Feasey, N. A., Hadfield, J., Keddy, K. H., Dallman, T. J., Jacobs, J., Deng, X., Wigley, P., Barquist Barquist, L., Langridge, G. C., Feltwell, T., Mather, A. E., Fookes, M., Aslett, M., Msefula, C., Kariuki, S., Maclennan, C. A., Onsare, R. S., Weill, F.-X., Le Hello, S., ... Thomson, N. R. (2016). Distinct Salmonella Enteritidis lineages associated with enterocolitis in high-income settings and invasive disease in low-income settings. *Nature Genetics*, *48*(10), 1211-1217. https://doi.org/10.1038/ng.3644

Peer reviewed version

Link to published version (if available):
10.1038/ng.3644

The dual emergence of *Salmonella* Enteritidis as a cause of enterocolitis in high-income settings and invasive nontyphoidal Salmonella disease in low-income settings

Authors
Nicholas A Feasey
James Hadfield
Karen H. Keddy
Jan Jacobs
Xiangyu Deng
Paul Wigley
Gemma Langridge
Theresa Feltwell
David Harris
Simon R Harris
Alison E Mather
Lars Barquist
Maria Fookes
Chisomo Msefula
Sam Kariuki
Calman Maclennan
Robert Onsare
François X Weill
Simon Le Hello
Anthony M. Smith
Michael McClelland
Prereak Desai
Christopher Parry
John Cheesbrough
Neil French
Josephina Campos
Jose A Chabalgoity

Katie Hopkins

Tom J Humphrey

Tristan Cogan

Sharon Tennant

Kristen Bornstein

Myron Levine

Lizeth Lacharme-Lora

Dean Everett

Robert A Kingsley

Julian Parkhill

Robert S Heyderman

Gordon Dougan

Melita A Gordon †

Nicholas R Thomson †

Abstract

Background

There is an epidemiological paradox surrounding *Salmonella enterica* serovar Enteritidis. It is responsible for a global epidemic of poultry associated, self-limiting gastroenteritis in the developed world where it is normally considered to be a host generalist. In sub-Saharan Africa (SSA), however, it is a major cause of invasive nontyphoidal *Salmonella* disease in humans where it is associated with high mortality if untreated.

Methods

Strains of *S*. Enteritidis originating from 45 countries were selected on the basis of temporal and host diversity, and antimicrobial susceptibility profile. Isolates were whole genome sequenced and the phylogeny was reconstructed using maximum-likelihood modelling by RAxML. The ages of major lineages were estimated using BEAST. The diversity was further evaluated by defining the evolutionary flux described in the pangenome as well as high-throughput phenotyping and an *in vivo* infection model.

Results

The phylogeny reveals the existence of two novel clades of *S*. Enteritidis that are geographically restricted in their distribution to Africa and associated with bloodstream infection and drug-resistance. Each of these clades has an associated characteristic accessory genome, including the prophage repertoire and virulence plasmid. The genomes of the African lineages show signatures of genomic degradation, a hallmark of differential host adaptation. These findings were corroborated both by the Biolog™ phenotyping microarray and the chick infection model, which demonstrated that representatives of this African lineage do not colonize the chicken caecum nor invade the spleen unlike the 'classical' global epidemic clones of *S*. Enteritidis which are associated with poultry.

Conclusions

Our analysis confirms *S*. Enteritidis as a further example of a *Salmonella* serotype that has become a prominent cause of bloodstream infection in SSA and which displays differential host-adaption. A phylogenetically distinct lineage of *S*. Enteritidis, traditionally associated solely with the consumption of poultry, has in the last few decades exploited a new niche. In this setting this African lineage is associated with invasive disease and increasing antimicrobial resistance. Our data, not only highlight the emergence of these new epidemic lineages, but also provide the means to track them.

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 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 hygiene and vaccination, epidemic *S*. Enteritidis has been in decline in many settings such as the United Kingdom [3]. *S.* Enteritidis has also been extensively used since the 1890s as a rodenticide. It was developed at Institute Pasteur and named the "Danysz virus" and has been used globally since the 1890s. Although these Salmonella-based rodenticides were banned in the US, Germany and the UK in the 1920s, 1930s and 1960s respectively, S. Enteritidis is still produced as a rodenticide in Cuba, under the name Biorat® [4].

Serovars of *Salmonella* that cause enterocolitis in industrial settings are strongly associated with non-focal sepsis or invasive nontyphoidal *Salmonella* (iNTS) disease in sub-Saharan Africa (SSA), a clinical syndrome responsible for an estimated 681,000 deaths per year [5,6]. *Salmonella* serovars *S*. Enteritidis and *S*. Typhimurium are the two leading causes of iNTS disease in SSA [7]. In order to investigate whether there are distinct bacterial characteristics or adaptions, in addition to (or in response to) host susceptibilities such as advanced HIV infection, whole genome sequence (WGS) based investigations of epidemic strains of *S*. Typhimurium from SSA have been undertaken. These revealed a novel pathotype of multilocus sequence type (MLST) ST313, which differ from strains that cause enterocolitis in industrialised settings by showing patterns of genomic degradation likely to be associated with more invasive disease and differential host adaptation [8-12].

There have been limited reports of *S.* Enteritidis isolated from African patients living in Europe that are multidrug resistant and which display a distinct Phage Type (PT 42) [13,14]. The broadest study of the phylogeny of *S*. Enteritidis to date revealed 5 major lineages, but contained only two African strains [15]. *Salmonella* is a key example of a

bacterial genus in which there is a recognizable genomic signature that distinguishes between a gastrointestinal and extra-intestinal lifestyle [16] whereby functions required for escalating growth in an inflamed gut are lost when the lineage becomes locked into an invasive lifestyle moving away from the lumen of the gut [17].

We hypothesize from the available epidemiological data that there were distinct lineages of *S*. Enteritidis circulating in both the developed and developing world with manifestly different aetiologies, likely distinct routes of spread and responsible for different patterns of disease that will display the distinct genomic signatures characteristic of differential adaptation. To investigate this we set out to collect a highly diverse collection of *S*. Enteritidis strains and to compare them using the highest possible resolution typing methodology, whole genome sequence data.

Methods

Strain selection

Strains were selected on the basis of diversity. Factors taken into account in order to distinguish isolates were; date of original isolation, antimicrobial susceptibility pattern, site of original location, source (human [invasive vs stool], animal or environmental), Phage type (where available), and Multilocus Variable Number Tandem Repeat (MLVA) type (where available). *S.* Enteritidis P125109 (EMBL accession no. AM933172) ioalted from a poultry farm from the UK was used as a reference [18]. The full metadata are in Supplementary Table 1.

Sequencing, construction of phylogeny and comparative genomics

PCR libraries were prepared from 500ng of DNA as previously described [19]. Isolates were sequenced using Illumina machines (Illumina, San Diego, CA, USA) and 150 bp paired-end reads were generated. Phylogeny was based on single nucleotide polymorphisms (SNPs) in conserved regions of the genome; WGS data for each of the isolates was mapped to a pseudo molecule including the reference strain P125109 plus

its plasmid pSENV using SMALT (http://www.sanger.ac.uk/resources/software/smalt/: version 0.5.8). Phylogenetic modelling is based on the assumption of a single common ancestor, therefore variable regions, where horizontal genetic transfer occurs were excluded [20] [21]. A maximum likelihood (ML) phylogenetic tree was then built from the SNP alignments of the isolates using RAxML (version 7.0.4) [22]. The maximum-likelihood phylogeny was supported by 100 bootstrap pseudo-replicate analyses of the alignment data. Clades were predicted using Hierarchical Bayesian Analysis of Population Structure (hierBAPS)[23]. This process was repeated to construct the plasmid phylogeny, using reads that aligned to pSENV. Temporal reconstruction was performed using Bayesian Evolutionary Analysis Sampling Trees (BEAST: http://beast.bio.ed.ac.uk/)[24]. Clock models were selected by first running an uncorrelated, relaxed lognormal clock and seeing if the posterior of uncorrelated lognormal relaxed clock standard deviation included 0, which would indicate that a strict clock might be appropriate. Skyline, exponential and constant population models were trialed. Three chains of 100 million states were run in parallel, as well as a fourth chain without genomic data to examine the influence of the prior. All priors and operators were left as their defaults.

In order to gain a detailed insight into genomic differences, a single high quality isolate from Malawi, D7795, was aligned against P125109 using ABACAS and annotated [25] . Differences were manually curated against the reference using the Artemis Comparison Tool (ACT)[26]. 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 prophage regions, they were annotated using the phage search tool PHAST and manually curated [27].

Accessory genome

The pangenome for the dataset was predicted using Bio-Roary [28]. Genes were considered to be core to *S*. Enteritidis if present in ≥90% of isolates. The remaining genes were considered to be core to the clades predicted by HierBAPS if present in ≥75% if isolates from the clade. These genes were then manually curated using ACT to

search for their presence and position in reference genome P125109 or the improved draft assembly of representative isolates of each of the other clades if not present in P125109.

Plasmid biology

Plasmid DNA was extracted from isolate D7795 using the Kado & Liu method and separated by gel-electrophoresis to estimate the number and size of plasmids present [29]. 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 (unpublished). For novel regions, genes were predicted using GLIMMER and manual annotations applied based on homology searches against the public databases, using both BLASTn and FASTA.

High throughput phenotyping

The Biolog™ platform (http://www.biolog.com) enables the simultaneous quantitative measurement of a number of cellular phenotypes, and therefore the creation of a phenotypic profile under a variety of growth conditions [30]. Incubation and recording of phenotypic data was performed using an OmniLog plate reader. In these experiments, two replicates of D7795 were compared to two replicates of PT4-like strain A1636 at 28˚C and 37˚C to represent environmental and human temperatures. Biolog™ plates PM1-4 and 9 (Carbon source [PM1,PM2], nitrogen source [PM3] and phosphor and sulphur source [PM4] metabolism and osmotic pressure [PM9]) were used. Each well was inoculated as described below, thereby testing 475 conditions at once (each plate has one negative control well). Plates were scanned every 15 min for 48 hours while incubated at 28°C and 37°C in air. Culture under anaerobic conditions was unavailable. Two paired replicates were performed for each of the two isolates.

After completion of the run, the signal data were compiled and analysed using the limma package in 'R' (www.R-project.org) as described previously [31]. A log-fold change of 0.5 controlled for a 5% false discovery rate was used as a cut-off for investigating a specific metabolite further using Pathway Tools[32] and the metabolic change was related to pseudogenes and non synonymous(NS)-SNPs in genes the genomes.

In vivo Infection Model

Two isolates were used in the animal models: *S*. Enteritidis P125109 and D7795. Unvaccinated commercial female egg-layer Lohmann Brown chicks were obtained from a commercial hatchery and housed in secure floor pens at a temperature of 25°C. Five chicks per strain were inoculated by gavage at 10 days (d) of age and received a dose of 10^8 *Salmonella* in a volume of 0.2 mL. Subsequently, four to five birds from each group were killed 3, 7 or 21 d post-infection (p.i.). At post mortem, the liver, spleen, and caecal contents were removed aseptically serially diluted and dispensed onto Brilliant Green agar (Oxoid) to quantify colony forming units (CFU) as described previously [33]. Statistical analysis was performed using SPSS, version 20 (IBM). Kruskal-Wallis was used to compare bacterial loads between infected groups.

Results

Strain collection

In total, 675 strains of *S.* Enteritidis isolated between 1948–2013 were sequenced. The collection originated from 45 countries and 6 continents (Table 1). 496/675 strains were from Africa, with 353 from SSA and a further 131 from the Republic of South Africa (RSA) and 12 from the Maghreb and Egypt. The full metadata are described in Supplementary Table 1.

Phylogeny

675 *S.* Enteritidis genomes and 1 *Salmonella enterica* serovar Gallinarum were mapped to reference isolate P125109, variable regions excluded and the remaining sites were screened for SNPs. This left an alignment file containing a total of 42,373 variable sites, from which ML-phylogeny was constructed using *S*. Gallinarum as an outgroup (Figure 1). HierBAPS was run over two rounds, providing excellent distinction between clades. The structure seen in the phylogeny of *S.* Enteritidis is striking for many reasons, with clear evidence of three epidemics, one which includes reference PT4-like isolate P125109, which we termed the 'global epidemic clade', 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 paraphyletic cluster from which the global epidemic clade emerged, and a further five smaller clades or clusters predicted by hierBAPs.

The clade containing isolates of Phage Type 4 and 1, associated with a global epidemic of poultry associated human enterocolitis, comprised 250 isolates from 28 countries, and these strains were isolated from across a 63-year period (1948-2013). 43 of these isolates were from Malawi and 82 from RSA. Of these 104/147 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 ESBL-producing isolates.

The global epidemic clade has emerged from a diverse cluster, which encompassed 134 isolates (Figure 1). 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 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 a German company [34]. There were also 41 isolates from RSA in this clade, where it has been a common cause of bloodstream infection.

There were two related, but phyletically and geographically distinct, epidemic clades associated with SSA. The Central/Eastern African clade included 167 isolates, all but 2 of which (from RSA) came from this region. 126/155 (82%) were MDR and 148/153 (97%) possessed between one and four antimicrobial resistance (AMR) phenotypes. 156/165 (95%) of these isolates were invasive. 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 significantly associated with drug resistance (4/66 strains susceptible to all antimicrobials tested) and 99% of these isolates were human-invasive (Supplementary Table 1).

The remaining 58 isolates included in this study were extremely diverse, phyletically, temporally and geographically. Only one was MDR and only two displayed any phenotypic resistance. 20 were human-invasive, and 6 were from stool. Three of these isolates were from stocks of rodenticide and these were phyletically remote from both global-epidemic and the two African epidemic clades.

Its apparent from the location of the archetypal reference isolate and archetypal phage-types in the phylogeny (Supplementary Figure 1) that the majority of *S*. Enteritidis studied previously belong to the global epidemic clade associated with developed countries and enterocolitis. The data presented here show the emergence of two

additional previously unrecognised lineages, largely restricted to Africa, that are strongly associated with MDR and invasive disease. To understand how recently these lineages emerged we used Bayesian Evolutionary Analysis by Sampling Trees (BEAST) to reconstruct the temporal history of the epidemic clades [35]. These data estimate the most recent common ancestor (MCRA) 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 MCRA of the global epidemic clade was 1918 (95% CrI: 1879-1942), 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 specialisation to their host bacterial species, whilst plasmids may confer a diverse array of virulence factors and AMR [36,37]. Therefore it is critical to evaluate the accessory genome in parallel with the core-genome. 622 sequenced genomes were used to create a pangenome, which yielded a core genome comprises 4,076 predicted genes present in ≥90% isolates and which carries all 12 *Salmonella* Pathogenicity Islands encoding functions essential for virulence across most pathogenic salmonellae as well as all 13 fimbrial operons found in the P125109 reference strain and required for colonisation and interaction with the host. The accessory genome consisted of 14,015 predicted genes present in <90% of the isolates. The Of the accessory genes, 324 genes were highly conserved across all of the epidemic clades and were almost all associated with the acquisition or loss of MGEs such as prophage or plasmids. These 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 [9,38].

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 2. 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 previously shown to be essential for invasion of chicken ova and mice [39]. There were a further 39 genes conserved in the global epidemic and clade 5 which were absent from both African clades, corresponding to region of difference (ROD) 21 [18]. The Central/Eastern Africa clade contained 77 predicted genes that were absent from the other clades. 33 were associated with the virulence plasmid and 40 were associated with a novel, Fels-2 like prophage region (Fels-BT). The West African clade had only 15 distinct 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 is degenerate in the reference strain P125109.

The *S.* Enteritidis plasmid is the smallest of the *Salmonella* virulence plasmids at 58 kb and is unusual in that it contains an incomplete set of *tra* genes responsible for horizontal transfer, by conjugation, of plasmids into other recipients. The phylogeny of the *S*. Enteritidis virulence plasmid was reconstructed using reads that mapped to the *S.* Enteritidis reference virulence plasmid, pSENV. 120/675 (18%) genomes lacked pSENV. The plasmid phylogeny is strikingly similar to that of the chromosome suggesting that it is highly conserved and has evolved alongside the chromosome (Supplementary Figure 3).

Plasmids from the African clades were much larger than those held in the other clades at ~100 kb. This is consistent with previous reports [13,14]. 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 xxxx). pSEN-BT was seen to be composed of a backbone of pSENV with additional regions that are highly similar to sequenced fragments of an *S.* Enteritidis virulence plasmid (pUO-SeVR) isolated from an African patient presenting with MDR invasive *S.* Enteritidis in Spain [14]. Plasmid pSEN-BT carries nine AMR genes, plus additional genes associated with virulence and a toxin/antitoxin plasmid addiction

system. Of note, plasmids from the West African clade have a plasmid containing resistance gene chloramphenicol acetyl transferase A1, whereas the Central/Eastern African strains carry *cat*A2 and tetracycline resistance gene *tet*A4. Like pSENV, the African virulence plasmid contained an incomplete set of *tra* genes and so is not self-transmissible. These observations suggest that the evolution of the *S.* Enteritidis plasmid mirrors that of the chromosome; it is not a 'novel' plasmid, but in different SSA locations has acquired different AMR genes in order to adapt to local selection pressures.

Multiple signatures of differential host adaptation

It is an emerging phenomenon 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 enteric to a more invasive lifestyle [9,18,40]. Accordingly we have looked for similar evidence within a representative example of a MDR, invasive, Central/Eastern African clade strain; strain 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).

In total there were 43 putative HDGs, many of which are involved in anaerobic metabolic pathways that enable *Salmonella* to utilize inflammation-derived nutrients core metabolic pathways, as part of an enteric lifestyle including cobalamine and propanediol utilisation. There is further evidence of reductive evolution in D7795 in the form of 363 genes containing NS-SNPS (which change the amino acid sequence), many of which are in genes within the same pathways as the HDGs (see Supplementary Results for detailed description). Supplementary Table 2 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 seen in D7795 is yet another signal of differential host adaptation seen in

both *S*. Typhi [40] and *S*. Gallinarum [18]. Many of these genes will be functional, but it is likely that some have become HDGs and that others have altered functions.

High-throughput phenotyping differences

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 growth, in the form of significantly different signal values. The entire output is graphically displayed in Figures 3.A (carbon sources) and 3.B (nitrogen, phosphorous and sulphur). Evaluation using Pathway Tools software revealed that 14/27 (52%) of pathways with evidence of decreased metabolic activity at 28°C in D7795 also displayed evidence of genomic degradation, and 12/30 (40%) of pathways with evidence of decreased metabolic activity at 37°C in D7795 also displayed evidence of genomic degradation.

Instances of reduced activity in a Central/Eastern African strain compared to a global epidemic strain, many of which have a corresponding signature of genomic degradation, included dulcitol and glycolic acid in the glycerol degradation pathway, propionic acid in the propanediol pathway and ethylamine and ethynolamine. These are all B12 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 [41] and HDGs relating to allantoin have been noted in *S.* Typhimurium ST313 [9]. The full list of differences is detailed in Supplementary Table 3 and 4. This is a further signal of decreased gut anaerobic metabolism.

In vivo infection model shows evolutionary divide in host range between key lineages

Given the genotypic and phenotypic differences observed in the global and African clades we predicted that these lineages would have differing infection phenotypes. We compared the infection profile of a member of the Central/Eastern African (D7795) to

the reference, global epidemic strain P125109 in the natural avian host. The groups infected with P125109 showed mild hepatosplenomegaly consistent with infection by this *Salmonella* serovar and caecal colonization. In contrast, the "African" strain failed to colonise the caeca or invade either liver or spleen. This reduced colonization phenotype, consistent with the differences in the accessory genome, suggests an entirely different transmission route for the "African" clades (Figure 4A-C). It also suggests that in the chick infection model of "African" *S*. Enteritidis, genomic degradation is a signal of differential host adaptation and not differential invasiveness. This is in marked contrast to the behavior of *S*. Typhimurium ST313, which is more invasive in a chick infection model [8].

Discussion

S. Enteritidis appears to be one of the most successful Salmonella lineages seen so far in its ability to adapt to different niches as and when opportunities for specialization have presented themselves. Langridge et al recently evaluated the Enteritidis/Gallinarum/Dublin lineage of *Salmonella*, revealing the nature and order of events necessary for both broad host-range and for host restriction [42]. In this study, we have highlighted the human niche plasticity of *S*. Enteritidis, providing evidence of three distinct epidemics of human disease, plus a further seven clades that demonstrate the huge reservoir of diversity amongst *S.* Enteritidis from which future human epidemics might emerge.

Two African epidemics are both significantly associated with a novel prophage repertoire, an expanded, MDR virulence plasmid and the genomic degradation that is the signature of host-restriction in multiple *Salmonella* serotypes [9,18,40]. This genomic degradation is too concentrated in pathways specifically associated with an enteric lifestyle to have arisen through a stochastic system; instead a selective advantage through exploiting a specific niche is implicated. It is noteworthy that both the MDR Central/Eastern African and the drug susceptible global epidemic clade have featured as prominent causes of iNTS disease at the same time in Blantyre, Malawi, consistent with the hypothesis that the two clades occupy different ecological niches outside the human host.

The evolution of the *S*. Enteritidis virulence plasmid is intriguing; pSENV is the smallest of the *Salmonella* virulence plasmids. The global epidemic clade has thrived without acquisition of drug resistance, whereas in SSA, the plasmid has nearly doubled in size. The absence of *tra* genes necessary for conjugal transfer either indicates that MDR status has evolved through acquisition of MGEs multiple times or through clonal expansion and vertical transmission of the plasmid to progeny. The available data suggest that former scenario has happened twice, once in West

Africa, and once in Central/Eastern Africa. The sequential acquisition of drug resistance in the two clades from SSA suggests that this has been critical to the expansion of these lineages/ecotypes, as appears to have been the case with ST313 strains of *S*. Typhimurium [43,44].

Phenotyping of these strains is consistent with the signatures of host-adaptation seen in other *Salmonella* serovars predicted in the genome of the "African" *S*. Enteritidis, for example revealing reduced activity in pathways for which there is no obvious signature in the genome i.e. allantoin. It is possible that these pathways have been affected by more subtle changes in the genome, for example changes in transcriptional start sites, the positions of which are extremely hard to predict from primary sequence alone. The chick infection model in which African *S*. Enteritidis does not colonise chick caeca, a feature critical to the emergence of PT4-like *S*. Enteritidis, confirms the different host-range and supports the hypothesis of differing ecological niches for the clades.

The most important question posed by this study is why have two clades of *Salmonella* from serotypes normally considered to be weakly invasive emerged in Africa? HIV is highly likely to have been an important factor, but in contrast to the picture seen with *S*. Typhimurium, fully susceptible *S.* Enteritidis isolates from the global epidemic clade have continued to cause disease alongside the novel clade. In Malawi, where there has been robust, long-term bloodstream infection (BSI) surveillance, serotype replacement of classical/susceptible *S.* Enteritidis by the African clade did not occur. This strongly suggests that the two clades are not in competition outside of the human host, instead there are three alternative possibilities:

1. That they occupy different ecological niches and are indeed different "ecotypes"
2. They share the same niche without detriment to transmission
3. They demonstrate symbiosis within the same niche

This study provides compelling evidence of a need to understand what these ecological niches might be, and what the transmission pathways of African/invasive Salmonellae are.

Just as the presence of these strains in Africa is striking, so too is their absence from industrial settings. There are three possible explanations for this; that they have not had the opportunity to spread, that they have a restricted human niche only found in Africa such as highly prevalent advanced HIV-infection, or that they are only competitive where antibiotics are unrestricted. It is notable, therefore, that there were only two "African" *S.* Enteritidis isolates amongst the strains submitted from RSA. This is surprising since South Africa has large numbers of people living in extreme poverty, high HIV rates and a large migrant population from SSA. It might be relevant that South Africa has better water, sanitation and hygiene (WASH) infrastructure relative to much of the rest of the continent and also has the resources to import food, favoring the emergence of the global epidemic clade. These data suggest these strains may occupy not just a different ecological niche, but a different socio-economic one too and as both WASH infrastructure improves and the HIV pandemic declines, their niche will disappear and they will die out. Of note, 83% of HIV-infected individuals eligible for antiretroviral therapy in Malawi now receive it, and in Blantyre, Malawi the number of cases of iNTS disease due to MDR *S*. Enteritidis has declined from 237 in 2001 to zero in 2014 (Feasey CID 2015 in press).

The principal limitation of this study is that it was retrospective in design and restricted to the laboratories that culture and store bacterial isolates and there was strain-selection bias towards invasive isolates. Of note, however the Global Enteric Multicenter Study (GEMS) has recently reported that *Salmonella* serotypes are not a common cause of moderate to severe diarrhoea in African children less than 5-years of age [45], which supports the hypothesis that these strains are intrinsically more invasive in susceptible individuals.

In Summary, two clades of *S.* Enteritidis, which are rarely seen outside of Africa, have emerged and have different phenotypes from 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 evolutional history of *S.* Typhimurium ST313. They are likely to have different ecologies and/or host ranges, and are strongly associated with drug resistance have caused epidemics of BSI in at least three settings in SSA, yet are rarely responsible for disease in South Africa. An investigation into the environmental reservoirs and transmission of these pathogens is urgently required.

References

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

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

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

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

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

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

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

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

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

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

11. Ramachandran S, Clifton IJ, Collyns TA, Watson JP, Pearson SB (2005) The treatment of spinal tuberculosis: a retrospective study. Int J Tuberc Lung Dis 9: 541-544.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

34. Inns T, Lane C, Peters T, Dallman T, Chatt C, et al. (2015) 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 20.

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

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

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

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

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

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

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

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

43. Gordon MA, Graham SM, Walsh AL, Wilson L, Phiri A, et al. (2008) Epidemics of invasive Salmonella enterica serovar enteritidis and S. enterica Serovar typhimurium infection associated with multidrug resistance among adults and children in Malawi. Clin Infect Dis 46: 963-969.

44. (2001) First randomised trial of treatments for pulmonary disease caused by M avium intracellulare, M malmoense, and M xenopi in HIV negative patients: rifampicin, ethambutol and isoniazid versus rifampicin and ethambutol. Thorax 56: 167-172.

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

Tables

Table 1: Summary of metadata

Region	Total	Antimicrobial susceptibility					Site of isolation		
		Drug susceptible	Resistant to 1-2 1st line	MDR	FQ	ESBL	Normally sterile site	Food/enviro	Stool/non-invasive
Asia	11						5	1	5
Europe	61	2	0	0	0	0	0	24	16
North Africa	12	9	0	0	2	0	9	1	1
South America	27	8	0	0	0	0	3	7	6
Republic of South Africa	131	83	44	4	0	0	57	0	74
Sub-Saharan Africa	353	99	64	149	0	3	269	7	22
USA	75						4	29	19

Figures (see attached power point slide set)

Figure 1: Maximum likelihood phylogeny of *S*. Enteritidis rooted to *S*. Gallinarum. There are 3 epidemic clades, including clear evidence of 2 African epidemic clades and a global epidemic clade

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 metabolic activity of isolates global epidemic clade (represented by isolate A1636) and Central/Eastern African clade (represented by isolate D7795) in response to 190 carbon sources (3A, below) and to 95 nitrogen (left) and 95 sulphur & phosphorous sources (right) (3B next slide). This is a graphical representation of the level of reduction of tetrazolium blue in the presence of each substrate in each well. Reduction is represented a coloured bar (see key above) and has been averaged across pairs of replicates

Figure 4: Salmonella isolation from chick infection model demonstrates failure of Central/Eastern African clade isolate to invade chicken spleen (4A) or liver (4B) or to colonise chicken caeca (4C) at 7 days post infection (dpi) unlike the global epidemic clade

Supplemental Data:
Supplemental 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
Supplemental Table 2: Comparison of genomic degradation seen in African epidemic clade with that seen in *S*. Typhi and *S*. Gallinarum

Supplemental Table 3: Full list of phenotypic differences between an example of the Central/Eastern African clade (D7795) and an example of the global epidemic clades (A1636) at 37˚C and corresponding genetic differences

Supplemental Table 4: Full list of phenotypic differences between an example of the Central/Eastern African clade (D7795) and an example of the global epidemic clades (A1636) at 28˚C and corresponding genetic differences

Supplemental Figure 1: ML-phylogeny with strains of known phage type highlighted, demonstrating the lack of genomic diversity captured by phage typing

Supplemental Figure 2: Distribution of prophage regions across the strain collection highlighted. Red indicates presence, blue absence

Supplemental Figure 3: Maximum likelihood phylogeny of S. Enteritidis plasmids mirrors that of the chromosome

Supplemental Figures X-Y:BEAST trees

Supplementary methods and results

High throughput phenotyping

The Biolog™ platform (http://www.biolog.com) enables the simultaneous quantitative measurement of a number of cellular phenotypes, and therefore the creation of a phenotypic profile under a variety of growth conditions [1]. Incubation and recording of phenotypic data was performed using an OmniLog plate reader. In these experiments, two replicates of D7795 were compared to two replicates of a PT4 like strain at 28˚C and 37˚C to represent environmental and human temperatures. Biolog™ plates PM1-4 and 9 (Carbon source [PM1,PM2], nitrogen source [PM3] and phosphor and sulphur source [PM4] metabolism and osmotic pressure [PM9]) were used. Each well was inoculated as described below, thereby testing 475 conditions at once (each plate has one negative control well).

The isolates were cultured overnight on LB-agar at 37°C in air to exclude contamination [2]. Colonies were scraped off plates and dispensed into IF-0a solution (Biolog) to a cell density corresponding to 81% transmittance. For each plate used, 880 µL of this cell suspension was added to 10 mL IF-10b GP/GP solution (Biolog) and 120 µL dye mix G (Biolog). This was then supplemented with a 1 mL solution of 7.5 mM D-ribose (Sigma), 2 mM magnesium chloride, 1 mM calcium chloride, 2 mM sodium pyrophosphate (Sigma), 25 µM L-arginine (Sigma), 25 µM L-methionine (Sigma), 25 µM hypoxanthine (Sigma), 10 µM lipoamide (Sigma), 5 µM nicotine adenine dinucleotide (Sigma), 0.25 µM riboflavin (Sigma), 0.005% by mass yeast extract (Fluka) and 0.005% by mass Tween 80 (Sigma). 100µl of this mixture was dispensed into each well on the assay plate. Plates were then allowed to equilibrate in air for 5 min prior to being sealed in airtight bags and loaded into the Omnilog machine (Biolog). Plates were scanned every 15 min for 48 hours while incubated at 28°C and 37°C in air. Culture under anaerobic conditions was unavailable. Two paired replicates were performed for each of the two isolates.

After completion of the run, the signal data were compiled and analysed. The data were first evaluated by examination of redox curves and discrepancies between replicates of the same isolate at the same temperature noted and repeated. The subsequent analysis was conducted using 'R' (http://www.R-project.org). As described previously[2], data were transformed to signal values, and the Bioconductor package limma was use to determine significant differences in tetrazolium reduction over time.

Significant differences in the data were defined as those changes with a Benjamini-Hochberg corrected p-value of less than 0.05 and a log fold change in signal value of at least 0.5. Significant changes were then manually reviewed by examining the graph of the tetrazolium reduction to ensure there was no obvious artefact i.e. two flat lines with one slight upward deviation at the end of the incubation period representing spontaneous reduction of the dye. The remaining differences were then visualised using Pathway Tools [4] and the metabolic change was related to pseudogenes and NS-SNPs in the genomes.

Chick Infection Models.

Two isolates were used in the animal models: *S*. Enteritidis P125109 and D7795. Unvaccinated commercial female egg-layer Lohmann Brown chicks were obtained from a commercial hatchery and housed in secure floor pens at a temperature of 25°C. Chicks were given ad libitum access to water and to a laboratory grade vegetable protein-based pellet diet (SDS). Five chicks per strain were inoculated by gavage at 10 days (d) of age and received a dose of 10^8 Salmonella in a volume of 0.2 mL. Subsequently, four to five birds from each group were killed 3, 7 or 21 d post infection (p.i.).

At post mortem, the liver, spleen, and caecal contents were removed aseptically from each bird and diluted 1:10 (wt/vol) in sterile PBS. Tissue and cecal content samples were then homogenized in a Colworth 80 BioMaster microstomacher (A. J. Seward & Co. Ltd.). Samples were serially diluted and dispensed onto Brilliant Green

agar (Oxoid) to quantify Salmonella numbers as described previously [5]. Statistical analysis was performed using SPSS, version 20 (IBM). Comparison of bacterial load between infected groups was tested using Kruskal Wallis, an equivalent nonparametric test to ANOVA, because the data were not distributed normally. Pathology scoring on hematoxylin-and-eosin stained sections at days 3, 5 and 7 following the methods described previously [6]. Statistical comparison between groups was performed using the Mann Whitney U-test.

Supplementary Results

Detailed comparison of an African MDR *S*. Enteritidis with P125109

Malawian invasive strain D7795 was selected for detailed analysis. D7795 was isolated from the blood of a Malawian child in 2000, and is MDR (resistant to amoxicillin, chloramphenicol, cotrimoxazole and tetracycline). Unsurprisingly, the draft genome sequence of D7795 closely resembles that of P125109. It does, however harbour a number of predicted pseudogenes, a different prophage repertoire and a substantially larger plasmid with an MDR region. In total D7795 shares 95.2% of its coding sequences with P125109 and P125109 shares 97.0% of its CDS with D7795. There are a further 363 genes in D7795 which have NS-SNPs with respect to P125109 (Table x).

Evidence of recent gene acquisition in P125109 has previously been highlighted [7]; of the 130 CDSs specific to P125109, compared to *S.* Gallinarum strain 287/91, those associated with the novel genomic region denoted region of difference (ROD) 4 and prophages ΦSE10, ΦSE14 and ΦSE20 appear to be recent acquisitions with no evidence of them ever being present in *S.* Gallinarum strain 287/91. The same is partially true of D7795, which contains no remnant or genomic 'scar' to suggest the two neighbouring variable regions from P125109, ROD21 or ΦSE20 were ever present. It does, however contain the degraded remnant of a prophage Gifsy 2-like

region, ΦSE12, although this region is even more degraded than in P125109. In contrast, P125109 contains the scar or remnant of a different Gifsy 2-like prophage region, which appears complete in D7795. These signature events are consistent with D7795 and P125109 having a common ancestor but them not being directly evolved from one another.

The core genome of P125109 is largely conserved in D7795. All 12 SPIs and all 13 fimbrial operons associated with P125109 are present. The genes within the SPIs are also intact, however fimbrial protein Y in the *fim* operon is a pseudogene, and 7 other fimbrial operon genes contain non-synonymous SNPs. The fimbrial operons are otherwise intact with respect to P125109 and D7795 has maintained the multiple functional fimbrial operons common to host promiscuous Salmonellae [8]. These differences are put in detailed context in Supplementary Table 2, adapted from Thomson et al [9].

The principle abnormality between the two chromosomes is in the prophage regions and genomic islands (RODs). P125019 contains prophage ΦSE20 and near to it, genomic island ROD 21. Prophage region ΦSE20 encodes 51 CDS in ~41 kbp. Potential cargo genes including the type-three secretion-system (TTSS) secreted effector protein *sseK3* and fragments of other TTSS effector genes including *sopE*, *sspH2* and *gogA* [9]. It has recently been demonstrated in both a mouse model [10] and cell-line model [11] that genes associated with region ΦSE20 are likely to be necessary for invasion in mice and chickens. The function of ROD21 is less clear. It is the largest genomic island in P125109 (~ 26 kb) and its structure has previously been described in detail [9]. Of note, it has a low G+C content and most of this island encodes paralogues of H-NS (*hnsB* – a down regulator of transcription) and/or an H-NS antagonist, *hnsT* [12-14].

In addition to these large differences, some of the prophage regions and RODs have under gone further degradation involving specific important genes. ΦSE12 and ΦSE12A are prophage remnants lying adjacent to each other and carry remnants of

the PhoPQ-activated genes *pagK* and *pagM*, an intact copy of *sopE* known to stimulate cytoskeletal re-organisation. This region has undergone further degradation in D7795, which does not carry gene *sodCI*, encoding a Cu/Zn superoxide dismutase known to be an important colonisation factor for *S.* Typhimurium [15]. Genomic island ROD 30 has lost the nickel/cobalt transporter *rcnA*. Both of these regions are likely to be ancient acquisitions as they are present in *S*. Typhimurium LT2.

D7795 contains three complete prophage regions that are absent from P125109, containing a total of 150 predicted CDS regions and including one Gifsy 2-like region, highly degraded in P125109, but retaining a further 10 genes in D7795. The three complete regions were annotated using Phast (see Table 8.3 for summary) and are closely related to phages Entero P2 (33.1 kb, 47 predicted CDS, including 3 cargo genes), *Yersinia* 413 (38.1 kb, 52 CDS, including 13 cargo genes) and a Gifsy 2-like prophage (33.1 kb, 52 CDS, including 5 cargo genes). The regions largely contain hypothetical genes or essential phage machinery, however the *Yersinia* 413 prophage carries a gene predicted to encode a leucine rich repeat (LRP) protein. These LRP-domains are protein interaction regions and have been implicated in suppression of gut-inflammation through inhibition of NF-κB-dependent gene expression by both *S*. Typhimurium and *Shigellae* [16].

Inspection of the draft assembly of D7795 reveals 43 putative pseudogenes (see Table 1 [summary] and Supplementary Table 3[full list]). Whilst this only represents 0.1% of the genome, it is a striking feature is how many are involved in core metabolic pathways, specifically in genes critical for an enteric lifestyle, and how many genes from the same pathway contain NS-SNPs. Some of them have catabolic roles in colonic mucus degradation, including SEN1434 (hexonate metabolism) and the sulphatase SEN 0035. Colonic mucus contains sugar acids, a source of nutrients that gut dwelling organisms must be able to utilize [17]. Several of the apparent pseudogenes are involved in anaerobic metabolism, including SEN3902, a glycerol dehydrogenase. Glycerol metabolism is necessary for anaerobic

growth of enterobacteria on non-fermentable sources. Dimethyl sulphoxide reductase, a gene necessary for the use of terminal electron acceptors in anaerobic metabolism, has also been lost, together with two other sulphur metabolism genes. Two genes involved in formate metabolism have become degenerate. There are a further four metabolic pseudogenes, including mono-pentose pathway (L-fuculose kinase [*fucK*]*)*.

Gene *pocR* is a pseudogene in D7795. This is one of the most important regulatory genes for the B12 synthesis and propanediol utilisation pathway [18], a key enteric-adaptation pathway. Many *Enterobacteriaceae* have lost the capacity to synthesise cobalamine and therefore to degrade 1,2-propanediol. However, Salmonella re-acquired a 40 kb metabolic island encoding both the *cbi* and *pdu* loci [19-21]. 1,2-propanediol is an important source of energy for *S.* Typhimurium, especially within the intracellular compartment [22]. Of note, the B12 transport gene *yncD* is also a pseudogene and a number of downstream genes in this pathway have acquired NS-SNPs including two tetrathionate and three propanediol utilisation genes, which suggests the possibility of degradation of this pathway. Degradation in this pathway is also seen in invasive, host-adapted serotype *S*. Gallinarum [9].

Like *S.* Typhi and *S.* Gallinarum 287/91, *S*. Enteritidis D7795 carries a number of mutations in a gene which is centrally involved with shedding - *shdA*. This gene is a surface-localized, fibronectin-binding protein whose expression is induced in vivo in the murine caecum [23,24] and is carried on a 25-kb genetic island named centisome 54 (CS54 island) in *S.* Typhimurium. Absence of this gene is associated with reduced faecal carriage and shedding of *S*. Typhimurium in mice, but not pigs [23] [25]. This gene is also a pseudogene in *S.* Typhi and *S.* Gallinarum and whilst it is intact in the African invasive *S*. Typhimurium ST313, the related gene *ratB* from CS54 has become a pseudogene [9,26]. As discussed above, there is one fimbrial gene (*fimY*), which is a pseudogene and one type III secretion system protein.

In addition to the metabolic signal suggestive of host adaptation, there are seven pseudogenes amongst membrane transport genes, including the two further genes involved in cobalamine metabolism. Cobalamine synthase incorporates Ni-Fe and Cobalt respectively, and it is interesting to note that D7795 has lost *rncA*, a high affinity Ni/Co transporter, through a deletion in ROD30 (see above)[27] and the gene encoding the cobalamine transporter *yncB*. Five other putative, uncharacterised transport genes are pseudogenes. A further seven membrane-protein associated genes have significant mutations.

Three DNA replication and repair genes carry mutations, including a *mutT*-family gene. Only one pseudogene lies in an amino acid catabolic or biosynthetic pathway, L-serine (*L*-serine and *L*-threonine catabolism), for which there is redundancy [28]. A further five hypothetical genes and one phage-protein gene have become pseudogenes. Table A6 in Appendix 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. Some of the overlapping traits are striking; including the loss of genes involved in common metabolic processes such as cobalamine and propanediol utilisation and electron transport acceptor function.

In addition to the large differences in terms of the variable region of the genome, there is further evidence of reductive evolution in D7795 in the form of 363 genes containing NS-SNPS (Table 8.5). The disproportionate clustering around membrane structures is yet another signal seen in host adaptation with both *S*. Typhi [29] and *S*. Gallinarum [9]. Many of these genes will be functional, but it is likely that some have become pseudogenes and that others have altered functions. Certainly there is evidence of clustering around pathways (Figure 8.5), for example the cobalamine biosynthetic pathway.

The MDR virulence plasmid of the Southern African *S*. Enteritidis clade

Plasmid extraction by Kado and Lui alkaline lysis, followed by gel-electrophoresis was performed on MDR isolate D7795 (Supplementary Figure 1). This revealed the presence of just one large plasmid, which was PacBio sequenced. This demonstrated a 100 kb plasmid, which, when aligned against finished P125109 plasmid pSENV revealed evidence of two deletions; the first contained genes *srgA, srgB and luxR*, part of the *lux* operon and the second was within gene pSENV0033, a putative integrase. All the other genes on pSENV including the virulence operon, were intact. The additional regions aligned against two sequenced fragments of an *S.* Enteritidis virulence plasmid (pUO-SeVR) isolated from an African patient presenting with MDR iNTS Enteritidis in Spain [30]. In total, nine antimicrobial resistance genes were identified, encoding resistance to amoxicillin (*bla*TEM-1), tetracycline (*tetR* and *tetA*), chloramphenicol (*cat*A2) the components of cotrimoxazole (*dfr*A7, *sul*1 and *sul*2) and aminoglycosides (*str*A and *str*B).

The plasmid contained a number of additional genes that might be associated with virulence and a toxin/antitoxin plasmid addiction system (Supplementary Table 4). These included *tir*, associated with gut wall attachment in enteropathogenic *E. coli*, and associated with virulence in 0157:H7 strain [31]; a *pecM* gene which is a membrane spanning efflux protein which may have drug efflux properties [32]; and *pncA* which is a nicotinamidase whose function is unclear, however it is essential to the pathogenesis of *Borrelia burgdorferi [33]*. Lastly genes *mucA* and *mucB* are present. The latter has been shown to be a lesion bypass polymerase and although their precise role is unclear, together they may repair damaged plasmid DNA, or facilitate mutagenesis. This would be beneficial for an MGE transmitted through a number of bacterial hosts, by allowing faster adaptation to foreign intracellular environments [34].

Like pSENV, the virulence plasmid contained an incomplete set of conjugal transfer genes (*tra* genes) and it was therefore predicted that this plasmid would not be capable of conjugal transfer. This was tested by conjugation experiments at 26˚C and 37˚C using the plasmid from *S*. Typhimurium A54560 [35] and IncHI1 plasmid

PHCM1 as positive controls and PT4-like *S*. Enteritidis A1636 as a negative control. As predicted, the plasmids of A1636 and D7795 failed to transfer, whilst the positive control successfully transferred.

Interpretation of high-throughput phenotyping

The findings from the pairs of replicates were compared using principal component analysis and found to be highly consistent. In total, there were 200 statistically significant differences between log-fold change in adjusted SVs between the 2 isolates; 97 at 28°C and 103 at 37°C. Following manual review of the signal from each well, 80 of the results were considered genuinely significant and pursued further. The entire output is graphically displayed in Figures x.A (PM1 and 2 [carbon sources]) and x.B (both PM 3 [nitrogen] and 4 [phosphorous and sulphur]).

Evaluation using Pathway Tools software to identify genes involved in each specific pathway with evidence of increased or decreased metabolic activity revealed that 14/27 (52%) of pathways with evidence of decreased metabolic activity at 28°C in D7795 also displayed evidence of genomic degradation, and 12/30 (40%) of pathways with evidence of decreased metabolic activity at 37°C in D7795 also displayed evidence of genomic degradation.

The two replicates of D7795 showed enhanced metabolism of three simple carbohydrates, including glucose, sucrose and lactulose, also of D-saccharic acid, D-alanine, mucic (galactaric) acid and formic acid. There were far more instances of reduced activity, many of which have a corresponding signature of genomic degradation, such as pseudogene formation or the presence of NS-SNPs in genes corresponding to the pathway. These included dulcitol and glycolic acid in the glycerol degradation pathway, propionic acid in the propanediol pathway and ethylamine and ethynolamine. Also there was reduced activity in response to three forms of butyric acid, a short chain fatty acid. This is another signature of loss of enteric adaptation, as the ability to metabolize short chain fatty acids is an

extremely useful trait for enteric adaptation [36]. The full list of differences is detailed in Supplementary Table 5.

There was no increased metabolism of any nitrogen sources in D7795 compared to A1636, however metabolism was reduced in several nitrogen sources (Supplementary Table 5). The most interesting sources noted to be down regulated were alloxan and allantoic acid. Allantoin can be found in the serum of birds, and is utilised as a carbon source during *S.* Enteritidis infection of chickens [37] and pseudogenes relating to allantoin have been noted in *S.* Typhimurium ST313 [26]. D7795 showed increased metabolism in the presence of 6% sodium chloride solution at both temperatures, but decreased metabolism in the presence of 4% sodium formate, which is consistent with the numerous mutations in formate catabolizing genes in D7795.

Table 8-1: Summary of differences between P12509 and D7795

Differences with respect to P125109	Variable regions with differences	Genes gained in D7795	Genes lost in D7795
Additional prophage regions	3	150	0
"Missing" prophage regions	2	0	79
Other insertions	3	25	
Other deletions	3	0	3
Likely pseudogenes	43		45
Plasmid		56	4
Total		+221	-131

Table 8-3: Summary of contents of prophages of D7795

	Prophage 1	Prophage2	Prophage3
Length	33.1	38.1	30.1
Closest relative	Entero 2 (83%)	Yersinia 413C	Gifsy1
GC	53%	52	51%
Related Salmonellae Prophage	RE_2010	RE_2010	Gifsy1
Total CDS	47	51	52
Phage proteins	44	34	46
Hypothetical proteins	3	13	5
Phage +hypothetical protein %	100	92	98

Table 8-4: Function categories of pseudogenes in D7795

Functional category	Number of pseudogenes
Core metabolism	17
Virulence, host adaptation	3
Membrane	7
Transporter	7
DNA reproduction/repair	3
Phage protein	1
Hypothetical genes	5

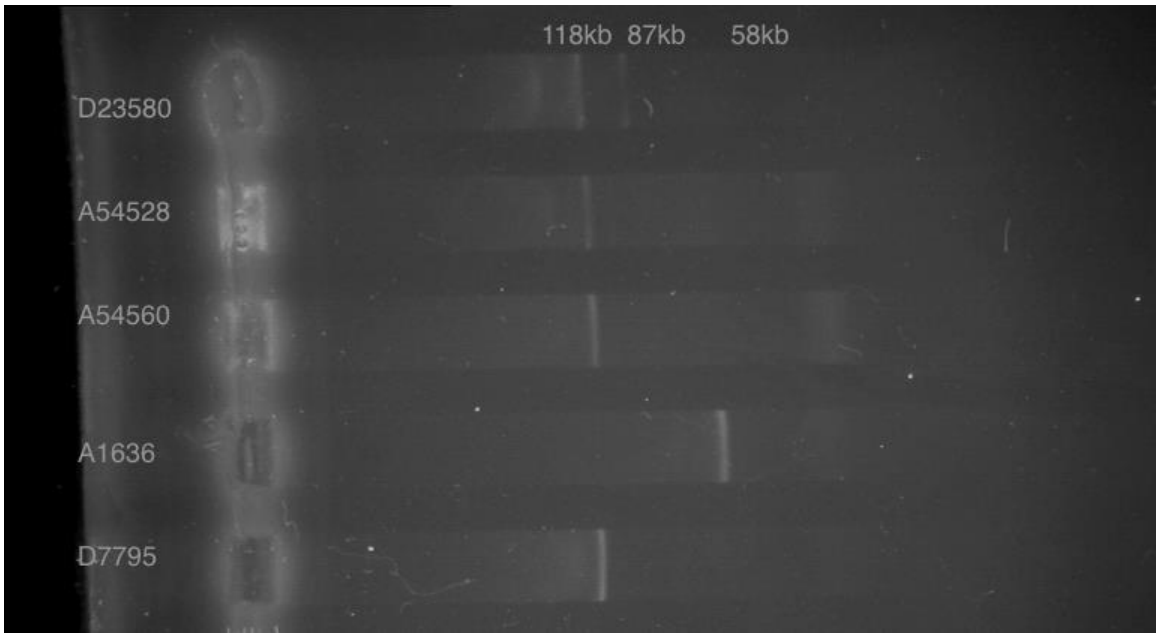
Table 8-5 Functional breakdown of genes containing NS-SNPs

Functional class	Number of genes
Replication and repair	16
Metabolism	41
Membrane/surface structures	122
Degradation of macromolecules	12
Degradation of small molecules	18
Regulators	67
Conserved hypothetical	83
Phage	4
Total	363

Table 8-6: Summary of phenotypic differences between A1636 and D7795. "Increased" and "decreased" refer to tetrazolium dye reduction by D7795 when compared to A1636

Condition (temp ^{o}C)		Number of conditions	
		28	**37**
Carbon sources (190)	Activity increased	7	8
	Activity decreased	13	20
Nitrogen sources (95)	Activity increased	0	0
	Activity decreased	13	10
Phosphorous & sulphur (95)	Activity increased	1	2
	Activity decreased	0	0
Osmotic stress (95)	Activity increased	3	2

	Activity decreased	1	0
Total differences		38	42



1. Tracy BS, Edwards KK, Eisenstark A (2002) Carbon and nitrogen substrate utilization by archival Salmonella typhimurium LT2 cells. BMC Evol Biol 2: 14.
2. Croucher NJ, Mitchell AM, Gould KA, Inverarity D, Barquist L, et al. (2013) Dominant role of nucleotide substitution in the diversification of serotype 3 pneumococci over decades and during a single infection. PLoS Genet 9: e1003868.
3. Homann OR, Cai H, Becker JM, Lindquist SL (2005) Harnessing natural diversity to probe metabolic pathways. PLoS Genet 1: e80.
4. Karp PD, Paley SM, Krummenacker M, Latendresse M, Dale JM, et al. (2010) Pathway Tools version 13.0: integrated software for pathway/genome informatics and systems biology. Brief Bioinform 11: 40-79.
5. Salisbury AM, Bronowski C, Wigley P (2011) Salmonella Virchow isolates from human and avian origins in England--molecular characterization and infection of epithelial cells and poultry. J Appl Microbiol 111: 1505-1514.
6. Parsons BN, Humphrey S, Salisbury AM, Mikoleit J, Hinton JC, et al. (2013) Invasive Non-Typhoidal Salmonella Typhimurium ST313 Are Not Host-

Restricted and Have an Invasive Phenotype in Experimentally Infected Chickens. PLoS Negl Trop Dis 7: e2487.
7. Kotakis C, Petropoulou Y, Stamatakis K, Yiotis C, Manetas Y (2006) Evidence for active cyclic electron flow in twig chlorenchyma in the presence of an extremely deficient linear electron transport activity. Planta 225: 245-253.
8. Townsend SM, Kramer NE, Edwards R, Baker S, Hamlin N, et al. (2001) Salmonella enterica serovar Typhi possesses a unique repertoire of fimbrial gene sequences. Infect Immun 69: 2894-2901.
9. Thomson NR, Clayton DJ, Windhorst D, Vernikos G, Davidson S, et al. (2008) Comparative genome analysis of Salmonella Enteritidis PT4 and Salmonella Gallinarum 287/91 provides insights into evolutionary and host adaptation pathways. Genome Res 18: 1624-1637.
10. Silva CA, Blondel CJ, Quezada CP, Porwollik S, Andrews-Polymenis HL, et al. (2012) Infection of mice by Salmonella enterica serovar Enteritidis involves additional genes that are absent in the genome of serovar Typhimurium. Infect Immun 80: 839-849.
11. Shah DH, Zhou X, Kim HY, Call DR, Guard J (2012) Transposon mutagenesis of Salmonella enterica serovar Enteritidis identifies genes that contribute to invasiveness in human and chicken cells and survival in egg albumen. Infect Immun 80: 4203-4215.
12. Park HS, Kim DW, Kim CC, Kim HK, Kim JS, et al. (1996) Induction chemotherapy with idarubicin plus N4-behenoyl-1-beta-D-arabinofuranosylcytosine in acute myelogenous leukemia: a newly designed induction regimen--a prospective, cooperative multicenter study. Semin Hematol 33: 24-29.
13. Navarre WW, Porwollik S, Wang Y, McClelland M, Rosen H, et al. (2006) Selective silencing of foreign DNA with low GC content by the H-NS protein in Salmonella. Science 313: 236-238.
14. Williamson HS, Free A (2005) A truncated H-NS-like protein from enteropathogenic Escherichia coli acts as an H-NS antagonist. Mol Microbiol 55: 808-827.
15. Figueroa-Bossi N, Uzzau S, Maloriol D, Bossi L (2001) Variable assortment of prophages provides a transferable repertoire of pathogenic determinants in Salmonella. Mol Microbiol 39: 260-271.
16. Haraga A, Miller SI (2003) A Salmonella enterica serovar typhimurium translocated leucine-rich repeat effector protein inhibits NF-kappa B-dependent gene expression. Infect Immun 71: 4052-4058.
17. Sweeney NJ, Laux DC, Cohen PS (1996) Escherichia coli F-18 and E. coli K-12 eda mutants do not colonize the streptomycin-treated mouse large intestine. Infect Immun 64: 3504-3511.
18. Bobik TA, Ailion M, Roth JR (1992) A single regulatory gene integrates control of vitamin B12 synthesis and propanediol degradation. J Bacteriol 174: 2253-2266.
19. Lawrence JG, Roth JR (1996) Evolution of coenzyme B12 synthesis among enteric bacteria: evidence for loss and reacquisition of a multigene complex.
. Genetics 142: 11-24.

20. Porwollik S, Wong RM, McClelland M (2002) Evolutionary genomics of *Salmonella*: gene acquisitions revealed by microarray analysis.
. Proc Natl Acad Sci U S A 99: 8956-8961.
21. Prentice MB, Cuccui J, Thomson N, Parkhill J, Deery E, et al. (2003) Cobalamin synthesis in Yersinia enterocolitica 8081. Functional aspects of a putative metabolic island. Adv Exp Med Biol 529: 43-46.
22. Klumpp J, Fuchs TM (2007) Identification of novel genes in genomic islands that contribute to Salmonella typhimurium replication in macrophages. Microbiology 153: 1207-1220.
23. Kingsley RA, Santos RL, Keestra AM, Adams LG, Baumler AJ (2002) Salmonella enterica serotype Typhimurium ShdA is an outer membrane fibronectin-binding protein that is expressed in the intestine. Mol Microbiol 43: 895-905.
24. Kingsley RA, Humphries AD, Weening EH, De Zoete MR, Winter S, et al. (2003) Molecular and phenotypic analysis of the CS54 island of Salmonella enterica serotype typhimurium: identification of intestinal colonization and persistence determinants. Infect Immun 71: 629-640.
25. Boyen F, Pasmans F, Donne E, Van Immerseel F, Morgan E, et al. (2006) The fibronectin binding protein ShdA is not a prerequisite for long term faecal shedding of Salmonella typhimurium in pigs. Vet Microbiol 115: 284-290.
26. Kingsley RA, Msefula CL, Thomson NR, Kariuki S, Holt KE, et al. (2009) Epidemic multiple drug resistant Salmonella Typhimurium causing invasive disease in sub-Saharan Africa have a distinct genotype. Genome Res 19: 2279-2287.
27. Rodrigue A, Effantin G, Mandrand-Berthelot MA (2005) Identification of rcnA (yohM), a nickel and cobalt resistance gene in Escherichia coli. J Bacteriol 187: 2912-2916.
28. Sawers G (1998) The anaerobic degradation of L-serine and L-threonine in enterobacteria: networks of pathways and regulatory signals. Arch Microbiol 171: 1-5.
29. Parkhill J, Dougan G, James KD, Thomson NR, Pickard D, et al. (2001) Complete genome sequence of a multiple drug resistant Salmonella enterica serovar Typhi CT18. Nature 413: 848-852.
30. Rodriguez I, Rodicio MR, Guerra B, Hopkins KL (2012) Potential international spread of multidrug-resistant invasive Salmonella enterica serovar enteritidis. Emerg Infect Dis 18: 1173-1176.
31. Bono JL, Keen JE, Clawson ML, Durso LM, Heaton MP, et al. (2007) Association of Escherichia coli O157:H7 tir polymorphisms with human infection. BMC Infect Dis 7: 98.
32. Rouanet C, Nasser W (2001) The PecM protein of the phytopathogenic bacterium Erwinia chrysanthemi, membrane topology and possible involvement in the efflux of the blue pigment indigoidine. J Mol Microbiol Biotechnol 3: 309-318.
33. Jewett MW, Jain S, Linowski AK, Sarkar A, Rosa PA (2011) Molecular characterization of the Borrelia burgdorferi in vivo-essential protein PncA. Microbiology 157: 2831-2840.

34. Goldsmith M, Sarov-Blat L, Livneh Z (2000) Plasmid-encoded MucB protein is a DNA polymerase (pol RI) specialized for lesion bypass in the presence of MucA', RecA, and SSB. Proc Natl Acad Sci U S A 97: 11227-11231.
35. Feasey NA, Cain AK, Msefula CL, Pickard D, Alaerts M, et al. (2014) Drug Resistance in Salmonella enterica ser. Typhimurium Bloodstream Infection, Malawi. Emerg Infect Dis 20: 1957-1959.
36. Keeney KM, Finlay BB (2011) Enteric pathogen exploitation of the microbiota-generated nutrient environment of the gut. Curr Opin Microbiol 14: 92-98.
37. Dhawi AA, Elazomi A, Jones MA, Lovell MA, Li H, et al. (2011) Adaptation to the chicken intestine in Salmonella Enteritidis PT4 studied by transcriptional analysis. Vet Microbiol 153: 198-204.