inflamed gut are lost when the lineage becomes invasive [11]. In order to investigate whether there were distinct bacterial characteristics explaining the very

different epidemiological and clinical profile of epidemic isolates of serotype *S*. Typhimurium from SSA and industrialised settings, whole-genome sequence (WGS) investigations of this serovar were previously undertaken. These revealed a novel pathotype of multilocus sequence type (MLST) ST313 from SSA, which differed from clades that cause enterocolitis in industrialised settings, by showing patterns of genomic degradation potentially associated with more invasive disease and differential host adaptation [12-17].

In relation to *S.* Enteritidis, there is a growing body of literature on the evolutionary history, phylogeny and utility of WGS for surveillance of *S*. Enteritidis outbreaks [18-20]. The broadest study of the phylogeny to date revealed five major lineages, but contained only two African isolates [21]. There have also been limited reports of isolates of *S.* Enteritidis from African patients living in Europe that are MDR and which display a distinct phage type (PT 42) [22,23]. We therefore hypothesized that there are distinct lineages of *S*. Enteritidis circulating in both the industrialised and developing world with different origins, likely distinct routes of spread and that are associated with different patterns of disease, which will display the distinct genomic signatures characteristic of differential adaptation. To investigate this we have collected a highly diverse global collection of *S*. Enteritidis isolates and compared them using whole-genome sequencing, the highest possible resolution typing methodology.

Results

Isolate collection

In total, 675 isolates of *S.* Enteritidis isolated between 1948 and 2013 were sequenced. The collection originated from 45 countries and six continents (Table 1). 496/675 isolates were from Africa, with 131 from the Republic of South Africa (RSA), a further 353 from the rest of SSA, and 12 from North Africa (Table 1). There were 343 isolates from normally sterile human sites (invasive), 124 non-invasive human isolates (predominantly stool samples) and 40 from animal, food or environmental sources. The full metadata are described in Supplementary Table 1 and have been uploaded to the publically available database Enterobase (https://enterobase.warwick.ac.uk/).

Phylogeny

675 *S.* Enteritidis genomes and one *Salmonella enterica* serovar Gallinarum were mapped to the *S.* Enteritidis strain P125109 reference sequence, variable regions excluded and the remaining sites were screened for single nucleotide polymorphisms (SNPs). This left an alignment containing a total of 42,373 variable sites, from which a maximum likelihood (ML)-phylogeny was constructed using *S.* Gallinarum, which is a closely related serovar, as an out-group (Figure 1). HierBAPS was run over two rounds, which provided clear distinction between clades/clusters [24]. The phylogeny of *S.* Enteritidis revealed evidence of three clades associated with epidemics, one which we have termed the 'global epidemic clade' and includes the reference PT4 isolate P125109 and two African clades: one predominantly composed of West African isolates (labeled the 'West African clade') and a second composed of isolates predominantly originating in Central and Eastern Africa, called the 'Central/Eastern African clade'). Figure 1 also shows the other clades and clusters predicted by HierBAPS, the largest of which is a paraphyletic cluster from

which the global epidemic clade emerged (Outlier Cluster in Figure 1), and a further five smaller clades or clusters predicted by HierBAPS.

The global epidemic clade contains isolates of multiple phage types, including 4 and 1, which have been linked to the global epidemic of poultry associated human enterocolitis [25]. It comprised 250 isolates from 28 countries, including 43 from Malawi and 82 from RSA. They were isolated from across a 63-year period (1948-2013). Antimicrobial susceptibility testing had been performed on 144 isolates and 104 were susceptible to all antimicrobials tested, five were multidrug resistant (MDR: resistant to 3 or more antimicrobial classes), one was nalidixic acid resistant and none were extended-spectrum beta-lactamase (ESBL)-producing isolates. Database comparison of the genomes from this clade revealed that 221 (88%) of them contained no predicted antimicrobial resistance (AMR) genes apart from the cryptic resistance gene *aac(6')-Iy [26]*.

The global epidemic clade has emerged from a diverse cluster previously described by Zheng [27], which encompassed 131 isolates (Figure 1: 'Outlier Cluster'). In addition to being paraphyletic, this group was geographically and temporally diverse, and predominantly drug susceptible (59/71 isolates). Whilst the majority of the diversity of phage typed isolates was contained within the global epidemic clade, this cluster alone contained isolates of phage type 14b, which was recently associated with a multi-country outbreak of *S*. Enteritidis enterocolitis in Europe associated with chicken eggs from Germany [28]. There were also 41 isolates from RSA in this clade, where it has been a common cause of bloodstream infection, and 39 bloodstream isolates from Malawi. Database comparison of the genomes from this clade revealed that 122 (82%) of these genomes contained no predicted AMR genes apart from the cryptic resistance gene *aac(6')-Iy*.

There were two related, but phylogenetically and geographically distinct, epidemic clades that largely originated from SSA. The Central/Eastern African clade included

166 isolates, all but two of which (from RSA) came from this region. Of these, 126/155 (82%) were MDR and 148/153 (97%) displayed phenotypic resistance to between one and four antimicrobial classes. All of these genomes contained at least five predicted resistance genes and 128 (77%) contained nine (Table 2 and Supplementary Table 2). 155/165 (94%) of these isolates were cultured from a normally sterile compartment of a human (i.e. blood or cerebrospinal fluid) and were considered to be causing invasive disease (Table 2). The second African epidemic clade was significantly associated with West Africa with 65/66 isolates coming from this region and one isolate from USA. This clade was also associated with drug resistance (62 [94%] resistant to ≥1 antimicrobial class by phenotype and genotype) and human invasive disease (61 [92%]). It also included two isolates that were subtyped as phage type 4.

The remaining 58 isolates included in this study were extremely diverse, phylogenetically, temporally and geographically. Only two displayed any phenotypic AMR, one of which was MDR. Inspection of the genome revealed that five had predicted AMR genes in addition to *aac(6')-Iy*, four of which were isolated in sub-Saharan Africa. Twenty were associated with invasive human disease, and six were recovered from stool. Three isolates were from stocks of rodenticide and these were phylogenetically remote from both global-epidemic and the two African epidemic clades.

To add further context to these findings we screened the entire publically available Public Health England (PHE) sequenced *Salmonella* routine surveillance collection, which includes 2,986 *S*. Enteritidis genomes, 265 of which were associated with travel to Africa (Supplementary Figure 1). Within this huge collection, including 61 (2.0%) bloodstream isolates and 2670 (89.4%) stool isolates, only 6 isolates (4 from blood culture, 1 from stool) fell within to the West African clade and 1 (from stool) belonged to the Central/Eastern African clade. Notably, these isolates were all either associated with travel to Africa and/or taken from patients of African origin.

It is apparent from the location of the archetypal reference isolate and archetypal phage types in the phylogeny (Supplementary Figure 2) that the majority of *S*. Enteritidis studied previously belonged to the global epidemic clade associated with enterocolitis in industrialised countries. Furthermore, its also clear that two additional, previously unrecognized *S*. Enteritidis lineages have emerged, largely restricted to Africa, that are strongly associated with MDR and invasive disease.

To understand how recently these African-associated lineages emerged we used Bayesian Evolutionary Analysis by Sampling Trees (BEAST) to reconstruct the temporal history of the epidemic clades [29]. These data (Supplementary Figure 3) estimate the most recent common ancestor (MRCA) of the Central/Eastern African clade dates to 1945 (95% Credible Interval [CrI]: 1924-1951) and for the West African clade it was 1933 (95% CrI: 1901-1956). We estimate the MRCA of the global epidemic clade originated around 1918 (95% CrI: 1879-1942 – Supplementary Figure 4), with a modern expansion occurring in 1976 (95% CrI: 1968-1983), whereas the paraphyletic cluster from which it emerged dates to approximately 1711 (95% CrI: 1420-1868).

The contribution of the accessory genome to the emergence of the African clades

Prophages have the potential to carry non-essential "cargo" genes, which suggests they confer a level of specialization to their host bacterial species, whilst plasmids may confer a diverse array of virulence factors and AMR [30,31]. Therefore it is critical to evaluate the accessory genome in parallel with the core. 622 sequenced genomes were used to determine a pangenome, which yielded a core genome comprising 4,076 predicted genes present in ≥90% isolates, including all 12 recognised *Salmonella* Pathogenicity Islands as well as all 13 fimbrial operons found in the P125109 reference [32]. The core gene definition was set to minimize stochastic loss of genes from the core due to errors in individual assemblies across

such a large dataset. The accessory genome consisted of 14,015 predicted genes. Of the accessory genes, 324 were highly conserved across the global and two African epidemic clades, as well as the outlier cluster. Almost all were associated with the acquisition or loss of mobile genetic elements (MGEs) such as prophage or plasmids. Prophage regions have been shown to be stable in *Salmonella* genomes and are potential molecular markers, the presence of which has previously been used to distinguish specific clades [13,33].

The lineage-specific whole gene differences of the major clades are summarized in Figure 2 and plotted against the representatives of the four major clades in Supplementary Figure 3. The lineage specific sequence regions include 57 predicted genes found to be unique to the global epidemic clade (Figure 2), all of which were associated with prophage ϕSE20, a region shown to be essential for invasion of chicken ova and mice in one previous study [34]. There were a further 39 genes conserved in the global epidemic and the paraphyletic outlying cluster, which were absent from both African clades, 26 of which correspond to region of difference (ROD) 21 [32]. The Central/Eastern Africa clade contained 77 predicted genes that were absent in the other clades. 33 were associated with the virulence plasmid and a further 40 chromosomal genes were associated with a novel, Fels-2 like prophage region (ϕ*fels-BT*). The West African clade had only 15 distinct predicted genes, 11 of which were plasmid-associated. The two African clades shared a further 102 genes: 48, including a leucine-rich repeat region, were associated with a novel prophage region closely related to *Enterobacter* phage P88, 44 were associated with a Gifsy-1 prophage found in *S*. Bovismorbificans and eight were associated with a Gifsy-2 prophage which has degenerated in the reference P125109.

The *S.* Enteritidis plasmid is the smallest of the generic *Salmonella* virulence plasmids at 58 kb and is unusual in that it contains an incomplete set of *tra* genes that are responsible for conjugative gene transfer. The phylogeny of the *S*. Enteritidis virulence plasmid backbone was reconstructed using reads that mapped

to the *S.* Enteritidis reference virulence plasmid, pSENV. 120/675 (18%) genomes lacked pSENV. The virulence plasmid phylogeny is similar to that of the chromosome, suggesting that they have been stably maintained by each lineage and diversified alongside them (Supplementary Figure 6).

The virulence plasmids from the African clades were much larger than those held in the other clades at ~90 kb. A representative example was extracted from Malawian isolate D7795, sequenced using long read technology to accurately reconstruct it (PacBio; see methods) and denoted pSEN-BT (Accession number LN879484). pSEN-BT is composed of a backbone of pSENV with additional regions that are highly similar to recently sequenced fragments of an novel *S.* Enteritidis virulence plasmid (pUO-SeVR) isolated from an African patient presenting with MDR invasive *S.* Enteritidis in Spain [22]. Plasmid pSEN-BT harbours nine AMR genes (full list in Supplementary Table 2), plus additional genes associated with virulence and a toxin/antitoxin plasmid addiction system. Of note, plasmids from the West African isolates carry resistance gene chloramphenicol acetyl transferase A1 (*cat*A1), whereas the Central/Eastern African strains carry *cat*A2 and tetracycline resistance gene *tet*(A). Like pSENV, the African virulence plasmid contained an incomplete set of *tra* genes and so is not self-transmissible. This was confirmed by conjugation experiments and is consistent with previous reports [22,23]. These observations suggest that the evolution of the *S.* Enteritidis plasmid mirrors that of the chromosome; it is thus not a 'novel' plasmid, but in different SSA locations has acquired different AMR genes.

Multiple signatures of differential host adaptation

It has been observed in multiple serovars of *Salmonella* including *S*. Typhi, *S*. Gallinarum and *S*. Typhimurium ST313 that the degradation of genes necessary for the utilization of inflammation-derived nutrients is a marker of that lineage having moved from an intestinal to a more invasive lifestyle [13,14,32,35]. Accordingly, we

have looked for similar evidence within a representative example of a MDR, invasive, Central/Eastern African clade isolate, D7795, that was isolated from the blood of a Malawian child in 2000. The draft genome sequence of D7795 closely resembles that of P125109, however, in addition to the novel prophage repertoire and plasmid genes described above, it harbours a number of predicted pseudogenes or hypothetically disrupted genes (HDGs)[11].

In total, there were 42 putative HDGs in D7795, many of which are found in genes involved in gut colonisation and fecal shedding as well as various metabolic processes such as cobalamine biosynthesis which is a cofactor for anaerobic catabolism of inflammation-derived nutrients, such as ethanolamine, following infection [36]. Curation of the SNPs and insertions or deletions (indels) predicted to be responsible for pseudogenisation across the Central/Eastern African clade and West African clade revealed 37/42 predicted HDGs were fixed in other representatives of the Central/East African clade, with 27 of them being present in over 90% of isolates from that clade. Relatively fewer HDGs in D7795 (19/42) were present in representatives of the West African clade, although 13 were present in ≥90% of isolates (Supplementary Table 3).

In addition to this evidence of reductive evolution in D7795, there were 363 genes containing non-synonymous (NS)-SNPs, which change the amino acid sequence and so may have functional consequences [37]. The two African clades were screened for the presence of these NS-SNPs and 131 were found to be present and completely conserved across both clades, including NS-SNPs in 43 genes encoding predicted membrane proteins, 36 metabolic genes and 23 conserved hypothetical genes (Supplementary Table 4). Furthermore many of these NS-SNPs fall in genes within the same metabolic pathways as the HDGs (see Supplementary Results for detailed description). Supplementary Table 5 provides a list of some of the common traits identified amongst the functions of genes lost independently by D7795, *S.* Typhi and *S*. Gallinarum. The disproportionate clustering of mutations in membrane structures

observed in the African clades is yet another sign of differential host adaptation analogous to that reported in both *S*. Typhi [35] and *S*. Gallinarum [32].

Biolog™ growth substrate platform profiling

The Biolog™ platform was utilized to generate a substrate growth utilisation profile for selected *S*. Enteritidis isolates (see high throughput phenotyping protocol in Supplementary materials). Corresponding signal values of replicate pairs of a Central/Eastern African isolate (D7795) and a global epidemic isolate (A1636) were compared using principal component analysis and found to be highly consistent. In total, 80 metabolites showed evidence of differential metabolic activity (Figure 3). Evaluation of data from the Central/Eastern African isolate using Pathway Tools software revealed that 14/27 (52%) of pathways with evidence of decreased metabolic activity at 28°C had a corresponding component of genomic degradation. This was also true for 12/30 (40%) of pathways with evidence of decreased metabolic activity at 37°C.

Instances of reduced metabolic activity in a Central/Eastern African strain (D7795) compared to a global epidemic strain (A1636) included dulcitol and glycolic acid in the glycerol degradation pathway, propionic acid in the propanediol pathway and ethylamine and ethanolamine. These are all vitamin B12 (cobalamin) dependent reactions, for which there was a corresponding signature of genomic degradation. Also there was reduced activity in response to three forms of butyric acid, alloxan and allantoic acid metabolism. Allantoin can be found in the serum of birds, but not humans and is utilised as a carbon source during *S.* Enteritidis infection of chickens, [38] and HDGs relating to allantoin have been noted in *S.* Typhimurium ST313 [13]. The full list of differences is detailed in Supplementary Table 6 and 7. This is a further sign of decreased metabolism of the Central/Eastern African isolate in the anaerobic environment of the gut.

Chicken infection model suggests evolutionary divide in host range between global epidemic and African lineages

Given the phenotypic differences observed in the genotypically distinct global and African clades, we hypothesized that these lineages could have differing infection phenotypes in an *in vivo* challenge model. We compared the infection profile of a member of the Central/Eastern African clade (D7795) to the reference global epidemic strain P125109 in an avian host. The chicken group infected with P125109 showed mild hepatosplenomegaly consistent with infection by this *Salmonella* serovar and cecal colonization (Figure 4A-C). In contrast, the Central/Eastern African strain displayed significantly reduced invasion at 7 dpi of both liver ($p=0.027$) and spleen ($p=0.007$), however cecal colonization was not significantly reduced ($p=0.160$). This is in marked contrast to the behavior of *S*. Typhimurium ST313, which is more invasive in a chick infection model [12].

Discussion

S. Enteritidis is an example of a successful *Salmonella* lineage with the apparent ability to adapt to different hosts and transmission niches as and when opportunities for specialization have presented. Langridge *et al* recently evaluated the Enteritidis/Gallinarum/Dublin lineage of *Salmonella*, revealing components of the nature and order of events associated with host-range and restriction [39]. In the present study, we have highlighted the plasticity of *S*. Enteritidis, providing evidence of three distinct epidemics of human disease. In addition we show multiple additional clades and clusters that demonstrate the huge reservoir of diversity amongst *S.* Enteritidis from which future epidemics might emerge.

An important question posed by this study is why have distinct clades of *Salmonella* emerged to become prominent causes of iNTS disease in Africa, from a serotype

415 normally considered to be weakly invasive? The presence of a highly
416 immunosuppressed population due to the HIV pandemic is clearly a key host factor
417 that facilitates the clinical syndrome iNTS disease [40,41]. In addition to human host
418 factors, there are two distinct African epidemic lineages that have emerged in the
419 last 90 years. Both lineages are significantly associated with a novel prophage
420 repertoire, an expanded, MDR-augmented virulence plasmid, and patterns of
421 genomic degradation with similarity to other host-restricted invasive *Salmonella*
422 serotypes including *S*. Typhi and *S*. Gallinarum and to clades of *S.* Typhimurium
423 associated with invasive disease in Africa [13,32,35]. This pattern of genomic
424 degradation is concentrated in pathways specifically associated with an enteric
425 lifestyle, however it is noteworthy that in the chick infection model, the African *S*.
426 Enteritidis invaded the chick liver and spleen less well than the global pandemic
427 clade. This raises the possibility that the two clades occupy different ecological
428 niches outside the human host or that they behave differently within the human
429 host and screening of the huge *S.* Enteritidis collection from routine *Salmonella*
430 surveillance by PHE supports the assertion that these lineages are geographically
431 restricted to Africa. This study therefore indicates a need to understand what these
432 ecological niches might be, and then to define the transmission pathways of African
433 clades of *S*. Enteritidis, in order to facilitate public health interventions to prevent
434 iNTS disease.
435
436 The evolution of the *S*. Enteritidis virulence plasmid is intriguing; pSENV is the
437 smallest of the known *Salmonella* virulence-associated plasmids, but in SSA, the
438 plasmid has nearly doubled in size partly through the acquisition of AMR genes. The
439 absence of *tra* genes necessary for conjugal transfer either indicates that MDR status
440 has evolved through acquisition of MGEs multiple times or through clonal expansion
441 and vertical transmission of the plasmid to progeny. The available data suggest that
442 the former scenario has happened twice, once in West Africa, and once in
443 Central/Eastern Africa.
444

Despite *S*. Enteritidis being reported as a common cause of bloodstream infection (BSI) in Africa [6,7] the Global Enteric Multicenter Study (GEMS) found that *Salmonella* serotypes were an uncommon cause of moderate to severe diarrhoea in African children less than 5-years of age [42]. Our data associating the African lineages with invasive disease is also consistent with data presented in a recent Kenyan study comparing a limited number and diversity of *S*. Enteritidis isolates from blood and stool. Applying the lineages defined in this study to the genome data reported from Kenya showed that 20.4% of isolates from that study belonging to the global clade were associated with invasive disease, whereas 63.2% of the isolates in that study belonging to our Central/Eastern African clade were associated with invasive disease [43]. The remaining isolates were associated with cases of enterocolitis or asymptomatic carriage, confirming that the Central/Eastern African clade can also cause enterocolitis. The association of *S*. Enteritidis clades circulating in sub-Saharan Africa with iNTS disease may reflect the fact that their geographical distribution permits them to behave as opportunistic invasive pathogens in a setting where advanced immunosuppressive disease is highly prevalent in human populations.

In summary, two clades of *S.* Enteritidis have emerged in Africa, which have different phenotypes and genotypes to the strains of *S*. Enteritidis circulating in the industrial world. These strains display evidence of changing host adaptation, different virulence determinants and multi-drug resistance, a parallel situation to the evolutionary history of *S*. Typhimurium ST313. They may have different ecologies and/or host ranges to global strains and have caused epidemics of BSI in at least three countries in SSA, yet are rarely responsible for disease in South Africa. An investigation into the environmental reservoirs and transmission of these pathogens is warranted and urgently required.

Methods

475 Bacterial Isolates

476

477 *S*. Enteritidis isolates were selected on the basis of six factors; date of original
478 isolation, antimicrobial susceptibility pattern, geographic site of original isolation,
479 source (human [invasive vs stool], animal or environmental), phage type (where
480 available), and multilocus variable number tandem repeat (MLVA) type (where
481 available). *S.* Enteritidis P125109 (EMBL accession no. AM933172) isolated from a
482 poultry farm from the UK was used as a reference [32]. The full metadata are in
483 Supplementary Table 1. Isolates have been attributed to region according to United
484 Nations statistical divisions
485 (http://unstats.un.org/unsd/methods/m49/m49regin.htm).

486

487 Sequencing, SNP-calling, construction of phylogeny and comparative genomics

488

489 PCR libraries were prepared from 500 ng of DNA as previously described [44].
490 Isolates were sequenced using Illumina GA II, HiSeq 2000 and MiSeq machines
491 (Illumina, San Diego, CA, USA) and 150 bp paired-end reads were generated. The
492 strains were aligned to *Salmonella* Enteritidis reference genome P125109 using a
493 pipeline developed in-house at the Wellcome Trust Sanger Institute (WTSI). For
494 each isolate sequenced, the raw sequence read pairs were split to reduce the overall
495 memory usage and allow reads to be aligned using more than one CPU. The reads
496 were then aligned using SMALT (www.sanger.ac.uk/science/tools/smalt-0), a
497 hashing based sequence aligner. The aligned and unmapped reads were combined
498 into a single BAM file. Picard (https://broadinstitute.github.io/picard) was used to
499 identify and flag optical duplicates generated during the making of a standard
500 Illumina library, which reduces possible effects of PCR bias. All of the alignments
501 were created in a standardized manner, with the commands and parameters stored
502 in the header of each BAM file, allowing for the results to be easily reproduced.

503

504 The combined BAM file for each isolate was used as input data in the SAMtools

505 mpileup program to call SNPs and small indels, producing a BCF file describing all of
506 the variant base positions [45]. A pseudo-genome was constructed by substituting
507 the base call at each variant or non-variant site, defined in the BCF file, in the
508 reference genome. Only base calls with a depth of coverage >4 or quality >50 were
509 considered in this analysis. Base calls in the BCF file failing this quality control filter
510 were replaced with the "N" character in the pseudo-genome sequence.
511
512 All of the software developed is freely available for download from GitHub under an
513 open source license, GNU GPL 3.
514
515 Phylogenetic modelling was based on the assumption of a single common ancestor,
516 therefore variable regions where horizontal genetic transfer occurs were excluded
517 [46] [47]. A maximum likelihood (ML) phylogenetic tree was then built from the
518 alignments of the isolates using RAxML (version 7.0.4) using a GTR+I+G model [48].
519 The maximum-likelihood phylogeny was supported by 100 bootstrap pseudo-
520 replicate analyses of the alignment data. Clades were predicted using Hierarchical
521 Bayesian Analysis of Population Structure (HierBAPS)[24]. This process was
522 repeated to construct the plasmid phylogeny, using reads that aligned to pSENV.
523 To ascertain the presence of the clusters defined by HierBAPs in the Public Health
524 England (PHE) routine *Salmonella* surveillance collection, seventeen isolates
525 representing the diversity of the collection were compared against 2986 *S*.
526 Enteritidis PHE genomes. Single linkage SNP clustering was performed as
527 previously described [49]. A maximum-likelihood phylogeny showing the
528 integration of the seventeen isolates with 50-SNP cluster representatives of the PHE
529 *S*. Enteritidis collection was constructed as above. FASTQ reads from all PHE
530 sequences in this study can be found at the PHE Pathogens BioProject at the
531 National Center for Biotechnology Information (Accession PRJNA248792).
532
533
534 Temporal reconstruction was performed using Bayesian Evolutionary Analysis

Sampling Trees (BEAST: http://beast.bio.ed.ac.uk/ version 1.8.2)[50]. A relaxed lognormal clock model was initially employed. The results of this model indicated that a constant clock model was not appropriate, as the posterior of the standard deviation of the clock rate did not include zero. A range of biologically plausible population models (constant, exponential and skyline) was investigated. Skyline models can be biased by non-uniform sampling and we observed a strong similarity between reconstructed skyline population and the histogram of sampling dates and so this model was excluded. The exponential models consistently failed to converge and were excluded. Thus, for all datasets, lognormal clock and constant population size models were used. The computational expense required for this analysis precluded running estimators for model selection. However, we note that Deng et al used the same models in their analysis of 125 *S*. Enteritidis isolates. Default priors were used except for ucld.mean, Gamma(0.001,1000), initial: 0.0001; exponential.popSize, LogNormal(10,1.5), initial: 1[21].

Three chains of 100 million states were run in parallel for each clade of the four major HierBAPS clades, as well as a fourth chain without genomic data to examine the influence of the prior, which in all cases was uninformative. The final results, as used here, all had effective sample sizes (ESS) of over 200 and had convergence between all three runs. For the Global and Global Outlier lineages, the datasets were not computationally feasible to analyse. We thus created 3 further random subsets of the data by drawing n isolates from each sampled year where n was sampled from a Poisson distribution where $\lambda=2$. The posteriors of all subsets were extremely similar and runs were combined to produce the final most recent common ancestor (MRCA) estimates.

In order to gain a detailed insight into genomic differences, a single high quality sequence from Malawian *S*. Enteritidis isolate D7795 was aligned against the P125109 using ABACAS and annotated [51]. Differences were manually curated against the reference using the Artemis Comparison Tool (ACT)[52]. Sections of

contigs which were incorporated into the alignment, but which did not align with P125109 were manually inspected and compared to the public databases using BLASTn (http://blast.ncbi.nlm.nih.gov). When these regions appeared to be novel prophages, they were annotated using the phage search tool PHAST and manually curated [53]. In order to investigate whether the SNPs and/or indels that were predicted to be responsible for pseudogene formation in D7795 were distinct to that isolate or conserved across both African epidemic clades, all isolates were aligned to P12509 and the relevant SNPs/indels investigated using *in-silico* PCR of the aligned sequences. Manual curation was performed to confirm the nature of all pseudogene-associated SNPs/indels. NS-SNPs identified in D7795 were sorted throughout the African clades by extracting and aligning the appropriate gene sequences from P125109 and D7795. The coordinates of the NS-SNPs were then used to identify the relevant sequence and determine the nature of the base.

Accessory genome

The pangenome for the dataset was predicted using ROARY [54]. Genes were considered to be core to *S*. Enteritidis if present in ≥90% of isolates. A relaxed definition of core genome was used as assemblies were used to generate it and the more assemblies one uses, the more likely it is that a core gene will be missed in one sample due to an assembly error. The remaining genes were considered to be core to the clades/clusters predicted by HierBAPS if present in ≥75% if isolates from within each clade/cluster. These genes were then curated manually using ACT to search for their presence and position in P125109 or the improved draft assembly of representative isolates of each of the other clades if not present in P125109. Any large accessory regions identified were blasted against the assembled genomes of the entire collection to confirm they were grossly intact.

Plasmid identification

Plasmid DNA was extracted from isolate D7795 using the Kado & Liu method and separated by gel-electrophoresis alongside plasmids of known size, to estimate the

number and size of plasmids present [55]. Plasmid conjugation was attempted by mixing 100 µL of overnight culture of donor and recipient strains (rifampicin resistant *Escherichia coli* C600) on Luria-Bertani agar plates and incubating overnight at 26°C and 37°C. The plasmid was sequenced using the PacBio platform (http://www.pacificbiosciences.com/) to gain long reads and a single improved draft assembly, which was aligned against P125109 plasmid pSENV (Accession Number HG970000). For novel regions of the plasmid from isolate D7795, genes were predicted using GLIMMER and manual annotations applied based on homology searches against the public databases, using both BLASTn and FASTA. The plasmid phylogeny was reconstructed using the same methodology as the chromosome; a maximum likelihood (ML) phylogenetic tree was built from the alignments of the isolates using RAxML (version 7.0.4) using a GTR+I+G model

Identification of AMR genes

A manually curated version of the Resfinder database was used to investigate the isolates for the presence of AMR genes [56]. To reduce redundancy, the database was clustered using CD-HIT-EST [57], with the alignment length of the shorter sequence required to be 90% the length of the longer sequence. All other options were left as the defaults. The representative gene of each cluster was then mapped with SMALT (http://www.sanger.ac.uk/science/tools/smalt-0) to the assemblies of each isolate to identify and matches with an identity of 90% or greater were considered significant, in line with the default clustering parameters of CD-HIT-EST. Where partial matches were identified at the ends of contigs, having an identity of 90% or greater to the matched region of the gene, potential AMR gene presence was recorded. To confirm presence of these partial matches, raw sequencing reads of the pertinent isolates were mapped using SMALT to these genes to check for 90% identity across the entire gene.

625 Biolog™ growth substrate platform profiling
626
627 The Biolog™ platform (http://www.biolog.com) enables the simultaneous
628 quantitative measurement of a number of cellular phenotypes, and therefore the
629 creation of a phenotypic profile of a variety of assay conditions [58]. Incubation and
630 recording of phenotypic data were performed using an OmniLog® plate reader. In
631 these experiments, two replicates of D7795 were compared to two of PT4-like strain
632 A1636 at 28 and 37˚C to represent environmental and human temperatures.
633 Biolog™ plates PM1-4 and 9 (Carbon source [PM1, PM2], nitrogen source [PM3] and
634 phosphor and sulphur source [PM4] metabolism and osmotic pressure [PM9]) were
635 used. Each well was inoculated as described in the high throughput phenotyping
636 protocol, thereby testing 475 conditions at once (each plate has one negative control
637 well). Plates were scanned every 15 min for 48 hours while incubated at 28°C and
638 37°C in air. Two paired replicates were performed for each of the two isolates.
639
640 After completion of the run, the signal data were compiled and analysed using the
641 limma package (www.bioconductor.org) in 'R' (www.R-project.org) as described
642 previously [59]. A log-fold change of 0.5 controlling for a 5% false discovery rate
643 was used as a cut-off for investigating a specific metabolite further using Pathway
644 Tools [60] and whether the metabolic change was related to pseudogenes and non
645 synonymous(NS)-SNPs in genes in the respective genomes.
646
647 *In vivo* Infection Model
648
649 Two isolates were used in the animal models: *S*. Enteritidis P125109 and D7795.
650 Unvaccinated commercial female egg-layer Lohmann Brown chicks (Domestic Fowl
651 [*Gallus gallus*]) were obtained from a commercial hatchery and housed in secure
652 floor pens at a temperature of 25°C. Eight chicks per strain per time point were
653 inoculated by gavage at 10 days (d) of age and received a dose of ~10^8 *Salmonella*
654 colony forming units (CFU) in a volume of 0.2 mL. Subsequently, four to five birds

655 from each group were humanely killed at 3, 7 or 21 d post-infection (p.i.). At post
656 mortem, the liver, spleen, and caecal contents were removed aseptically,
657 homogenised, serially diluted and dispensed onto Brilliant Green agar (Oxoid) to
658 quantify colony forming units (CFU) as described previously [61]. Statistical analysis
659 was performed using SPSS, version 20 (IBM). Kruskal-Wallis was used to compare
660 bacterial loads between infected groups.
661
662 All work was conducted in accordance with the UK legislation governing
663 experimental animals, Animals (Scientific Procedures) Act 1986, under project
664 licence 40/3652 and was approved by the University of Liverpool ethical review
665 process prior to the award of the project license. The licensing procedure requires
666 power calculations to determine minimal group sizes for each procedure to ensure
667 results are significant. For these experiments a group size of 8 birds per time point
668 was chosen, based on a variation in 1.0 $\log_{10}$ in bacterial count between groups as
669 being significant along with prior experience of *Salmonella* infection studies. Groups
670 were randomly selected on receipt from the hatchery and investigators conducting
671 animal experiments were not blinded, as the current UK code of practice requires all
672 cages or pens to be fully labeled with experimental details. No animals were
673 excluded from the analysis. All animals were checked a minimum of twice daily to
674 ensure their health and welfare.
675
676 Code availability
677
678 Software is referenced and URLS are provided in the text of the methods, all
679 software is open source.
680

ACKNOWLEDGEMENTS

This work was supported by The Wellcome Trust. We would like to thank the members of the Pathogen Informatics Team and the core sequencing teams at the Wellcome Trust Sanger Institute (Cambridge, UK). We are grateful to David Harris for work in managing the sequence data.

FUNDING

This work was supported by a number of organizations. The Wellcome Trust Sanger Institute authors were funded by Wellcome Trust Award 098051; NF was supported by the Wellcome Trust Research Fellowship WT092152MA. NF, RSH and this work were supported by a strategic award from the Wellcome Trust for the MLW Clinical Research Programme (101113/Z/13/Z). The Institut Pasteur (IP) authors were funded by the IP, the Institut de Veille Sanitaire, and by the French Government "Investissement d'Avenir" program (Integrative Biology of Emerging Infectious Diseases" Laboratory of Excellence, grant no. ANR-10-LABX-62-IBEID). JJ was supported by the antibiotic resistance surveillance project in DR Congo, funded by Project 2.01 of the Third Framework Agreement between the Belgian Directorate General of Development Cooperation and the Institute of Tropical Medicine, Antwerp, Belgium. SK was supported by the NIH Grant Number R01 AI099525-02. AEM was supported by Wellcome Trust grant 098051 whilst at the Wellcome Trust Sanger Institute, and Biotechnology and Biological Sciences Research Council grant BB/M014088/1 at the University of Cambridge. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

AUTHOR CONTRIBUTIONS

Study design: NAF, NRT, MAG, GD, RAK, JP. Data analysis: NAF, NRT, JH, TF, LB, PQ LLL, GL, SRH, AEM, MF, MA. Isolate acquisition and processing and clinical data collection: NAF, KHK, JJ, XD, CMe, SK, CMl, RSO, FXW, SLH AMS, MM, PD, CMP, JC, NF, JC, JAC, LBe, KLH, TJH, OL, TAC, M T, SS, SMT, KB, MML, DBE, RSH. Manuscript writing: NAF, JH, NRT, MAG. All authors contributed to manuscript editing.

COMPETING FINANCIAL INTERESTS

711 The authors declare no competing financial interests.

712

713

714 References

715 1. Ward LR, Threlfall J, Smith HR, O'Brien SJ (2000) Salmonella enteritidis epidemic.
716 Science 287: 1753-1754; author reply 1755-1756.

717 2. Jones TF, Ingram LA, Cieslak PR, Vugia DJ, Tobin-D'Angelo M, et al. (2008)
718 Salmonellosis outcomes differ substantially by serotype. J Infect Dis 198: 109-
719 114.

720 3. O'Brien SJ (2013) The "decline and fall" of nontyphoidal salmonella in the United
721 kingdom. Clin Infect Dis 56: 705-710.

722 4. Braden CR (2006) Salmonella enterica serotype Enteritidis and eggs: a national
723 epidemic in the United States. Clin Infect Dis 43: 512-517.

724 5. Friedman CR, Malcolm G, Rigau-Perez JG, Arambulo P, 3rd, Tauxe RV (1996) Public
725 health risk from Salmonella-based rodenticides. Lancet 347: 1705-1706.

726 6. Reddy EA, Shaw AV, Crump JA (2010) Community-acquired bloodstream infections
727 in Africa: a systematic review and meta-analysis. Lancet Infect Dis 10: 417-432.

728 7. Feasey NA, Masesa C, Jassi C, Faragher EB, Mallewa J, et al. (2015) Three
729 Epidemics of Invasive Multidrug-Resistant Salmonella Bloodstream Infection in
730 Blantyre, Malawi, 1998-2014. Clin Infect Dis 61 Suppl 4: S363-371.

731 8. Feasey NA, Dougan G, Kingsley RA, Heyderman RS, Gordon MA (2012) Invasive
732 non-typhoidal salmonella disease: an emerging and neglected tropical disease in
733 Africa. Lancet 379: 2489-2499.

734 9. Ao TT, Feasey NA, Gordon MA, Keddy KH, Angulo FJ, et al. (2015) Global burden of
735 invasive nontyphoidal salmonella disease, 2010(1). Emerg Infect Dis 21.

736 10. Nuccio SP, Baumler AJ (2015) Reconstructing pathogen evolution from the ruins.
737 Proc Natl Acad Sci U S A 112: 647-648.

738 11. Nuccio SP, Baumler AJ (2014) Comparative analysis of Salmonella genomes
739 identifies a metabolic network for escalating growth in the inflamed gut. MBio 5:
740 e00929-00914.

741 12. Parsons BN, Humphrey S, Salisbury AM, Mikoleit J, Hinton JC, et al. (2013) Invasive
742 non-typhoidal Salmonella typhimurium ST313 are not host-restricted and have an
743 invasive phenotype in experimentally infected chickens. PLoS Negl Trop Dis 7:
744 e2487.

745 13. Kingsley RA, Msefula CL, Thomson NR, Kariuki S, Holt KE, et al. (2009) Epidemic
746 multiple drug resistant Salmonella Typhimurium causing invasive disease in sub-
747 Saharan Africa have a distinct genotype. Genome Res 19: 2279-2287.

748 14. Okoro CK, Barquist L, Connor TR, Harris SR, Clare S, et al. (2015) Signatures of
749 adaptation in human invasive Salmonella Typhimurium ST313 populations from
750 sub-Saharan Africa. PLoS Negl Trop Dis 9: e0003611.

751 15. Herrero-Fresno A, Wallrodt I, Leekitcharoenphon P, Olsen JE, Aarestrup FM, et al.
752 (2014) The role of the st313-td gene in virulence of Salmonella Typhimurium
753 ST313. PLoS One 9: e84566.

754 16. Ramachandran G, Perkins DJ, Schmidlein PJ, Tulapurkar ME, Tennant SM (2015)
755 Invasive Salmonella Typhimurium ST313 with naturally attenuated flagellin elicits
756 reduced inflammation and replicates within macrophages. PLoS Negl Trop Dis 9:
757 e3394.

758 17. Okoro CK, Kingsley RA, Connor TR, Harris SR, Parry CM, et al. (2012)
759 Intracontinental spread of human invasive Salmonella Typhimurium
760 pathovariants in sub-Saharan Africa. Nat Genet 44: 1215-1221.

761 18. Allard MW, Luo Y, Strain E, Pettengill J, Timme R, et al. (2013) On the evolutionary
762 history, population genetics and diversity among isolates of Salmonella
763 Enteritidis PFGE pattern JEGX01.0004. PLoS One 8: e55254.

764 19. den Bakker HC, Allard MW, Bopp D, Brown EW, Fontana J, et al. (2014) Rapid
765 whole-genome sequencing for surveillance of Salmonella enterica serovar
766 enteritidis. Emerg Infect Dis 20: 1306-1314.

767 20. Taylor AJ, Lappi V, Wolfgang WJ, Lapierre P, Palumbo MJ, et al. (2015)
768 Characterization of Foodborne Outbreaks of Salmonella enterica Serovar
769 Enteritidis with Whole-Genome Sequencing Single Nucleotide Polymorphism-
770 Based Analysis for Surveillance and Outbreak Detection. J Clin Microbiol 53:
771 3334-3340.

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

775 22. Rodriguez I, Guerra B, Mendoza MC, Rodicio MR (2011) pUO-SeVR1 is an
776 emergent virulence-resistance complex plasmid of Salmonella enterica serovar
777 Enteritidis. J Antimicrob Chemother 66: 218-220.

778 23. Rodriguez I, Rodicio MR, Guerra B, Hopkins KL (2012) Potential international spread
779 of multidrug-resistant invasive Salmonella enterica serovar enteritidis. Emerg
780 Infect Dis 18: 1173-1176.

781 24. Cheng L, Connor TR, Siren J, Aanensen DM, Corander J (2013) Hierarchical and
782 spatially explicit clustering of DNA sequences with BAPS software. Mol Biol Evol
783 30: 1224-1228.

784 25. Rodrigue DC, Tauxe RV, Rowe B (1990) International increase in Salmonella
785 enteritidis: a new pandemic? Epidemiol Infect 105: 21-27.

786 26. Magnet S, Courvalin P, Lambert T (1999) Activation of the cryptic aac(6')-Iy
787 aminoglycoside resistance gene of Salmonella by a chromosomal deletion
788 generating a transcriptional fusion. J Bacteriol 181: 6650-6655.

789 27. Zheng J, Pettengill J, Strain E, Allard MW, Ahmed R, et al. (2014) Genetic diversity
790 and evolution of Salmonella enterica serovar Enteritidis strains with different
791 phage types. J Clin Microbiol 52: 1490-1500.

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

797 29. Drummond AJ, Rambaut A (2007) BEAST: Bayesian evolutionary analysis by
798 sampling trees. BMC Evol Biol 7: 214.

799 30. Thomson N, Baker S, Pickard D, Fookes M, Anjum M, et al. (2004) The role of
800 prophage-like elements in the diversity of Salmonella enterica serovars. J Mol
801 Biol 339: 279-300.

802 31. Rotger R, Casadesus J (1999) The virulence plasmids of Salmonella. Int Microbiol 2:
803 177-184.

804 32. Thomson NR, Clayton DJ, Windhorst D, Vernikos G, Davidson S, et al. (2008)
805 Comparative genome analysis of Salmonella Enteritidis PT4 and Salmonella
806 Gallinarum 287/91 provides insights into evolutionary and host adaptation
807 pathways. Genome Res 18: 1624-1637.

808 33. Cooke FJ, Wain J, Fookes M, Ivens A, Thomson N, et al. (2007) Prophage
809 sequences defining hot spots of genome variation in Salmonella enterica serovar
810 Typhimurium can be used to discriminate between field isolates. J Clin Microbiol
811 45: 2590-2598.

812 34. Shah DH, Zhou X, Kim HY, Call DR, Guard J (2012) Transposon mutagenesis of
813 Salmonella enterica serovar Enteritidis identifies genes that contribute to
814 invasiveness in human and chicken cells and survival in egg albumen. Infect
815 Immun 80: 4203-4215.

816 35. Parkhill J, Dougan G, James KD, Thomson NR, Pickard D, et al. (2001) Complete
817 genome sequence of a multiple drug resistant Salmonella enterica serovar Typhi
818 CT18. Nature 413: 848-852.

819 36. Thiennimitr P, Winter SE, Winter MG, Xavier MN, Tolstikov V, et al. (2011) Intestinal
820 inflammation allows Salmonella to use ethanolamine to compete with the
821 microbiota. Proc Natl Acad Sci U S A 108: 17480-17485.

822 37. Kingsley RA, Kay S, Connor T, Barquist L, Sait L, et al. (2013) Genome and
823 transcriptome adaptation accompanying emergence of the definitive type 2 host-
824 restricted Salmonella enterica serovar Typhimurium pathovar. MBio 4: e00565-
825 00513.

826 38. Dhawi AA, Elazomi A, Jones MA, Lovell MA, Li H, et al. (2011) Adaptation to the
827 chicken intestine in Salmonella Enteritidis PT4 studied by transcriptional analysis.
828 Vet Microbiol 153: 198-204.

829 39. Langridge GC, Fookes M, Connor TR, Feltwell T, Feasey N, et al. (2015) Patterns of
830 genome evolution that have accompanied host adaptation in Salmonella. Proc
831 Natl Acad Sci U S A 112: 863-868.

832 40. Feasey NA, Everett D, Faragher EB, Roca-Feltrer A, Kang'ombe A, et al. (2015)
833 Modelling the Contributions of Malaria, HIV, Malnutrition and Rainfall to the
834 Decline in Paediatric Invasive Non-typhoidal Salmonella Disease in Malawi.
835 PLoS Negl Trop Dis 9: e0003979.

836 41. Gordon MA, Banda HT, Gondwe M, Gordon SB, Boeree MJ, et al. (2002) Non-
837 typhoidal salmonella bacteraemia among HIV-infected Malawian adults: high
838 mortality and frequent recrudescence. AIDS 16: 1633-1641.

839 42. Kotloff KL, Nataro JP, Blackwelder WC, Nasrin D, Farag TH, et al. (2013) Burden
840 and aetiology of diarrhoeal disease in infants and young children in developing
841 countries (the Global Enteric Multicenter Study, GEMS): a prospective, case-
842 control study. Lancet 382: 209-222.

843 43. Kariuki S, Onsare RS (2015) Epidemiology and Genomics of Invasive Nontyphoidal
844 Salmonella Infections in Kenya. Clin Infect Dis 61 Suppl 4: S317-324.

845 44. Harris SR, Cartwright EJ, Torok ME, Holden MT, Brown NM, et al. (2013) Whole-
846 genome sequencing for analysis of an outbreak of meticillin-resistant
847 Staphylococcus aureus: a descriptive study. Lancet Infect Dis 13: 130-136.

848 45. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, et al. (2009) The Sequence
849 Alignment/Map format and SAMtools. Bioinformatics 25: 2078-2079.

850 46. Holt KE, Parkhill J, Mazzoni CJ, Roumagnac P, Weill FX, et al. (2008) High-
851 throughput sequencing provides insights into genome variation and evolution in
852 Salmonella Typhi. Nat Genet 40: 987-993.

853 47. Croucher NJ, Harris SR, Fraser C, Quail MA, Burton J, et al. (2011) Rapid
854 pneumococcal evolution in response to clinical interventions. Science 331: 430-
855 434.

856 48. Stamatakis A (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic
857 analyses with thousands of taxa and mixed models. Bioinformatics 22: 2688-
858 2690.

859 49. Ashton P, Peters T, Tewolde R, Day M, Doumith M et al. (2015) Revolutionising
860 Public Health Reference Microbiology using Whole Genome Sequencing:
861 Salmonella as an exemplar. bioRxiv.

862 50. Drummond AJ, Suchard MA, Xie D, Rambaut A (2012) Bayesian phylogenetics with
863 BEAUti and the BEAST 1.7. Mol Biol Evol 29: 1969-1973.

864 51. Assefa S, Keane TM, Otto TD, Newbold C, Berriman M (2009) ABACAS: algorithm-
865 based automatic contiguation of assembled sequences. Bioinformatics 25: 1968-
866 1969.

867 52. Carver TJ, Rutherford KM, Berriman M, Rajandream MA, Barrell BG, et al. (2005)
868 ACT: the Artemis Comparison Tool. Bioinformatics 21: 3422-3423.

869 53. Zhou Y, Liang Y, Lynch KH, Dennis JJ, Wishart DS (2011) PHAST: a fast phage
870 search tool. Nucleic Acids Res 39: W347-352.

871 54. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, et al. (2015) Roary: Rapid
872 large-scale prokaryote pan genome analysis. Bioinformatics.

873 55. Kado CI, Liu ST (1981) Rapid procedure for detection and isolation of large and
874 small plasmids. J Bacteriol 145: 1365-1373.

875 56. Zankari E, Hasman H, Cosentino S, Vestergaard M, Rasmussen S, et al. (2012)
876 Identification of acquired antimicrobial resistance genes. J Antimicrob Chemother
877 67: 2640-2644.

878 57. Li W, Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets
879 of protein or nucleotide sequences. Bioinformatics 22: 1658-1659.

880 58. Tracy BS, Edwards KK, Eisenstark A (2002) Carbon and nitrogen substrate
881 utilization by archival Salmonella typhimurium LT2 cells. BMC Evol Biol 2: 14.

882 59. Croucher NJ, Mitchell AM, Gould KA, Inverarity D, Barquist L, et al. (2013) Dominant
883 role of nucleotide substitution in the diversification of serotype 3 pneumococci
884 over decades and during a single infection. PLoS Genet 9: e1003868.

885 60. Karp PD, Paley SM, Krummenacker M, Latendresse M, Dale JM, et al. (2010)
886 Pathway Tools version 13.0: integrated software for pathway/genome informatics
887 and systems biology. Brief Bioinform 11: 40-79.

888 61. Salisbury AM, Bronowski C, Wigley P (2011) Salmonella Virchow isolates from
889 human and avian origins in England--molecular characterization and infection of
890 epithelial cells and poultry. J Appl Microbiol 111: 1505-1514.

891 62. Betancor L, Yim L, Fookes M, Martinez A, Thomson NR, et al. (2009) Genomic and
892 phenotypic variation in epidemic-spanning Salmonella enterica serovar Enteritidis
893 isolates. BMC Microbiol 9: 237.

894

895

896

Tables

Table 1: Summary of metadata (n) by region in numbers

Region	Total	Site of isolation			Antimicrobial resistance phenotype				
		Human Invasive	Human non-invasive	Food/Animal/ Environment	Drug susceptible	Resistant to 1-2 1st line	MDR*	Fluoroquinolone	ESBL†
Asia	11	5	5	1	0	0	0	0	0
Europe	61	0	16	24	2	0	0	0	0
South America	27‡	3	6	7	8	0	0	0	0
North Africa	12	9	1	1	9	0	0	2	0
Sub-Saharan Africa	353	269	22	7	99	64	149	0	3
Republic of South Africa	131	57	74	0	83	44	4	0	0

*Multidrug resistant: resistant to ≥3 antimicrobials

†Extended spectrum beta lactamase producing

‡Uruguay strains previously characterised by Betancor [62]

905 Table 2: Metadata summarised by clade

Major Clade/cluster	Site of isolation N (%)				Number (%) of antimicrobial resistance genes*		
	Human Invasive	Human non-invasive	Food/Animal/ Environment	Unknown	1-3	4-6	7-9
West African	61 (92)	1 (2)	0 (0)	4 (6)	22 (33)	9 (14)	35 (66)
Central/Eastern African	155 (93)	7 (4)	0 (0)	5 (3)	0 (0)	11 (7)	156 (93)
Global epidemic	94 (38)	95 (38)	31 (12)	30 (12)	243 (97)	7 (3)	0 (0)
Outlier cluster	51 (38)	36 (27)	27 (20)	20 (15)	128 (96)	3 (2)	3 (2)

906

907 *All isolates contained cryptic aminoglycoside acetyltransferase gene *aac(6')-ly*[26]

908

Figures

Figure 1: Maximum likelihood phylogeny of *S*. Enteritidis based on 675 isolates rooted to *S*. Gallinarum. There are 3 epidemic clades; 2 African epidemic clades and a global epidemic clade. Scale bar indicates nucleotide substitutions per site.

Figure 2: Differences in accessory genomes of 4 major clades. Approximate position of prophages in chromosome is depicted, although prophages are not drawn to scale

Figure 3: Heat map revealing changes in metabolic activity of Central/Eastern African clade isolate D7795 when compared to global epidemic isolate A1636 at 28 and 37˚C. The figure also displays whether there are corresponding mutations in genes related to the affected metabolic pathway. (NSSNP=Non-synonymous single nucleotide polymorphism, HDG = Hypothetically disrupted gene)

Figure 4: *Salmonella* isolation from a chick infection model demonstrates failure of Central/Eastern African clade isolate to invade chicken spleen (4A) and liver (4B) or to colonize chicken caeca (4C) at 7 days post infection (dpi) (n=24 at this time point) compared to the global epidemic clade. Numbers are expressed as colony forming units (CFU) per gram of tissue

Supplementary Data

Supplementary Table 1: Metadata associated with each individual strain including date, place, and source of isolation plus antimicrobial susceptibility data where known. Predicted antimicrobial resistance genes are also included

Supplementary Table 2: Full list of predicted antimicrobial resistance genes

Supplementary Table 3: List of pseudogenes identified in D7795 and confirmation of presence/absence across African clades

Supplementary Table 4: List of genes in both Central/Eastern and West African clades with non-synonymous SNPs present throughout both clades

Supplementary Table 5: Comparison of genomic degradation seen in African epidemic clade with that seen in *S*. Typhi and *S*. Gallinarum

Supplementary Table 6: Full list of phenotypic differences between an example of the Central/Eastern African clade (D7795) and an example of the global epidemic clade (A1636) at 37˚C and corresponding genetic differences

Supplementary Table 7: Full list of phenotypic differences between an example of the Central/Eastern African clade (D7795) and an example of the global epidemic clade (A1636) at 28˚C and corresponding genetic differences

Supplementary Figure 1: Maximum likelihood phylogeny placing representative isolates from current study within the context of the diversity of *S*. Enteritidis genomes in the PHE collection. Black taxa labels represent PHE 50-SNP cluster representatives, red taxa labels represent seventeen representative strains from this study.

Supplementary Figure 2: Maximum likelihood phylogeny with strains of known phage type highlighted, demonstrating the lack of genomic diversity captured by phage typing

Supplementary Figure 3: BEAST tree of Central/Eastern African Clade and West African Clades revealing estimated age of clades

Supplementary Figure 4: Histograms of dates and BEAST results (treeHeight) for the subsets of the outlier cluster (S3A and S3C) and global epidemic clade (S3B and S3D)

Supplementary Figure 5: Distribution of prophage regions across the isolate collection highlighted. Red indicates presence, blue absence. Gaps indicate isolates not sequenced at Sanger Institute

Supplementary Figure 6: Maximum likelihood phylogeny of *S*. Enteritidis plasmids

High throughput phenotyping protocol

Supplementary results