

1 **Evolution of *Salmonella enterica* serotype Typhimurium driven by**
2 **anthropogenic selection and niche adaptation**

3

4

5 Matt Bawn^{1,2}, Gaetan Thilliez¹, Mark Kirkwood¹, Nicole Wheeler³, Liljana Petrovska⁴,

6 Timothy J. Dallman⁵, Evelien M. Adriaenssens¹, Neil Hall² and Robert A. Kingsley^{1,6*}

7

8 ¹ Quadram Institute Biosciences, Norwich Research Park, Norwich, UK

9 ² Earlham Institute, Norwich Research Park, Norwich, UK

10 ³ Centre for Genomic Pathogen Surveillance, Wellcome Sanger Institute, Cambridge, UK

11 ⁴ Animal and Plant Health Agency, Addlestone, UK

12 ⁵ Gastrointestinal Bacteria Reference Unit, National Infection Service, Public Health

13 England, London, UK

14 ⁶ University of East Anglia, Norwich, UK

15

16

17 *Correspondence: rob.kingsley@quadram.ac.uk (R.A. Kingsley)

18

19

20

21
22

Abstract

23 *Salmonella enterica* serotype Typhimurium (*S. Typhimurium*) is a leading cause of
24 gastroenteritis and disseminated disease worldwide. Two *S. Typhimurium* strains (SL1344
25 and ATCC14028) are widely used to study host-pathogen interactions, yet genotypic
26 variation results in strains with diverse host range, pathogenicity and risk to food safety. A
27 robust fully parsimonious phylogenetic tree constructed from recombination purged variation
28 in the whole genome sequence of 131 diverse strains of *S. Typhimurium* revealed population
29 structure composed of two high order clades (α and β) and multiple subclades on extended
30 internal branches, that exhibited distinct signatures of host adaptation and anthropogenic
31 selection. Clade α contained a number of subclades composed of strains from well
32 characterized epidemics in domesticated animals, while clade β predominantly contained
33 subclades associated with wild avian species, with the notable exception of a subclade
34 containing the DT204/49 complex. The contrasting epidemiology of α and β strains was
35 reflected in a distinct distribution of antimicrobial resistance (AMR) genes, accumulation of
36 hypothetically disrupted coding sequences (HDCS), and signatures of functional
37 diversification associated with invasiveness of host adapted serotypes. Gene flux was
38 predominantly driven by acquisition, loss or recombination of prophage. The acquisition of
39 large genetic islands (SGI-1 and 4) was limited to two recent pandemic clones (DT104 and
40 monophasic *S. Typhimurium* ST34) in clade α . Together, our data are consistent with the
41 view that a broad host range common ancestor of *S. Typhimurium* diversified with clade α
42 lineages remained largely associated with multiple domesticated animal species, while clade
43 β spawned multiple lineages that underwent diversifying selection associated with adaptation
44 to various niches, predominantly in wild avian species.

45

46

Introduction

47

48 Bacteria of the genus *Salmonella* are a common cause of foodborne disease. Most of the
49 approximately 2500 serovars cause gastroenteritis in humans and other animals, while some
50 have evolved host adaptation associated with extra intestinal disseminated infections in
51 specific host species ¹. For example, *Salmonella enterica* serotype Typhimurium (*S.*
52 *Typhimurium*) and *S. Enteritidis* circulate in multiple vertebrate host species and cause food
53 borne infections in the human population resulting in an estimated 75 million cases and 27
54 thousand deaths from gastroenteritis worldwide ². *S. Typhi* and *S. Paratyphi A* circulate
55 exclusively in the human population and cause an estimated 2.5 million infections resulting
56 in 65 thousand deaths each year as a result of the disseminated disease typhoid and
57 paratyphoid disease ². Similarly, other serotypes evolved host adaptation to specific non-
58 human host species, such as *S. Gallinarum* with poultry, *S. Dublin* with cattle, and *S.*
59 *Choleraesuis* with pigs, where they are associated with disseminated infections ¹.

60

61 Although *S. Typhimurium* is considered to be a broad host range serotype, the
62 epidemiological record of *S. Typhimurium* phage types identifies several *S. Typhimurium*
63 pathovariants with distinct host range, pathogenicity and risk to food safety ^{3,4}. The
64 pathovariant commonly associated with this serotype, has a broad host range and is
65 associated with gastroenteritis in the human population. Such broad host range strains of *S.*
66 *Typhimurium* account for the majority of those isolated by public health surveillance in
67 England, presumably because they are common in many species of livestock and poultry, the
68 primary zoonotic reservoir for human infections ⁵. The epidemiological record of this
69 pathovariant is characterised by successive waves of dominant clones identified historically
70 by their phage type, that account for up to 60% of all human infections for several years,

71 before being replaced by a subsequent strains ⁶. Dominant clonal groups have been
72 characterized by strains of phage types DT9, DT204/49 complex, DT104, and the current
73 monophasic *S. Typhimurium* (*S. 4,[5],12:i:-*) sequence type 34 (ST34), since around the
74 middle of the last century ⁷⁻¹⁰. In contrast, some phage types are common in clonal groups
75 typically associated with a restricted host range, and in some cases altered pathogenicity. For
76 example, clonal groups of *S. Typhimurium* DT8, DT2 and DT56 circulate in populations of
77 ducks, pigeon, and passerine birds, respectively, and only rarely cause gastroenteritis in the
78 human population ¹¹⁻¹³. Also, specific clonal groups of *S. Typhimurium* ST313 are associated
79 with disseminated disease (invasive non-typhoidal *Salmonella*, iNTS) in sub-Saharan Africa
80 ^{14,15}.

81

82 In this study we report the population structure, gene flux, recombination and signatures of
83 functional diversification in the whole genome sequence of 131 strains of *S. Typhimurium*
84 with well characterised epidemiology. To assist in our analysis, we also report high quality
85 complete and closed whole genome sequence of six additional reference genomes,
86 representing diversity within the population structure not represented by previously reported
87 sequence.

88

89 **Results**

90

91 **Population structure of *S. Typhimurium* consists of two high-order clades containing**
92 **strains with distinct epidemiology.** Variant sites (38739 SNPs) in the core genome
93 sequence of 134 *S. Typhimurium* strains representing commonly isolated phage types
94 revealed two diverse phylogroups composed of three strains of ST36 that clustered separately
95 from the remaining 131 *Typhimurium* isolates (Supplementary Figure 1). The two clusters

96 were more similar to one another than any other serotype, including the most closely related
97 serotype, *S. Heidelberg*. Since the majority of *S. Typhimurium* formed a large number of
98 relatively tightly clustered isolates, we focussed on the analysis of the population structure
99 and evolution of this phylogroup. A phylogenetic tree constructed using variant sites (8382
100 SNPs) in the core genome sequence of the 131 *S. Typhimurium* strains and rooted with *S.*
101 *Heidelberg*, revealed a ‘star’ topology with relatively long internal branches extending from a
102 hypothetical common ancestor, and diversification at the terminal branches (Figure 1). The
103 population structure determined using a three-level hierarchical Bayesian approach¹⁶
104 resolved *S. Typhimurium* into two major clades, designated α and β , and third-level clades (α
105 8-18 and β 1-7) that corresponded to known epidemic clades. Clade β was defined by an
106 internal branch defined by approximately 100 core genome SNPs that originated from a
107 common ancestor of the basal clade α .

108

109 Despite relatively few SNPs distinguishing clade α and β , these clades exhibited distinct
110 epidemiology characterised by association predominantly with livestock (clade α) or avian
111 species, including wild species (clade β). Strikingly all pig isolates from our sampling were
112 located in clade α . Cattle isolates were in both first order clades (11 in clade α 7 in clade β),
113 but in clade β they had a relatively limited distribution with five of the isolates from a
114 subclade containing the DT204/49 complex of strains associated with a cattle associated
115 epidemic in the 1970’s⁸. Clade α contained strains from several previously described
116 epidemics in livestock animal species, including in pigs (α 12, U288)¹⁷, two clades associated
117 with recent pandemic clonal groups associated with pigs, cattle and poultry (α 17,
118 monophasic *Typhimurium* ST34 and α 15, DT104)¹⁸⁻²⁰, and potentially epidemic clades not
119 previously described in the literature, consisting of isolates from pigs, cattle and poultry (α 8

120 and α 11). Clade- β was characterised by many long internal branches, indicative of a
121 relatively high level of sequence divergence, relative to those in Clade- α . In contrast, clade- β
122 contained several third-level clades previously described as host-adapted, particularly for
123 avian species such as passerine birds (β 5, DT56), duck (β 2, DT8) and pigeon (β 3 DT2)²¹,
124 and ST313 that includes two sub-clades specifically associated with disseminated disease in
125 sub-Saharan Africa²².

126

127

128 **Antimicrobial resistance genes and plasmid replicons are associated with strains**
129 **associated with livestock.** The presence of multiple third-level clades associated with recent
130 livestock associated epidemic strains in the first level clade α and the relative paucity in clade
131 β suggested that they may be under differential anthropogenic selection pressure. A key
132 anthropogenic selection pressure on microbial populations circulating in livestock is the
133 widespread use of antimicrobial drugs in animal husbandry. Consistent with their distinct
134 epidemiology, antimicrobial resistance (AMR) genes were common in clades α (mean of 2.7
135 per strain) and relatively rare in β (mean 0.38 per strain) (Figure 1A and Supplementary
136 Table 2). Indeed, AMR genes in clade- β were restricted to strains from DT204/49 complex
137 known to be associated with cattle⁸ and the ST313 associated with disseminated disease in
138 sub-Saharan Africa commonly treated with antibiotics²². AMR genes are commonly present
139 on plasmids and we therefore determined the presence of plasmid replicon sequence in short
140 read sequence data from the 131 strains in clades α and β . The IncF replicon corresponding
141 to the presence of the virulence plasmid pSLT²³ was widespread in *S. Typhimurium*, but
142 notably absent from a number of third-level clades including β 5 (DT56), α 11 and α 17
143 (monophasic *S. Typhimurium* ST34). The IncQ1 plasmid replicon, previously associated with

144 antibiotic resistance ²⁴, was also widespread, particularly in clade α , associated with
145 domesticated animal species.

146

147 **Distinct patterns of genome degradation and signatures of functional divergence and**
148 **invasiveness in clades α and β .** Hypothetically disrupted coding sequences (HDCS) due to
149 frameshift mutations or premature stop codons were determined in high quality finished and
150 closed genome sequence of 11 representative strains from major subclades (Figure 1B and
151 Supplementary Table 3). Representative strains of clade α generally contained fewer HDCS
152 than those from clade β , with the exception of NCTC13348 (DT104) and SL1344
153 (DT204/49) that had atypically high and low numbers of HDCS, respectively (Figure 1B).
154 However, none of the HDCS in NCTC13348 were in genes previously been implicated in
155 pathogenesis, while SL1344 contained two virulence gene HDCS (*lpfD* and *ratB*) (Figure
156 1A). In general, clade α strains had 0-3 HDCS in virulence genes (mean 0.8 SD 0.9), while
157 clade β was characterised by multiple lineage containing three or more virulence gene HDCS
158 (mean 5.0 SD 3.2) (Figure 1A and Supplementary Table 4). Isolates in clade β 5 (DT56,
159 passerine bird associated ¹³) contained up to eleven virulence gene HDCS (*lpfD*, *ratB*, *sseK3*,
160 *siiE*, *siiC*, *ttrB*, *sseJ*, *gogB*, *sseK2*, *fimH* and *katE*). The greatest number of HDCS in clade α
161 were observed in α 12 (U288, possibly pig adapted), in which three virulence genes were
162 affected (*avrA*, *sadA* and *tsr*). *LpfD* was found to be the only HDCS (Supplementary Table 3)
163 that segregated between clade- α and clade- β . A 10-nucleotide deletion causing a frameshift
164 mutation in *lpfD* resulting in a truncation approximately half way into the protein in all
165 isolates of clade- β .

166

167 To quantify the relative level of functional divergence in the proteome of isolates in each
168 clade we used a profile hidden Markov Model approach, delta-bitscore (DBS profiling) ²⁵

169 (Figure 1C and Supplementary Table 5). The method assigned a value (bitscore) to peptides
170 of the proteome that indicated how well each sequence fitted the HMM. We determined the
171 difference in bitscore of the proteome of each isolate relative to that of *S. Typhimurium* strain
172 SL1344 (DBS = bitscore SL1344 proteome - bitscore test proteome). A greater DBS is
173 therefore indicatives of excess of polymorphisms that potentially alter protein function, and
174 most likely a loss of function as it indicates divergence from the profile HMM. Mean DBS
175 was significantly greater ($p<0.05$, Wilcoxon test) for proteomes of strains in clade- β
176 compared with clade- α (Figure 1C). In general, third-level clades in clade- α exhibited DBS
177 of approximately zero, consistent with limited functional divergence. Notably, despite
178 considerable numbers of HDCS in strain NCTC13348 (α 15, DT104), DBS was only
179 moderately elevated in this clade. Proteomes of strains in clade β exhibited mean DBS of
180 0.03 and above with the exception of clades β 1 and β 2. The proportion of the proteome with
181 any deviation in DBS was also greater in clade- β than clade- α (Figure 1D).

182

183 We also used a machine learning approach to predict the ability of strains to cause
184 extraintestinal disease based on the DBS of 196 proteins that were recently determined as the
185 most predictive of the invasiveness phenotype in 13 serotypes (six extra-intestinal pathovars
186 and seven gastrointestinal pathovars) of *S. enterica* subspecies I^{26,27} (Figure 1E). The
187 invasiveness index metric is the fraction of decision trees in a random forest algorithm that
188 vote for an invasive phenotype. The invasiveness index was significantly greater ($p<0.05$,
189 Wilcoxon test) in clade- β than clade- α , consistent with the epidemiology and pathogenicity
190 of the isolates located in these clades^{3,4}.

191

192 **The clade-specific accessory genome is largely driven by acquisition of prophage genes**
193 **and integrative elements.** A pangenome analysis of *S. Typhimurium* (excluding the ST36

194 phylogroup) identified 9167 total gene families. The core genome (present in 99-100% of
195 strains) was 3672 genes, soft core genome (95-99%) 388 genes. Shell genome (15-95%) 792
196 genes, and cloud genes (0-15%) 4315 genes (Figure 2A and 2B). We defined gene families of
197 the accessory genome as non-prophage chromosomal, prophage, plasmid and undefined,
198 based on their location and annotation in the complete and closed genomes of eleven
199 reference strains phylogenetically distributed across *S. Typhimurium* (Table 1). Gene
200 families not present in reference genomes were classified as ‘undefined’. The accessory
201 genome was defined as genes present in 95% or fewer of isolates and thus represents the
202 major source of genetic variation between strains.

203

204 Some gene families exhibited a distinct distribution in clade α or β , or within individual
205 third-level clades of clade α or β (Supplementary Figure 2). Four non-phage chromosomal
206 genes were specifically associated with clade- β strains, STM0038 a putative arylsulfatase,
207 *tdcE* encoding a pyruvate formate lyase 4, *aceF* acetyltransferase and *dinI* a DNA damage
208 inducible protein. A series of plasmid genes in clade β 2 corresponded to a region of p2 seen
209 in LO1157-10. This region is likely a transposon as it contains an IS200 transposase and an
210 integrase. Also, a number of prophage-associated genes were present throughout clade- β due
211 to apparent recombination in the ST64B prophage. The rate of gene flux in clade α and clade
212 β was determined by computing the number of accessory genes as a function of SNPs in each
213 clade. By this measure, the rate of gene flux was nearly twice as high in clade α compared to
214 clade β (Supplementary Figure 3).

215

216 Generally, gene flux in the non-prophage and non-plasmids gene families that were specific
217 to individual third level clades was limited to individual genes or small blocks of genes
218 (Supplementary Figure 2). The exception was two large genetic islands in clades α 15 (DT104

219 complex) and α 17 (monophasic *S. Typhimurium* ST34) corresponding to the presence of
220 SGI1²⁸ and SGI4²⁹. In addition, a chromosomal block of genes in clade β 2 corresponding to
221 an insertion at the Thr-tRNA at 368274 containing a series of hypothetical proteins and a
222 gene with similarity to the *trbL* gene involved in conjugal transfer (A0A3R0DZN6) and a
223 site-specific integrase. Some of these genes were also present in isolates in clades β 1 and β 3.
224 The greatest contribution to third level clade specific gene families was in the those with
225 predicted functions in prophage (Figure 2C and Supplementary Figure 2).

226

227

228 **Extant prophage repertoire is the result of recombination and infrequent loss of**
229 **ancestral elements and acquisition of new phage.** In order to investigate the flux of
230 prophage genes resulting in clade-specific repertoires, we identified prophage in eleven
231 complete and closed reference genomes of *S. Typhimurium* sequences. A total of 83
232 complete or partial prophage elements were identified in the eleven reference genomes
233 (Figure 4). Prophage were present at twelve variably occupied chromosomal loci and the
234 number per strain ranged from five in DT2 (strain 94-213) to nine in monophasic *S.*
235 *Typhimurium* ST34 (strain SO4698-09) (Table 1). Clustering of gene families in the
236 prophage pangenome identified 23 prophage, although in some cases blocks of genes were
237 replaced resulted in mosaic prophage for example “*Salmonella* virus ST64BX” and
238 “*Salmonella* virus Gifsy1X” (supplementary table 6), and the definition of families of
239 prophage with a high confidence was consequently problematic (Figure 4). Thirteen
240 prophage elements encoded at least one identifiable cargo gene, capable of modifying the
241 characteristics of the host bacterial strain, including eleven genes previously implicated in
242 virulence (Supplementary Table 3). Ten prophage families contained no recognisable cargo
243 genes. The evolutionary history of prophage acquisition and loss was reconstructed based on

244 principles of maximum parsimony. Six prophage (*Salmonella* viruses “BcepMuX”,
245 “Gifsy1X”, “Gifsy2X”, “Fels1X”, “ST64BX” and “sal3X”, hereafter referred to as BcepMu,
246 Gifsy1, Gifsy2, Fels1, ST64B and sal3) (supplementary table 6), that together accounted for
247 61 of the prophage in these genomes, were most likely present in the common ancestor of *S.*
248 Typhimurium. Loss of two of these ancestral prophage by three isolates (NCTC13348,
249 L01157-10 and D23580) represented the only evidence for decrease in prophage repertoire in
250 the dataset.

251

252 A total of 22 additional prophage had a limited distribution within *S. Typhimurium* strains,
253 present in three or fewer genomes *Salmonella* viruses (“TmEGF”, “TmSEN34”, “mTmII”,
254 “mTmV”, BTP1, “TmST104”, “SPN9CC”, Fels2, “TmHP1/mTmHP1”, BTP5, “TmC3PO”,
255 “RE2010”, “TmR2D2” and “TmSEN1”) (Table 2 and supplementary table 6) and are
256 therefore likely to have been acquired during the evolution of *S. Typhimurium* (Figure 4).
257 *Salmonella* virus BTP1 that was reported to be specific to the ST313 strains associated with
258 epidemics of invasive NTS disease in sub-Saharan Africa ³⁰, was also present in strain
259 SO7676-03, a strain in clade β5 (DT56 complex) adapted to circulation in wild bird
260 (Passerine) species ¹². “*Salmonella* virus mTmV” of strain SO4698-09, that carries the *sopE*
261 virulence gene in some monophasic *S. Typhimurium* ST34 isolates ¹⁸, was absent from all
262 other *S. Typhimurium* reference strains. However, a second prophage “*Salmonella* virus
263 mTmII” with similarity to SJ46 was also in strain SO4698-09, and shared several clusters of
264 gene families in common with mTmV.

265

266 With the notable exception of BcepMu the prophage predicted to be present in the common
267 ancestor of *S. Typhimurium* exhibited considerable variation, potentially due to
268 recombination ³¹ (Figure 4). Recombination is a major source of genetic variation in bacteria,

269 although the level of recombination seen in other bacteria and indeed other *Salmonella*
270 serovars vary greatly. We identified potential recombination in the genome sequence of the
271 131 *S. Typhimurium* of the main phylogroup by the identification of atypical SNP density.
272 Recombination was almost exclusively present in prophage regions resulting in clade specific
273 sequence variation (Figure 5). Recombination resulted in replacement of large blocks of gene
274 families in ancestral prophage elements. Fels1, sal3 and Gifsy2 were conserved in the most
275 reference strains, with the exception of Fels1 in DT104 and Gifsy2 in monophasic *S.*
276 *Typhimurium* ST34 that had large alternative blocks of gene families. Gifsy1 was highly
277 variable in all strains, but retained a core set of genes suggesting common ancestry and
278 frequent recombination. Variation in ST64B was also present in most strains, and variable
279 blocks of genes distinguished strains in first order clades α from β , and resulted in the
280 acquisition of *sseK3* virulence gene by the common ancestor of the latter.

281

282 Discussion

283

284 The population structure of *S. Typhimurium* consisted of two relatively distantly related
285 clusters comprising strains of ST36 and a second (main) phylogroup the remainder of
286 *Typhimurium*, but predominantly ST19, consistent with previous reports of two distinct *S.*
287 *Typhimurium* phylogroups^{32,33}. These two groups were more closely related to each other
288 than to other serotypes of *S. enterica* subspecies I, including the nearest neighbour, serotype
289 Heidelberg. The main *Typhimurium* phylogroup exhibited a star shaped phylogeny with
290 multiple deeply rooted branches emerging from a common ancestor, with diversification at
291 the terminal branches in some cases, associated with expansion of epidemic clonal groups.
292 The topology of this phylogroup was similar to that of serotypes of *S. enterica* subspecies I
293^{34,35}, with internal branches radiating from a common ancestor, defined by the accumulation

294 of hundreds of SNPs in *S. Typhimurium* compared with tens of thousands of SNPs defining
295 lineages of representative strains of *S. enterica* subspecies I serotypes⁴.

296

297 The nested phylogenetic structure, rooted with the *S. Heidelberg* outgroup, was characterised
298 by two high order clades (α and β), in which clade α was basal to clade β . Several deeply
299 rooted lineages of clade α contained isolates almost entirely from livestock. A single lineage
300 originating from the common ancestor of the main *S. Typhimurium* phylogroup gave rise to
301 the common ancestor of clade β and diversification into multiple lineages was accompanied
302 by apparent host adaptation to diverse host species, but notably many more avian species,
303 compared with clade α . Indeed some β subclades such as those associated with the DT56,
304 DT2 and DT8 complexes that are well characterized host adapted clonal groups^{11,12,21},
305 contained exclusively isolates from avian species, and were present on relatively extended
306 internal branches. This general phylogenetic topology is consistent with that described for
307 distinct collections of *S. Typhimurium* strains from North America and Asia^{36,37}.

308

309 Isolates in clades α and β appeared to be under distinct anthropogenic selection pressure for
310 the acquisition and maintenance of AMR, that correlated with their distribution in livestock
311 or wild avian species³⁸. Clade α isolates that were predominantly from livestock contained
312 several lineages with multiple AMR genes, while most clade β isolates contained few or no
313 AMR genes. Differential selection pressure for acquisition and maintenance of AMR genes is
314 consistent with the idea that some *S. Typhimurium* genotypic variants are adapted to
315 circulation in specific host populations that exert different selection pressure for the
316 acquisition and maintenance of AMR genes. Antimicrobials have been used widely to control
317 infection or as growth promoters in livestock, but wild animals are unlikely to encounter
318 therapeutic levels of these drugs³⁹. However, *S. Typhimurium* strains of DT56 and DT40

319 that we report are present in clade β , known to be associated with passerine birds, and lacking
320 AMR genes have also been reported in cattle where they may be subject to selection for
321 antimicrobial resistance⁴⁰, and we might therefore expect to find AMR genes in these strains
322 also. One possibility is that DT56 and DT40 strains from clade β may transiently colonise the
323 cattle host, but are unable to circulate in this population and do not transmit back to the avian
324 population with high frequency. Host adaptation to avian species therefore appears to create
325 an effective barrier to circulation in livestock. Two clade β lineages did contain strains with
326 multiple AMR genes, but in each case, they were atypical for this clade in that they were
327 associated with an epidemic in cattle (DT204/49 complex) or invasive NTS in people in sub-
328 Saharan Africa (ST313)^{7,22}, and therefore were likely to be under selection for AMR.

329

330 The molecular basis of the barrier to circulation of several clade of the β isolates in livestock
331 is not known, but likely the result of genotypic changes affecting functional diversification of
332 the proteome. The proteome delta bitscore (DBS) of clade β isolates exhibited elevated
333 divergence from profile HMMs of protein families in gamma proteobacteria, compared to
334 clade α isolates, potentially resulting in loss or altered protein function²¹. Similarly,
335 divergence was reported in the *S. Gallinarum* proteome, a serotype highly host adapted to
336 poultry where it is associated with fowl typhoid²⁵. β subclades also exhibited an elevated
337 invasiveness index, the fraction of decision trees using random forests that vote for the
338 invasive (extraintestinal) disease outcome, a predictive score of host adaptation to an
339 extraintestinal lifestyle. In our analysis DBS of 300 protein families most predictive of
340 disseminated disease in *S. Typhi*, *S. Patatyphi A*, *S. Gallinarum*, *S. Dublin* and *S.*
341 *Choleraesuis*, also predicted an extraintestinal lifestyle in *S. Typhimurium* β subclades,
342 consistent with the epidemiological data⁴. The pattern of functional divergence in some *S.*
343 *Typhimurium* β subclades may therefore at least in part be by a process of convergent

344 evolution with that observed as a result of the evolution of several extraintestinal serotypes *S.*
345 *enterica*, including *S. Typhi*, *S. Paratyphi A*, *S. Gallinarum*, *S. Dublin* and *S. Choleraesuis*²⁶.
346

347 Host adaptation of bacteria is commonly associated with the accumulation of HDCS,
348 potential pseudogenes that contribute to genome degradation^{41,42}. A total of 24 genes
349 previously implicated in virulence, adhesion or multicellular behaviour were HDCS in one or
350 more clade β isolates. Notably, 15 of these genes (63%) were also HDCS in highly host
351 adapted *S. Typhi* or *S. Paratyphi A*, serotypes that are restricted to humans and cause a
352 disseminated disease⁴³. HDCS were especially common in strains of DT56/DT40 complex
353 (clade β 5), that are reported to be highly host adapted to passerine birds, in which eleven
354 HDCS were observed. In *S. Paratyphi A*, many mutations and gene flux that occurred was
355 reported to be neutral, indicated by their sporadic distribution within clades, and frequent loss
356 from the population by purifying selection⁴⁴. In *S. Typhimurium*, we also observed some
357 examples of potential neutral mutations, but in many cases HDCS in virulence genes were
358 present in multiple related strains from the same subclade, indicating that they were likely
359 under selection, and stably maintained in the population (Figure 1). In contrast to clade β ,
360 genome degradation affecting virulence-associated genes was less frequent in isolates clade
361 α . Just three virulence gene HDCS had a clade phylogenetic signature, *avrA*, *tsr* and *sadA*, in
362 α 10, α 11 and α 12. Subclade α 12 (U288 complex) was the only clade to contain all of these
363 HDCS, consistent with the U288 complex exhibiting apparent host-adaption to pigs^{17,29}.
364 Therefore, the presence of HDCS in virulence associated genes was almost exclusively
365 associated with subclades containing strains with strong epidemiological evidence of host
366 adaptation⁴. Despite the *S. Typhimurium* DT104 complex strain NCTC13348 (clade α 15)
367 exhibiting a high level of genome degradation that was uncharacteristic for α subclades and
368 the broad host range epidemiology of the clonal group⁴⁵, no virulence or multicellular

369 behaviour genes were HDCS⁴⁵. Furthermore, the mean DBS for the proteome of strains from
370 clade α 15 was similar to that of other α subclades, suggesting that functional divergence as a
371 whole was not atypical from that of other clade α isolates.

372

373

374 While genes encoding components of the type III secretion systems (T3SS) 1 and 2 apparatus
375 were never HDCS, several genes encoding effector proteins secreted by them were (*sseI*,
376 *sseK2*, *sseK3*, *avrA*, *sseL*, *sseJ* and *gtgE*). The *sseI* gene is inactivated in ST313 due to
377 insertion of a transposable element, and results in hyper dissemination of these strains to
378 systemic sites of the host via CD11b⁺ migratory dendritic cells⁴⁶. The *sseK2*, *sseK3*, *avrA*
379 and *sseL* genes each encode effectors that inhibit the NF κ B signalling pathway thereby
380 modulating the proinflammatory response during infection⁴⁷⁻⁴⁹. Furthermore, these effectors
381 are commonly absent or degraded in serotypes of *Salmonella* serotypes associated with
382 disseminated disease^{50,51}, suggesting that altered interaction with the macrophage is essential
383 for disseminated disease in diverse *Salmonella* variants and hosts. In addition, several genes
384 encoding components of fimbrial or afimbrial adhesin systems (*sadA*, *ratB*, *lpfD*, *stfD*, *stbD*,
385 *safC*, *fimH*, *siiC* and *siiE*), or anaerobic respiration (*ttrB*), and chemotaxis (*tsr*) were HDCS
386 in one or more isolates. Many of these genes have been implicated in intestinal colonisation
387⁵²⁻⁵⁸, suggesting that their inactivation in host adapted variants may be associated with a loss
388 of selection for functions no longer required in a reduced host range or where intestinal
389 colonisation is no longer critical to transmission. The *sadA* and *katE*, genes are involved in
390 multicellular behaviour, biofilm formation and catalase activity that protects against oxidative
391 stress during high density growth functions, respectively, and are commonly affected by
392 genome degradation in host adapted pathovars of *Salmonella*^{59,60}.

393

394 The only virulence-related HDCS that segregated clades α and β resulted from a 10 bp
395 insertion in the *lpfD* gene of all clade β isolates. Within the host *S. Typhimurium*
396 preferentially colonises Peyer's patches (PPs)⁶¹, due to long-polar fimbriae *Lpf* binding to
397 M-cells⁶². Similarly, *lpf* genes in *E. coli* pathotypes are required for interaction with Peyer's
398 patches and intestinal colonisation⁶³. Despite the disruption of *lpfD* in the clade β isolate
399 SL1344, deletion of *lpf* reduced colonisation on the surface of chicken intestinal tissue,⁶⁴,
400 suggesting that long polar fimbriae retain function. Differences in intestinal architecture of
401 avian species that have the lymphoid organ the bursa of Fabricius containing numerous M
402 cells compared with mammalian species that have Peyer's patches with relatively scarce M
403 cells^{65,66} may explain the pattern of *lpfD* HDCS in *S. Typhimurium*. The function of long
404 polar fimbriae expressing full length LpfD in clade α isolates has not been investigated, but
405 its distribution in isolates from livestock and human infections mark it as of potential
406 importance to human health.

407
408 Bacterial genome diversity is largely driven by the flux of genes resulting from acquisition by
409 horizontal gene transfer and deletion, rather than allelic variation⁶⁷. The accessory genome of
410 *S. Typhimurium* revealed few genes that segregated clade α and β , but distinct forms of
411 ST64B prophage resulting from recombination that replaced a large block of genes were
412 present in clade α and β , and resulted in the presence of *sseK3* specifically in clade β . The
413 accessory genome contributed significantly to genetic variation that distinguished third order
414 subclades in both clade α and β , especially phage and plasmid genes. Non-phage
415 chromosomal genes exhibited relatively little clade specific accessory genome suggesting
416 that the majority was the result of deletions or gene acquisition on small mobile genetic
417 elements that were neutral and subsequently lost, as observed previously in *S. Paratyphi*⁴⁴.
418 However, three large genetic elements were acquired on the chromosome in α 15 (DT104) or

419 α 17 (monophasic *S. Typhimurium* ST34), the two most recent dominant MDR pandemic
420 clonal groups that together account for over half of all *S. Typhimurium* infections in the
421 human population Europe in the past 30 years. The acquired genes corresponded to SGI1²⁸
422 in the DT104 complex, and SGI4 and a composite transposon in monophasic *S. Typhimurium*
423 ST34^{18,29}, highlighting the likely importance of horizontal gene transfer in the emergence of
424 epidemic clones.

425

426 Variable prophage repertoires are a major source of genetic diversity in *Salmonella*^{68,69}, and
427 may contribute to the emergence and spread by impacting the fitness during intra-niche
428 competition due to lytic killing or lysogenic conversion of competing strains⁷⁰. This view
429 was supported by the considerable phage-associated gene flux observed in *S. Typhimurium*.
430 Importantly, the phage component of the accessory genome in *S. Typhimurium* had a strong
431 correlation with the third level clade, suggesting that although transfer was frequent, the
432 acquisition or loss of prophage elements was not transient, consistent with selection within
433 each clonal group⁴⁴. Reconstruction of the evolutionary history of prophage elements in the
434 main *S. Typhimurium* phylogroup indicated that the common ancestor likely contained six
435 prophage, Gifsy1, Gifsy2, Fels1, ST64B, sal3 and BcepMu, that were well conserved during
436 subsequent diversification. Just two of these ancestral prophage, Fels1 and Gifsy1, were lost
437 from the genome, on three occasions in different lineages. The majority of the prophage flux
438 was from the acquisition of between one to three prophage in each lineage, with the
439 exception of a lineage containing clade β 3 (DT2), that only contained the ancestral prophage
440 repertoire.

441

442 Together, our analyses are consistent with the view that the common ancestor of the main *S.*
443 *Typhimurium* phylogroup was a broad host range pathogen with little genome degradation

444 capable of circulating within multiple species of livestock. The age of the common ancestor
445 of *S. Typhimurium* is not known and previous attempts to calculate this using Bayesian
446 approaches have been frustrated by a weak molecular clock signal ⁷¹. However, the common
447 ancestor of *S. Paratyphi A* was estimated to have existed approximately 500 years ago and
448 provides a frame of reference ⁴⁴. The main *S. Typhimurium* phylogroup exhibited greater
449 genetic diversity than reported for *S. Paratyphi A*, an estimated maximum root to tip SNP
450 accumulation of approximately 750 and 250, respectively. Therefore, the common ancestor
451 of the main *S. Typhimurium* phylogroup is likely to have existed substantially before that of
452 *S. Paratyphi A*. However, the *S. Typhimurium* MRCA is unlikely to have predated the
453 domestication of livestock, that began around ten thousand years ago ⁷², raising the
454 possibility that the emergence of this phylogroup was linked to the anthropogenic selection
455 provided by the domestication of species for livestock. Subsequent to the emergence of the
456 common ancestor of this phylogroup, a single lineage appears to have spawned multiple
457 lineages, some of which have become highly host adapted to various wild avian species, by a
458 process of convergent evolution with that observed in host adapted serotypes of *Salmonella*
459 such as *S. Typhi*.

460

461 Materials and Methods

462

463 **Bacterial strains and culture.** *S. Typhimurium* isolates and Illumina short read sequence
464 used in this study have been described previously ⁷³, selected based on phage type determined
465 during routine surveillance by Public Health England (PHE) and the Animal and Plant Health
466 Agency (APHA) in order to represent the diversity *S. Typhimurium* phage types as a proxy
467 for genetic diversity. A strain collection of 134 *S. Typhimurium* or monophasic variant
468 isolates was composed of 2 to 6 randomly selected strains from the top ten most frequent

469 phage types from PHE and the top 20 most frequent phage types from APHA surveillance,
470 from 1990-2010 (2-5 strains of each) were used in this analysis. In addition, commonly used
471 lab strain SL1344 ⁷⁴, two reference strains of ST313 (D23580 and A130)⁷⁵ and three DT2
472 strains isolated from pigeon ⁷⁶ were included. For routine culture, bacteria were stored in
473 25% glycerol at -80°C and recovered by culture on Luria Bertani agar plates, and single
474 colonies were selected to inoculate LB broth that was incubated at 37 °C for 18 hours with
475 shaking.

476 **Short-read *de novo* assembly.** Illumina generated fastq files were assembled using an in-
477 house pipeline adapted from that previously described ⁷⁷. For each paired end reads, Velvet ⁷⁸
478 (1.2.08) was used to generate multiple assemblies varying the k-mer size between 31 and 61
479 using Velvet Optimiser ⁷⁹ and selecting the assembly with the longest N50. Assemblies were
480 then improved using Improve_Assembly software ⁸⁰ that uses SSPACE (version 3.0) ⁸¹ and
481 GapFiller (version 1.0) ⁸² to scaffold and gap-fill. Ragout was used to order contigs ⁸³ based
482 on comparison to the long-read sequences. The finished genomes were then annotated using
483 Prokka (version 1.11) ⁸⁴.

484 **Long-Read sequencing using Pacbio and sequence assembly.** DNA for long-read
485 sequencing on the Pacbio platform was extracted from 10 ml of cultured bacteria as
486 previously described ⁷³. Data were assembled using version 2.3 of the Pacbio SMRT analysis
487 pipeline (<https://smrt-analysis.readthedocs.io/en/latest/SMRT-Pipe-Reference-Guide-v2.2.0/>
488). The structure of the initial assembly was checked against a parallel assembly using
489 Miniasm ⁸⁵ which showed general agreement. The Pacbio best practice for circularizing
490 contigs was followed using Minimus ⁸⁶ and the chromosomal contiguous sequence in each
491 assembly was re-orientated to begin at the *thrA* gene. Illumina short read sequence data were
492 used to correct for SNPs and indels using iCORN2 (<http://icorn.sourceforge.net/>). The
493 finished sequences were then annotated using Prokka ⁸⁴.

494 **Phylogenetic reconstruction and population structure analysis.** The paired-end sequence
495 files for each strain were mapped to the SL1344 reference genome (FQ312003)⁷⁴ using the
496 Rapid haploid variant calling and core SNP phylogeny pipeline SNIPPY (version 3.0)
497 (<https://github.com/tseemann/snippy>). The size of the core genome was determined using
498 snp-sites (version 2.3.3)⁸⁷, outputting monomorphic as well as variant sites and only sites
499 containing A,C,T or G. Variant sites were identified and a core genome variation multifasta
500 alignment generated. The core genome of 134 *S. Typhimurium* (3686476 nucleotides)
501 (Supplementary Table 2) contained 17823 variant sites. The core genome (3739972
502 nucleotides) of 131 *S. Typhimurium* non-ST36 contained 8382 variant sites. The sequence
503 alignment of variant sites was used to generate a maximum likelihood phylogenetic tree with
504 RAxML using the GTRCAT model implemented with an extended majority-rule consensus
505 tree criterion⁸⁸. The genome sequence of *S. Heidelberg* (NC_011083.1) was used as an
506 outgroup in the analysis to identify the root and common ancestor of all *S. Typhimurium*
507 strains. HierBaps (hierarchical Bayesian analysis of Population Structure)¹⁶ was used to
508 estimate population structure using three nested levels of molecular variation and 10
509 independent runs of the optimization algorithm as reported previously⁸⁹. The input for this
510 analysis was the same SNP variant matrix for the 131 strains with reference to SL1344 that
511 was used to generate the GTRCAT phylogeny above.

512 ***In-silico* genotyping.** The presence of antibiotic resistance, virulence and plasmid replicon
513 genes in short-read data was determined by the mapping and local assembly of short reads to
514 data-bases of candidate genes using Ariba⁹⁰. The presence of candidate genes from the
515 resfinder⁹¹, VFDB⁹² and PlasmidFinder⁹³ databases was determined. Reads were mapped
516 to candidate genes using nucmer with a 90% minimum alignment identity. This tool was also
517 used to determine the presence of specific genes or gene allelic variants. The results of the
518 ARIBA determination of the presence or absence of the *lpfD* gene were confirmed using

519 SRST2⁹⁴ setting each alternative form of the gene as a potential allele. SRST2 was also used
520 to verify the ARIBA findings of the VFDB data set, as the presence of orthologous genes in
521 the genome was found to confound the interpretation of results.

522 **Hypothetically disrupted coding sequences (HDCS).** HDCS were identified in high-quality
523 finished Pacbio sequences and previously published reference sequences by identifying
524 putative altered open reading frames using the RATT annotation transfer tool⁹⁵. The *S.*
525 *Typhimurium* strains SL1344 annotation (accession no. FQ312003) was transferred to each
526 assembled sequence and coding sequences identified as having altered length were manually
527 curated by comparison of aligned sequences visualised using Artemis comparison tool (ACT)
528⁹⁶. Genes that contained either a premature stop codon or a frameshift mutation were
529 classified as HDCS. The identified HDCS were used to construct a database that could be
530 used as a reference for SRST2 (above) to detect presence or absence in short-read sequence
531 data. Alleles were called based on matching to 99% sequence identity and allowing one miss-
532 match per 1000 nucleotides.

533 **Delta bitscore.** Illumina short-read sequences were mapped to the SL1344 reference genome
534 and annotated using PROKKA and then analysed in a pairwise fashion against SL1344 using
535 delta-bit-score (DBS), a profile hidden Markov model based approach²⁵ with Pfam hidden
536 Markov Models (HMMs)⁹⁷. The mean DBS per genome and percentage of genes with
537 mutations in Pfam domains (non-zero DBS) are reported.

538 **Invasiveness Index.** The invasiveness index²⁶ for each strain was calculated to scan for
539 patterns of mutation accumulation common to *Salmonella* lineages adapted to an invasive
540 lifestyle. To calculate the invasiveness index, Illumina reads were mapped to a core-genome
541 reference using the snippy pipeline above, and annotated using PROKKA. Protein sequences
542 were then screened using phmmer from the HMMER3.0 package⁹⁸ to identify the closest
543 homologs to the 196 predictive genes used by the invasiveness index model. These genes

544 were then scored against profile hidden Markov models (HMMs) for these protein families
545 from the eggNOG database ²⁶ using hmmsearch ⁹⁸, to test for uncharacteristic patterns of
546 sequence variation. Bitscores produced in the comparison of each protein sequence to its
547 respective protein family HMM were then used as input to the model.

548 **Phage location and cargo in long-read strains.** The location of prophage elements in
549 assembled long-read sequences and published reference genome was determined using
550 Phaster ⁹⁹, which identified regions as being intact, questionable or incomplete. This yielded
551 a total of 83 potential complete and partial sequences across the 11 representative strains.
552 Prophage sequences were annotated using PROKKA that identified terminase, tail fibre,
553 recombinase/integrase proteins capsid proteins, phage related proteins, and hypothetical
554 proteins.

555 **Determination of recombination.** Recombination was inferred by identifying regions of
556 high SNP density from whole genome alignments of short-read data to SL1344, using
557 Gubbins ¹⁰⁰. The results were visualised using Phandango ¹⁰¹ and related to the predicted
558 prophage locations in the SL1344 genome. Similar results were obtained using maximum
559 likelihood inference using clonal frame ¹⁰².

560 **Determination of the *S. Typhimurium* pangenome.** The annotated assemblies of 131
561 predominantly *S. Typhimurium* ST19 isolates were used as the input to the pangenome
562 pipeline ROARY ¹⁰³. The presence or absence of genes was determined without splitting
563 orthologues. In order to characterise the contribution of prophage and plasmids to the
564 pangenome, genes were assigned to one of four categories, non-prophage genes located on
565 the chromosome, prophage genes, plasmid genes and undefined genes, based on the
566 similarity to annotated genes of complete and closed whole genome sequence of eleven
567 reference strains. Orthologous genes were identified based on > 90% nucleotide sequence

568 identity using nucmer¹⁰⁴. A core-genome reference sequence (genes present in at least 99%
569 of reference strains), was also constructed and used to determine the invasiveness index.

570 **Estimation of gene flux rates.** Genes were assigned a score based on their presence in
571 strains within a specific clade. The clade gene score was compared to the score determined
572 for strains outside of the clade to determine whether the gene was more prevalent within the
573 clade than without. Genes were classed as associated with the clade if their score was greater
574 than the mean plus two standard deviations of the non-cladial score (corresponding to the top
575 95% in a normal distribution).

576 The number of clade associate genes was compared with the number of SNPs associated with
577 a clade (this gives a measure of evolutionary time) to determine the level of gene flux for the
578 clade. The level of gene flux in the two first-level clades was then compared.

579 **Prophage classification.** PHASTER⁹⁹ curated prophage sequences were classified into
580 species and genus-level groupings based on the current criteria used by the Bacterial and
581 Archaeal Viruses Subcommittee of the International Committee on Taxonomy of Viruses
582 (ICTV)¹⁰⁵. At the species level, genomes were clustered at 95% nucleotide identity over the
583 whole genome length, meaning that two genomes belong to two different species if they
584 differ in more than 5% of their genome. Clustering was performed with CD-HIT-EST at 95%
585 nucleotide identity over 95% of the alignment length (99% of alignment length of shorter
586 sequence)¹⁰⁶ and with Gegenees, a pairwise nucleotide comparison tool, using the accurate
587 settings of 200 bp fragment size and 100 bp step size¹⁰⁷. The Gegenees output was used in
588 combination with vConTACT2¹⁰⁸ to classify the prophage sequences into new or existing
589 genera. Briefly, coding sequences were predicted with PROKKA⁸⁴ and transformed into a
590 table linking genomes and their encoding proteins. This table was used as input into
591 vConTACT along with the Viral RefSeq database v85¹⁰⁹. vConTACT then used Diamond¹¹⁰,
592 Markov clustering MCL¹¹¹ and ClusterONE¹¹² to predict viral clusters based on shared

593 protein content. The output was visualised using Cytoscape¹¹³. Genera were defined as
594 vConTACT viral clusters which shared a significant (>50%) nucleotide identity. The clusters
595 were then compared to the current and pending ICTV taxonomic classification
596 (ictvonline.org) using blast and vConTACT viral cluster output containing reference genomes
597 and all prophage sequences were assigned to new or existing taxa.

598

599

600

601

602 **Declarations and Acknowledgements**

603 No procedures or data collection required ethics approval, consent to participate, or consent
604 for publication. All data is available in accessible databases as indicated in the text and
605 supplementary information tables. The authors declare that they have no competing interests.
606 The project was conceived by MB and RAK, data analysis was performed by MB, GT, MK,
607 NW and EMA, interpretation of the data was by MB, RAK, LP, TD and NH, the manuscript
608 was drafted by MB and RAK, and edited and approved by GT, MK, NW, LP, TJD, EMA,
609 NH and RAK. The author(s) gratefully acknowledge the support of the Biotechnology and
610 Biological Sciences Research Council (BBSRC); RAK was funded by the BBSRC Institute
611 Strategic Programme Microbes in the Food Chain BB/R012504/1 and its constituent
612 project(s) BBS/E/F/000PR10348 and BBS/E/F/000PR10352, and by projects BB/J004529/1,
613 BB/M025489/1 and BB/N007964/1. NH was supported by a BBSRC funding for the Earlham
614 Institute BB/CCG1720/1. EMA was supported by the BBSRC Institute Strategic Programme
615 Gut Microbes and Health BB/R012490/1 and its constituent project BBS/E/F/000PR10356.
616 The genome sequencing for this work was carried out by the Genomics Pipelines group at the
617 Earlham Institute which is funded as a BBSRC National Capability (BB/CCG1720/1). EMA

618 was funded by the BBSRC Institute Strategic Programme Gut Microbes & Health
619 BB/R012490/1 and its constituent projects BBS/E/F/000PR10353 and
620 BBS/E/F/000PR10356. This research was supported in part by the NBI Computing
621 infrastructure for Science (CiS) group through use of HPC resources. The authors also
622 acknowledge advice and informatics support from Andrew Page, Andrea Telatin and Nabil-
623 Fareed Alikhan from the Quadram Institute Bioscience informatics support group.

624

Table 1. Characteristics of complete and closed whole genome sequence of *S. Typhimurium* reference strains.

strain	ST	PT ^a	accession no.	clade	chrom size ^b	pSLT ^c	P1 ^c	P2 ^c	ECC ^d	CDS ^e	SNPs ^f	Prophage ^g	total HAC ^h	spec. HAC
SO9207-07	19	DT170B	PRJEB34598	α11	4.92	-	-	-	0	4592	973	7 (5:1:1)	48	7
SO1960-05	19	U288	PRJEB34597	α12	4.90	154.4	19.4	18.2	3	4832	1272	8 (5:2:1)	57	16
NCTC13348	19	DT104	HF937208.1	α15	4.93	94.0	-	-	1	4751	1083	9 (5:3:1)	63	25
SO4698-09	34	DT193	PRJEB10340	α17	5.04	-	-	-	0	4754	1010	9 (6:2:1)	43	7
SO7676-03	19	DT56	PRJEB34599	β5	4.88	-	-	-	0	4570	797	6 (4:1:1)	75	26
SO9304-02	19	DT41	PRJEB34596	β4	5.05	117.4	32.2		3	4685	943	8 (5:2:1)	78	15
SL1344	19	DT44	FQ312003.1	β1	4.88	93.8	86.9	8.7	3	4771	0	9 (4:3:2)	42	2
LO1157-10	19	DT8	PRJEB34595	β2	4.86	93.8	23.9	22.2	3	4706	1101	7 (4:2:1)	62	19
D23580	313	-	FN424405.1	β1	4.88	117.0	84.6	-	2	4790	901	7 (5:2:0)	57	9
A130	313	DT56v	PRJEB34594	β1	4.93	166.9	-	-	1	4812	901	8 (5:2:1)	62	13
94-213	98	DT2	HG326213.1	β3	4.82	93.8	-	-	1	4598	903	5 (3:1:1)	66	20

^a Phage type / Sequence type, ^b Chromosome size (Mbp), ^c contig size (Kbp), ^d extra chromosomal contigs (ECC), putative plasmid sequences, ^e Number of predicted coding sequences, ^f number of SNPs with reference to SL1344, ^g Number of prophage (intact, incomplete, candidate), ^h total number of pseudogenes with reference to SL1344 allele, ⁱ number of clade-specific pseudogenes (Table S1), ^j mean delta bitscore (DBS) with reference to SL1344 protein orthologue

Figure Legends

Figure 1. Phylogenetic relationship of the ST19 *Salmonella* Typhimurium phylogroup.

(A) Maximum likelihood phylogenetic tree and based on sequence variation (SNPs) in the core genome with reference to *S. Typhimurium* strain SL1344. The root was identified using *S. Heidelberg* (accession number NC_011083.1) as the outgroup. 1st (α and β) and 3rd (α 11-19 and β 1-7) are indicated (vertical bars). Phage type complexes associated with the third-level clusters are indicated (bold type) colour coded with the lineages and representative strains from third level clusters (italicized type). The source of each isolate in the tree is indicated by filled boxes colour coded as indicated in the inset key (arrow). The presence of replicon sequence (grey box), antimicrobial resistance genes (blue box) and hypothetically disrupted coding sequence (HDCS) of virulence related genes (red box) in short read sequence data are indicated. (B) Bars indicate the number of ancestral (black), phage or insertion sequence elements (grey), chromosomal gene (colour coded with lineages in Figure 1A) HDCS in the genome of representative strains from each third level clade. (C) Box plots indicate the mean dbitscore (DBS: bitscore SL1344 – test strain bitscore) of proteomes in third level clades. (D) Box plot indicates the percentage of the proteome of the proteome of isolates from each third level clade with a non-zero bitscore (bitscore SL1344 – test strain bitscore >0 or <0) as an estimate of function divergence. (E) Box plots indicate the mean invasiveness index per genome, the fraction of random forest decision trees voting for an invasiveness phenotype based on training on the DBS of a subset of the proteome of ten gastrointestinal and extraintestinal pathovar serotypes.

Figure 2. The pan genome of 131 *S. Typhimurium* isolates. Gene families were identified based on sequence alignment with a cut off of 90% sequence identity and assigned to non-prophage chromosomal (red), prophage (green), plasmid (blue), or undefined (grey), based on

their genome context in eleven annotated reference genomes from each third level clade. (A)

Number of genome families in the core, softcore, shell and cloud components of the

pangenome. (B) Number of genome families of each pan genome component in isolates from

each *S. Typhimurium* third level clade. (C) Accessory genome (shell and cloud) in each

isolate. Gene families present in more than 130 or less than 5 strains were excluded.

Maximum Likelihood tree based on variation (SNPs) in the core genome with reference to *S.*

Typhimurium SL1344. Third-level clades are indicated in colour coded in common with the

phylogeny vertical bars.

Figure 3. Genome alignment and phylogenetic relationship of complete and closed

reference strains of *S. Typhimurium* or used in this study. Sequence with >90%

nucleotide sequence identity are indicated where this is direct alignment (green) or reverse

and complement (red). The location of prophage sequence (red bars) or integrative elements

(blue bars) are indicated. A maximum likelihood tree based on sequence variation (SNPs) in

the core genome with reference to *S. Typhimurium* strain SL1344 (left) is annotated with the

most likely order of acquisition (black arrow) or loss (red arrow) of prophage and integrative

elements, based on the principle of parsimony.

Figure 4. Clustering of prophage genes based on sequence identity indicates related

families and potential recombination. Genes from all prophage identified in complete and

closed whole genome sequence of eleven reference strains of *S. Typhimurium* were assigned

to families based on sequence identity (>90% identity). Prophage genes (columns) were

clustered to identify related prophage. The presence of a gene is indicated with a box

predicted function based on *in silico* annotation are colour coded based on annotation,

terminase (black), capsid (green), recombinase/integrase (purple), tail fibre (blue), other

phage associated (red), and hypothetical protein (grey). A cladogram showing the relationship of prophage is based on the pattern of gene presence or absence is indicated (top).

Figure 5 Recombination inferred by high SNP density in 131 *S. Typhimurium* strains.

Regions of high SNP density (red) are indicated for each of the 131 isolates in the *S. Typhimurium* ST19 cluster with reference to the *S. Typhimurium* strain SL1344 genome. Recombination is shown with reference to the population structure and phylogeny of *Typhimurium* shown in figure 1. The position of predicted prophage (blue) in the *S. Typhimurium* strain SL1344 genome are indicated (top).

Supplementary Figure 1. Phylogenetic relationship of *S. Typhimurium* and diverse *S. enterica* serotypes. Mid-point rooted maximum likelihood phylogenetic tree based on the variation (SNPs) in the core genome of 18 strains of *Salmonella* Typhimurium and 14 representative strain of diverse *S. enterica* subspecies *enterica* serotypes, with reference to *S. Typhimurium* strain SL1344 genome sequence. *S. Typhimurium* strains (red lineages and text) are present in two clusters, composed of 15 strains with isolates that are ST19, ST34, ST313, ST98 and ST568 and three more divergent isolates of ST36. The phylogeny is rooted with respect to *S. Heidelberg* and was calculated using SL1344 as a reference to create a core-genome variant-site alignment and the GTRCAT model in RAxML.

Supplementary figure 2. Accessory genome with strong clade association. Gene families with a strong clade association in clade a or b (A), or in one of the third level clades (B). Maximum likelihood phylogenetic tree and based on sequence variation (SNPs) in the core genome with reference to *S. Typhimurium* strain SL1344 (left). Third-level clades are

indicated and colour coordinated with that in Figure 1. Genes in each clade were assigned a score based on the number of strains containing the gene within the clade. This score was also calculated for the strains outside the clade. Clade associated genes were defined as genes that had scores greater than the mean plus two SD of the score for all other clades. Genes are colour coded based assignment to non-prophage chromosomal (red), prophage (green), plasmid (blue), or undefined (grey).

Supplementary Figure 3. Gene flux rate metrics determined for non-singleton gene families first-level clades.

Supplementary Table Legends

Supplementary Table 1. S. Typhimurium strain collection used in this study related to determine population structures. Table can be viewed at
<https://www.dropbox.com/sh/dh84yycc4tguirw3/AABnbbrSPtqEcjGY6IR6GggLa?dl=0>

Supplementary Table 2. Presence of plasmid operons, AMR genes (Resfinder) and virulence genes (VFDB) determined by *in-silico* genotyping using ARIBA software. The identifier column corresponds to the study-identifier column in Supplementary Table 1. Presence of genes are indicated by '1'. Table can be viewed at
<https://www.dropbox.com/sh/dh84yycc4tguirw3/AABnbbrSPtqEcjGY6IR6GggLa?dl=0>

Supplementary Table 3. Genome degradation in long-read reference strains. Potential hypothetically disrupted coding sequences (HDCS) were identified in reference genomes though anomalies in annotation transfer from the SL1344 reference sequence using RATT

software and manual curation to exclude false positive HDCS. Table can be viewed at
<https://www.dropbox.com/sh/dh84yycc4tguirw3/AABnbbrSPtqEcjGY6IR6GggLa?dl=0>

Supplementary Table 4. Summary of HDCS in *S. Typhimurium* main phylogroup. The presence of HDCS alleles was determined *in-silico* using SRST2 reported in Figure 1. The study identifier refers to isolates in Supplementary Table 1. Alleles are specified as wild-type (WT) or HDCS. In some cases, multiple HDCS forms were determined to be present and these are denoted as HDCS1 or HDCS2 etc. Table can be viewed at

<https://www.dropbox.com/sh/dh84yycc4tguirw3/AABnbbrSPtqEcjGY6IR6GggLa?dl=0>

Supplementary Table 5. Summary of DBS and Invasiveness Index analysis. Isolate identifier corresponds to the identifier column in Supplementary Table 1. The mean delta bitscore (DBS) for the proteome of each strain, the number of proteins with a DBS greater than ten. Invasiveness index for each isolate is indicated. Table can be viewed at
<https://www.dropbox.com/sh/dh84yycc4tguirw3/AABnbbrSPtqEcjGY6IR6GggLa?dl=0>

Supplementary Table 6. Characteristics of prophage elements present in complete and closed whole genome sequence of *S. Typhimurium* reference strains. Table can be viewed at <https://www.dropbox.com/sh/dh84yycc4tguirw3/AABnbbrSPtqEcjGY6IR6GggLa?dl=0>

References

- 1 Kingsley, R. & Bäumler, J. Host adaptation and the emergence of infectious disease: the *Salmonella* paradigm. *Mol Micro* **36**, doi:10.1046/j.1365-2958.2000.01907.x (2000).
- 2 Kirk, M. D. *et al.* World Health Organization Estimates of the Global and Regional Disease Burden of 22 Foodborne Bacterial, Protozoal, and Viral Diseases, 2010: A Data Synthesis. *PLoS Med* **12**, e1001921, doi:10.1371/journal.pmed.1001921 (2015).
- 3 Rabsch, W. *et al.* *Salmonella enterica* serotype *Typhimurium* and its host-adapted variants. *Infect Immun* **70**, 2249-2255. (2002).

4 Branchu, P., Bawn, M. & Kingsley, R. A. Genome variation and molecular epidemiology of *Salmonella* Typhimurium pathovariants. *Infect Immun* **86**, e00079-00018, doi:10.1128/IAI.00079-18 (2018).

5 Anonymous. *Salmonella in livestock production in Great Britain, 2017*, <<https://www.gov.uk/government/publications/salmonella-in-livestock-production-in-great-britain-2017>> (2018).

6 Rabsch, W. in *Salmonella, Methods and Protocols Methods in Molecular Biology* (eds H. Schatten & A. Eisenstark) Ch. 10, 177-212 (Humana Press, 2007).

7 Threlfall, E. J., Ward, L. R. & Rowe, B. Spread of multiresistant strains of *Salmonella* typhimurium phage types 204 and 193 in Britain. *Br Med J* **2**, 997 (1978).

8 Rabsch, W., Tschepe, H. & Baumler, A. J. Non-typhoidal salmonellosis: emerging problems. *Microbes Infect* **3**, 237-247 (2001).

9 Rabsch, W., Truepschuch, S., Windhorst, D. & Gerlach, R. G. *Typing phages and prophages of Salmonella*. 25-48 (Caister Academic Press, 2011).

10 Tassinari, E. *et al.* Microevolution of antimicrobial resistance and biofilm formation of *Salmonella* Typhimurium during persistence on pig farms. *Sci Rep* **9**, 8832, doi:10.1038/s41598-019-45216-w (2019).

11 Ashton, P. M. *et al.* Whole Genome Sequencing for the Retrospective Investigation of an Outbreak of *Salmonella* Typhimurium DT 8. *PLoS Curr* **7**, doi:10.1371/currents.outbreaks.2c05a47d292f376afc5a6fcdd8a7a3b6 (2015).

12 Mather, A. E. *et al.* Genomic Analysis of *Salmonella* enterica Serovar Typhimurium from Wild Passerines in England and Wales. *Appl Environ Microbiol* **82**, 6728-6735, doi:10.1128/AEM.01660-16 (2016).

13 Hughes, L. A. *et al.* Characterisation of *Salmonella* enterica serotype Typhimurium isolates from wild birds in northern England from 2005 - 2006. *BMC Vet Res* **4**, 4, doi:10.1186/1746-6148-4-4 (2008).

14 Kingsley, R. A. *et al.* Epidemic multiple drug resistant *Salmonella* typhimurium causing invasive disease in sub-Saharan Africa have a distinct genotype. *Genome Res* **19**, doi:10.1101/gr.091017.109 (2009).

15 Feasey, N. A., Dougan, G., Kingsley, R. A., Heyderman, R. S. & Gordon, M. A. Invasive non-typhoidal salmonella disease: an emerging and neglected tropical disease in Africa. *Lancet* **379**, 2489-2499, doi:10.1016/S0140-6736(11)61752-2 (2012).

16 Cheng, L., Connor, T. R., Siren, J., Aanensen, D. M. & Corander, J. Hierarchical and spatially explicit clustering of DNA sequences with BAPS software. *Mol Biol Evol* **30**, 1224-1228, doi:10.1093/molbev/mst028 (2013).

17 Hooton, S. P., Atterbury, R. J. & Connerton, I. F. Application of a bacteriophage cocktail to reduce *Salmonella* Typhimurium U288 contamination on pig skin. *Int J Food Microbiol* **151**, 157-163, doi:10.1016/j.ijfoodmicro.2011.08.015 (2011).

18 Petrovska, L. *et al.* Microevolution of monophasic *Salmonella* Typhimurium during epidemic, United Kingdom, 2005–2010. *Emerging infectious diseases* **22**, 617 (2016).

19 Mather, A. E. *et al.* Distinguishable Epidemics of Multidrug-Resistant *Salmonella* Typhimurium DT104 in Different Hosts. *Science* **341**, 1514-1517, doi:10.1126/science.1240578 (2013).

20 Leekitcharoenphon, P. *et al.* Global Genomic Epidemiology of *Salmonella* enterica Serovar Typhimurium DT104. *Appl Environ Microbiol* **82**, 2516-2526, doi:10.1128/AEM.03821-15 (2016).

21 Kingsley, R. A. *et al.* Genome and Transcriptome Adaptation Accompanying Emergence of the Definitive Type 2 Host-Restricted *Salmonella* enterica Serovar Typhimurium Pathovar. *mBio* **4**, doi:10.1128/mBio.00565-13 (2013).

22 Okoro, C. K. *et al.* Intracontinental spread of human invasive *Salmonella* Typhimurium pathovariants in sub-Saharan Africa. *Nat Genet* **44**, 1215-1221, doi:10.1038/ng.2423 (2012).

23 Lobato-Marquez, D., Molina-Garcia, L., Moreno-Cordoba, I., Garcia-Del Portillo, F. & Diaz-Orejas, R. Stabilization of the Virulence Plasmid pSLT of *Salmonella* Typhimurium by Three Maintenance Systems and Its Evaluation by Using a New Stability Test. *Frontiers in molecular biosciences* **3**, 66, doi:10.3389/fmolb.2016.00066 (2016).

24 Oliva, M. *et al.* A novel group of IncQ1 plasmids conferring multidrug resistance. *Plasmid* **89**, 22-26, doi:10.1016/j.plasmid.2016.11.005 (2017).

25 Wheeler, N. E., Barquist, L., Kingsley, R. A. & Gardner, P. P. A profile-based method for identifying functional divergence of orthologous genes in bacterial genomes. *Bioinformatics* **32**, 3566-3574, doi:10.1093/bioinformatics/btw518 (2016).

26 Wheeler, N. E., Gardner, P. P. & Barquist, L. Machine learning identifies signatures of host adaptation in the bacterial pathogen *Salmonella enterica*. *PLoS genetics* **14**, e1007333-e1007333, doi:10.1371/journal.pgen.1007333 (2018).

27 Van Puyvelde, S. *et al.* An African *Salmonella* Typhimurium ST313 sublineage with extensive drug-resistance and signatures of host adaptation. *Nat Commun* **10**, 4280, doi:10.1038/s41467-019-11844-z (2019).

28 Boyd, D. *et al.* Complete nucleotide sequence of a 43-kilobase genomic island associated with the multidrug resistance region of *Salmonella enterica* serovar Typhimurium DT104 and its identification in phage type DT120 and serovar Agona. *J Bacteriol* **183**, 5725-5732, doi:10.1128/JB.183.19.5725-5732.2001 (2001).

29 Branchu, P. *et al.* SGI-4 in Monophasic *Salmonella* Typhimurium ST34 Is a Novel ICE That Enhances Resistance to Copper. *Front Microbiol* **10**, 1118, doi:10.3389/fmicb.2019.01118 (2019).

30 Owen, S. V. *et al.* Characterization of the Prophage Repertoire of African *Salmonella* Typhimurium ST313 Reveals High Levels of Spontaneous Induction of Novel Phage BTP1. *Front Microbiol* **8**, 235, doi:10.3389/fmicb.2017.00235 (2017).

31 Summer, E. J. *et al.* Burkholderia cenocepacia phage BcepMu and a family of Mu-like phages encoding potential pathogenesis factors. *J Mol Biol* **340**, 49-65, doi:10.1016/j.jmb.2004.04.053 (2004).

32 Gymoese, P. *et al.* Investigation of Outbreaks of *Salmonella enterica* Serovar Typhimurium and Its Monophasic Variants Using Whole-Genome Sequencing, Denmark. *Emerg Infect Dis* **23**, 1631-1639, doi:10.3201/eid2310.161248 (2017).

33 Sun, J. *et al.* The molecular epidemiological characteristics and genetic diversity of *salmonella* typhimurium in Guangdong, China, 2007-2011. *PLoS One* **9**, e113145, doi:10.1371/journal.pone.0113145 (2014).

34 Alikhan, N. F., Zhou, Z., Sergeant, M. J. & Achtman, M. A genomic overview of the population structure of *Salmonella*. *PLoS Genet* **14**, e1007261, doi:10.1371/journal.pgen.1007261 (2018).

35 Lan, R., Reeves, P. R. & Octavia, S. Population structure, origins and evolution of major *Salmonella enterica* clones. *Infect Genet Evol* **9**, 996-1005, doi:10.1016/j.meegid.2009.04.011 (2009).

36 Zhang, S. *et al.* Zoonotic Source Attribution of *Salmonella enterica* Serotype Typhimurium Using Genomic Surveillance Data, United States. *Emerg Infect Dis* **25**, 82-91, doi:10.3201/eid2501.180835 (2019).

37 Mather, A. E. *et al.* New Variant of Multidrug-Resistant *Salmonella enterica* Serovar Typhimurium Associated with Invasive Disease in Immunocompromised Patients in Vietnam. *mBio* **9**, doi:10.1128/mBio.01056-18 (2018).

38 Van Boeckel, T. P. *et al.* Global trends in antimicrobial use in food animals. *Proc Natl Acad Sci U S A* **112**, 5649-5654, doi:10.1073/pnas.1503141112 (2015).

39 McEwen, S. A. & Fedorka-Cray, P. J. Antimicrobial use and resistance in animals. *Clin Infect Dis* **34 Suppl 3**, S93-S106, doi:10.1086/340246 (2002).

40 Horton, R. A. *et al.* Wild birds carry similar *Salmonella enterica* serovar Typhimurium strains to those found in domestic animals and livestock. *Res Vet Sci* **95**, 45-48, doi:10.1016/j.rvsc.2013.02.008 (2013).

41 Abraham, S. *et al.* Isolation and plasmid characterization of carbapenemase (IMP-4) producing *Salmonella enterica* Typhimurium from cats. *Sci Rep* **6**, 35527, doi:10.1038/srep35527 (2016).

42 Parkhill, J. *et al.* Complete genome sequence of a multiple drug resistant *Salmonella enterica* serovar Typhi CT18. *Nature* **413**, 848-852 (2001).

43 Holt, K. E. *et al.* Pseudogene accumulation in the evolutionary histories of *Salmonella enterica* serovars Paratyphi A and Typhi. *BMC Genomics* **10**, 36 (2009).

44 Zhou, Z. *et al.* Transient Darwinian selection in *Salmonella enterica* serovar Paratyphi A during 450 years of global spread of enteric fever. *Proc Natl Acad Sci U S A* **111**, 12199-12204, doi:10.1073/pnas.1411012111 (2014).

45 Threlfall, E. J. Epidemic *Salmonella typhimurium* DT 104 - a truly international multiresistant clone. *Journal of Antimicrobial Chemotherapy* **46**, 7-10, doi:10.1093/Jac/46.1.7 (2000).

46 Carden, S. E. *et al.* Pseudogenization of the Secreted Effector Gene *sseI* Confers Rapid Systemic Dissemination of *S. Typhimurium* ST313 within Migratory Dendritic Cells. *Cell Host Microbe* **21**, 182-194, doi:10.1016/j.chom.2017.01.009 (2017).

47 Collier-Hyams, L. S. *et al.* Cutting edge: *Salmonella* AvrA effector inhibits the key proinflammatory, anti-apoptotic NF-kappa B pathway. *J Immunol* **169**, 2846-2850, doi:10.4049/jimmunol.169.6.2846 (2002).

48 Yang, Z. *et al.* SseK3 Is a *Salmonella* Effector That Binds TRIM32 and Modulates the Host's NF-kappaB Signalling Activity. *PLoS One* **10**, e0138529, doi:10.1371/journal.pone.0138529 (2015).

49 Geng, S. *et al.* The SseL protein inhibits the intracellular NF-kappaB pathway to enhance the virulence of *Salmonella* Pullorum in a chicken model. *Microb Pathog* **129**, 1-6, doi:10.1016/j.micpath.2019.01.035 (2019).

50 Nuccio, S.-P. & Bäumler, A. J. Comparative Analysis of *Salmonella* Genomes Identifies a Metabolic Network for Escalating Growth in the Inflamed Gut. *mBio* **5**, doi:10.1128/mBio.00929-14 (2014).

51 Johnson, R., Mylona, E. & Frankel, G. Typhoidal *Salmonella*: Distinctive virulence factors and pathogenesis. *Cell Microbiol* **20**, e12939, doi:10.1111/cmi.12939 (2018).

52 Kingsley, R. A. *et al.* 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 (2003).

53 Bäumler, A. J., Tsolis, R. M. & Heffron, F. The *lpf* fimbrial operon mediates adhesion to murine Peyer's patches. *Proc. Natl. Acad. Sci. USA* **93**, 279-283 (1996).

54 Weening, E. H. *et al.* The *Salmonella enterica* serotype Typhimurium *lpf*, *bcf*, *stb*, *stc*, *std*, and *sth* fimbrial operons are required for intestinal persistence in mice. *Infect Immun* **73**, 3358-3366 (2005).

55 Gerlach, R. G. *et al.* *Salmonella* Pathogenicity Island 4 encodes a giant non-fimbrial adhesin and the cognate type 1 secretion system. *Cell Microbiol* **9**, 1834-1850 (2007).

56 Winter, S. E. *et al.* Gut inflammation provides a respiratory electron acceptor for *Salmonella*. *Nature* **467**, 426-429, doi:nature09415 [pii] 10.1038/nature09415 (2010).

57 Rivera-Chavez, F. *et al.* Energy Taxis toward Host-Derived Nitrate Supports a Salmonella Pathogenicity Island 1-Independent Mechanism of Invasion. *mBio* **7**, doi:10.1128/mBio.00960-16 (2016).

58 Bourret, T. J., Liu, L., Shaw, J. A., Husain, M. & Vazquez-Torres, A. Magnesium homeostasis protects Salmonella against nitrooxidative stress. *Sci Rep* **7**, 15083, doi:10.1038/s41598-017-15445-y (2017).

59 MacKenzie, K. D. *et al.* Parallel evolution leading to impaired biofilm formation in invasive Salmonella strains. *PLoS Genet* **15**, e1008233, doi:10.1371/journal.pgen.1008233 (2019).

60 Singletary, L. A. *et al.* Loss of Multicellular Behavior in Epidemic African Nontyphoidal Salmonella enterica Serovar Typhimurium ST313 Strain D23580. *mBio* **7**, e02265, doi:10.1128/mBio.02265-15 (2016).

61 Gonzales, A. M., Wilde, S. & Roland, K. L. New Insights into the Roles of Long Polar Fimbriae and Stg Fimbriae in Salmonella Interactions with Enterocytes and M Cells. *Infect Immun* **85**, doi:10.1128/iai.00172-17 (2017).

62 Bäumler, A. J. *et al.* Identification of a new iron regulated locus of *Salmonella typhi*. *Gene* **193**, 207-213 (1996).

63 Lloyd, S. J. *et al.* A double, long polar fimbria mutant of *Escherichia coli* O157:H7 expresses Curli and exhibits reduced in vivo colonization. *Infect Immun* **80**, 914-920, doi:10.1128/IAI.05945-11 (2012).

64 Ledeboer, N. A., Frye, J. G., McClelland, M. & Jones, B. D. *Salmonella enterica* Serovar Typhimurium Requires the Lpf, Pef, and Tafi Fimbriae for Biofilm Formation on HEp-2 Tissue Culture Cells and Chicken Intestinal Epithelium. *Infection and Immunity* **74**, 3156-3169, doi:10.1128/IAI.01428-05 (2006).

65 Nakato, G. *et al.* New approach for m-cell-specific molecules screening by comprehensive transcriptome analysis. *DNA Res* **16**, 227-235, doi:10.1093/dnares/dsp013 (2009).

66 Kozuka, Y., Nasu, T., Murakami, T. & Yasuda, M. Comparative studies on the secondary lymphoid tissue areas in the chicken bursa of Fabricius and calf ileal Peyer's patch. *Vet Immunol Immunopathol* **133**, 190-197, doi:10.1016/j.vetimm.2009.08.003 (2010).

67 Rodriguez-Valera, F. *et al.* Explaining microbial population genomics through phage predation. *Nat Rev Micro* **7**, 828-836, doi:http://www.nature.com/nrmicro/journal/v7/n11/supplinfo/nrmicro2235_S1.html (2009).

68 Mottawea, W. *et al.* *Salmonella enterica* Prophage Sequence Profiles Reflect Genome Diversity and Can Be Used for High Discrimination Subtyping. *Front Microbiol* **9**, 836, doi:10.3389/fmicb.2018.00836 (2018).

69 Figueroa-Bossi, N., Uzzau, S., Maloriol, D. & Bossi, L. Variable assortment of prophages provides a transferable repertoire of pathogenic determinants in *Salmonella*. *Mol Microbiol* **39**, 260-272. (2001).

70 Bossi, L., Fuentes, J. A., Mora, G. & Figueroa-Bossi, N. Prophage contribution to bacterial population dynamics. *J Bacteriol* **185**, 6467-6471 (2003).

71 Hawkey, J. *et al.* Evidence of microevolution of *Salmonella* Typhimurium during a series of egg-associated outbreaks linked to a single chicken farm. *BMC Genomics* **14**, 800, doi:10.1186/1471-2164-14-800 (2013).

72 Zeder, M. A. Domestication and early agriculture in the Mediterranean Basin: Origins, diffusion, and impact. *Proc Natl Acad Sci U S A* **105**, 11597-11604, doi:10.1073/pnas.0801317105 (2008).

73 Petrovska, L. *et al.* Microevolution of Monophasic *Salmonella* Typhimurium during Epidemic, United Kingdom, 2005–2010. *Emerging Infectious Diseases* **22**, 617-624, doi:10.3201/eid2204.150531 (2016).

74 Kroger, C. *et al.* The transcriptional landscape and small RNAs of *Salmonella enterica* serovar Typhimurium. *Proc Natl Acad Sci U S A* **109**, E1277-1286, doi:10.1073/pnas.1201061109 (2012).

75 Kingsley, R. A. *et al.* Epidemic multiple drug resistant *Salmonella* Typhimurium causing invasive disease in sub-Saharan Africa have a distinct genotype. *Genome research* **19**, 2279-2287, doi:10.1101/gr.091017.109 (2009).

76 Kingsley, R. A. *et al.* Genome and transcriptome adaptation accompanying emergence of the definitive type 2 host-restricted *Salmonella enterica* serovar Typhimurium pathovar. *MBio* **4**, e00565-00513, doi:10.1128/mBio.00565-13 (2013).

77 Makendi, C. *et al.* A Phylogenetic and Phenotypic Analysis of *Salmonella enterica* Serovar Weltevreden, an Emerging Agent of Diarrheal Disease in Tropical Regions. *PLoS Negl Trop Dis* **10**, e0004446, doi:10.1371/journal.pntd.0004446 (2016).

78 Zerbino, D. & Birney, E. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research* **18**, 821-829 (2008).

79 Zerbino, D. R. Using the Velvet de novo assembler for short-read sequencing technologies. *Current protocols in bioinformatics Chapter 11*, Unit 11.15, doi:10.1002/0471250953.bi1105s31 (2010).

80 Page, A. J. *et al.* Robust high-throughput prokaryote de novo assembly and improvement pipeline for Illumina data. *Microb Genom* **2**, e000083, doi:10.1099/mgen.0.000083 (2016).

81 Boetzer, M., Henkel, C. V., Jansen, H. J., Butler, D. & Pirovano, W. Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* **27**, 578-579, doi:10.1093/bioinformatics/btq683 (2011).

82 Nadalin, F., Vezzi, F. & Policriti, A. GapFiller: a de novo assembly approach to fill the gap within paired reads. *BMC Bioinformatics* **13**, 1-16, doi:10.1186/1471-2105-13-s14-s8 (2012).

83 Kolmogorov, M., Raney, B., Paten, B. & Pham, S. Ragout-a reference-assisted assembly tool for bacterial genomes. *Bioinformatics* **30**, i302-309, doi:10.1093/bioinformatics/btu280 (2014).

84 Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068-2069, doi:10.1093/bioinformatics/btu153 (2014).

85 Li, H. Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. *Bioinformatics* **32**, 2103-2110, doi:10.1093/bioinformatics/btw152 (2016).

86 Sommer, D. D., Delcher, A. L., Salzberg, S. L. & Pop, M. Minimus: a fast, lightweight genome assembler. *BMC Bioinformatics* **8**, 64, doi:10.1186/1471-2105-8-64 (2007).

87 Page, A. J. *et al.* SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. *Microb Genom* **2**, e000056, doi:10.1099/mgen.0.000056 (2016).

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

89 Hayden, H. S. *et al.* Genomic Analysis of *Salmonella enterica* Serovar Typhimurium Characterizes Strain Diversity for Recent U.S. Salmonellosis Cases and Identifies Mutations Linked to Loss of Fitness under Nitrosative and Oxidative Stress. *MBio* **7**, doi:10.1128/mBio.00154-16 (2016).

90 Hunt, M. *et al.* ARIBA: rapid antimicrobial resistance genotyping directly from sequencing reads. doi:10.1101/118000 (2017).

91 Zankari, E. *et al.* Identification of acquired antimicrobial resistance genes. *J Antimicrob Chemother* **67**, 2640-2644, doi:10.1093/jac/dks261 (2012).

92 Chen, L., Zheng, D., Liu, B., Yang, J. & Jin, Q. VFDB 2016: hierarchical and refined dataset for big data analysis--10 years on. *Nucleic Acids Res* **44**, D694-697, doi:10.1093/nar/gkv1239 (2016).

93 Carattoli, A. *et al.* In silico detection and typing of plasmids using PlasmidFinder and plasmid multilocus sequence typing. *Antimicrob Agents Chemother* **58**, 3895-3903, doi:10.1128/aac.02412-14 (2014).

94 Inouye, M. *et al.* SRST2: Rapid genomic surveillance for public health and hospital microbiology labs. *Genome Med* **6**, 90, doi:10.1186/s13073-014-0090-6 (2014).

95 Otto, T. D., Dillon, G. P., Degrave, W. S. & Berriman, M. RATT: Rapid Annotation Transfer Tool. *Nucleic Acids Res* **39**, e57, doi:10.1093/nar/gkq1268 (2011).

96 Carver, T. J. *et al.* ACT: the Artemis Comparison Tool. *Bioinformatics* **21**, 3422-3423 (2005).

97 El-Gebali, S. *et al.* The Pfam protein families database in 2019. *Nucleic Acids Research*, gky995-gky995, doi:10.1093/nar/gky995 (2018).

98 Finn, R. D., Clements, J. & Eddy, S. R. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res* **39**, W29-37, doi:10.1093/nar/gkr367 (2011).

99 Arndt, D., Marcu, A., Liang, Y. & Wishart, D. S. PHAST, PHASTER and PHASTEST: Tools for finding prophage in bacterial genomes. *Brief Bioinform*, doi:10.1093/bib/bbx121 (2017).

100 Croucher, N. J. *et al.* Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Research* **43**, e15-e15, doi:10.1093/nar/gku1196 (2015).

101 Hadfield, J. *et al.* Phandango: an interactive viewer for bacterial population genomics. *Bioinformatics*, doi:10.1093/bioinformatics/btx610 (2017).

102 Didelot, X. & Wilson, D. J. ClonalFrameML: Efficient Inference of Recombination in Whole Bacterial Genomes. *PLOS Computational Biology* **11**, e1004041, doi:10.1371/journal.pcbi.1004041 (2015).

103 Page, A. J. *et al.* Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* **31**, 3691-3693, doi:10.1093/bioinformatics/btv421 (2015).

104 Kurtz, S. *et al.* Versatile and open software for comparing large genomes. *Genome Biol* **5**, R12 (2004).

105 Adriaenssens, E. & Brister, J. R. How to Name and Classify Your Phage: An Informal Guide. *Viruses* **9**, doi:10.3390/v9040070 (2017).

106 Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150-3152, doi:10.1093/bioinformatics/bts565 (2012).

107 Agren, J., Sundstrom, A., Hafstrom, T. & Segerman, B. Gegenees: fragmented alignment of multiple genomes for determining phylogenomic distances and genetic signatures unique for specified target groups. *PLoS One* **7**, e39107, doi:10.1371/journal.pone.0039107 (2012).

108 Bin Jang, H. *et al.* Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. *Nat Biotechnol* **37**, 632-639, doi:10.1038/s41587-019-0100-8 (2019).

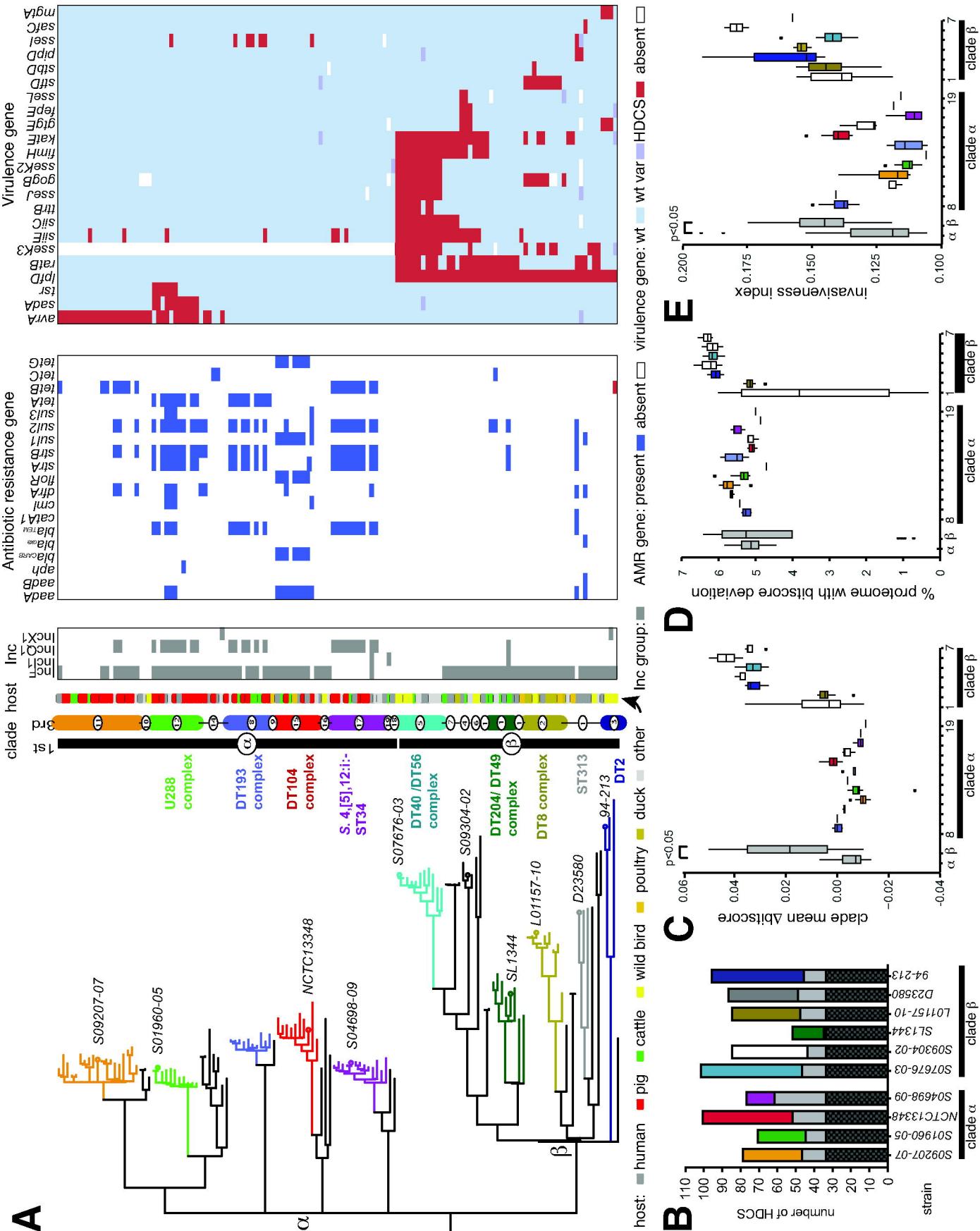
109 O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res* **44**, D733-745, doi:10.1093/nar/gkv1189 (2016).

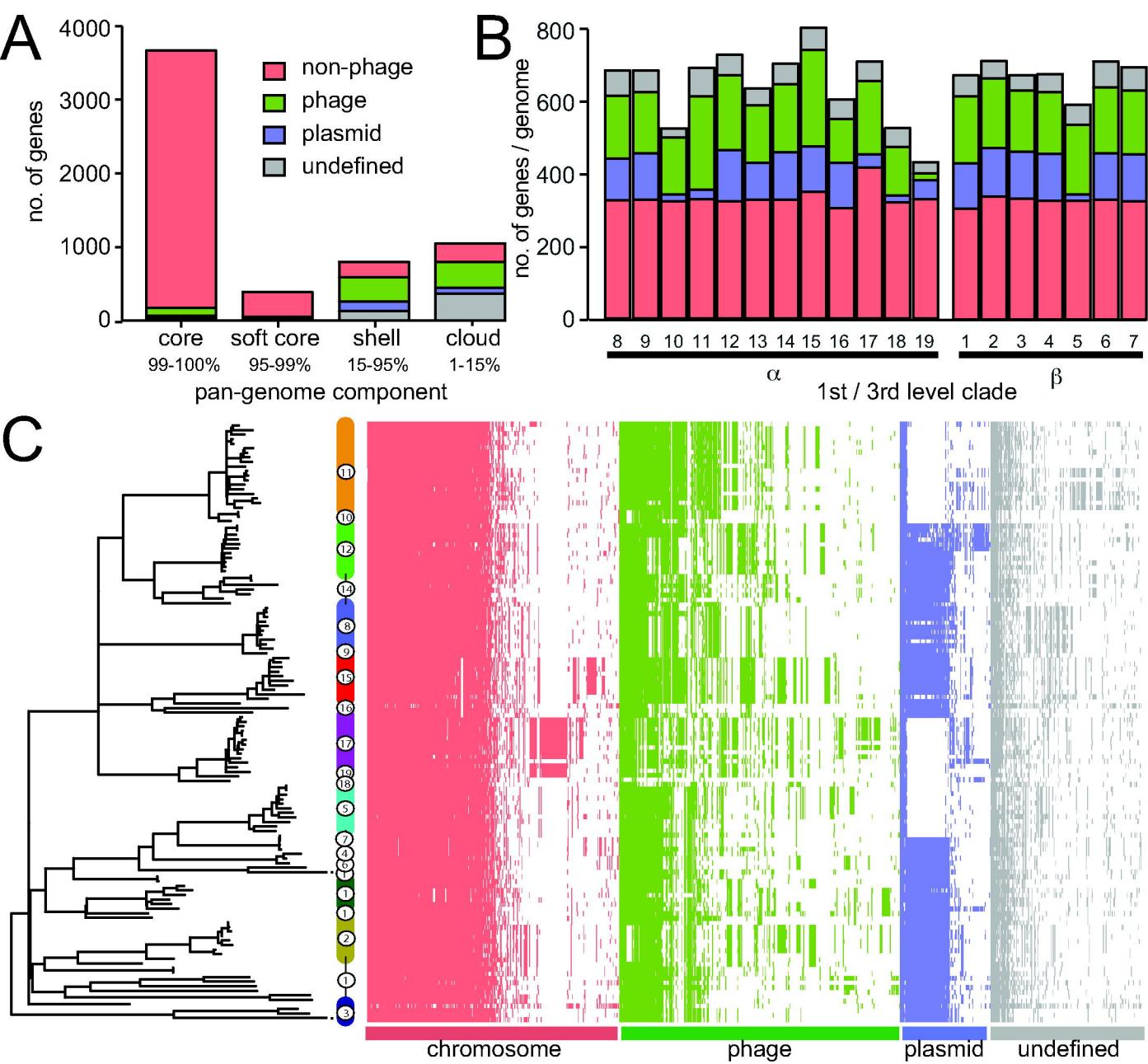
110 Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nature methods* **12**, 59-60, doi:10.1038/nmeth.3176 (2015).

111 Enright, A. J., Van Dongen, S. & Ouzounis, C. A. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res* **30**, 1575-1584, doi:10.1093/nar/30.7.1575 (2002).

112 Nepusz, T., Yu, H. & Paccanaro, A. Detecting overlapping protein complexes in protein-protein interaction networks. *Nat Methods* **9**, 471-472, doi:10.1038/nmeth.1938 (2012).

113 Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* **13**, 2498-2504, doi:10.1101/gr.1239303 (2003).





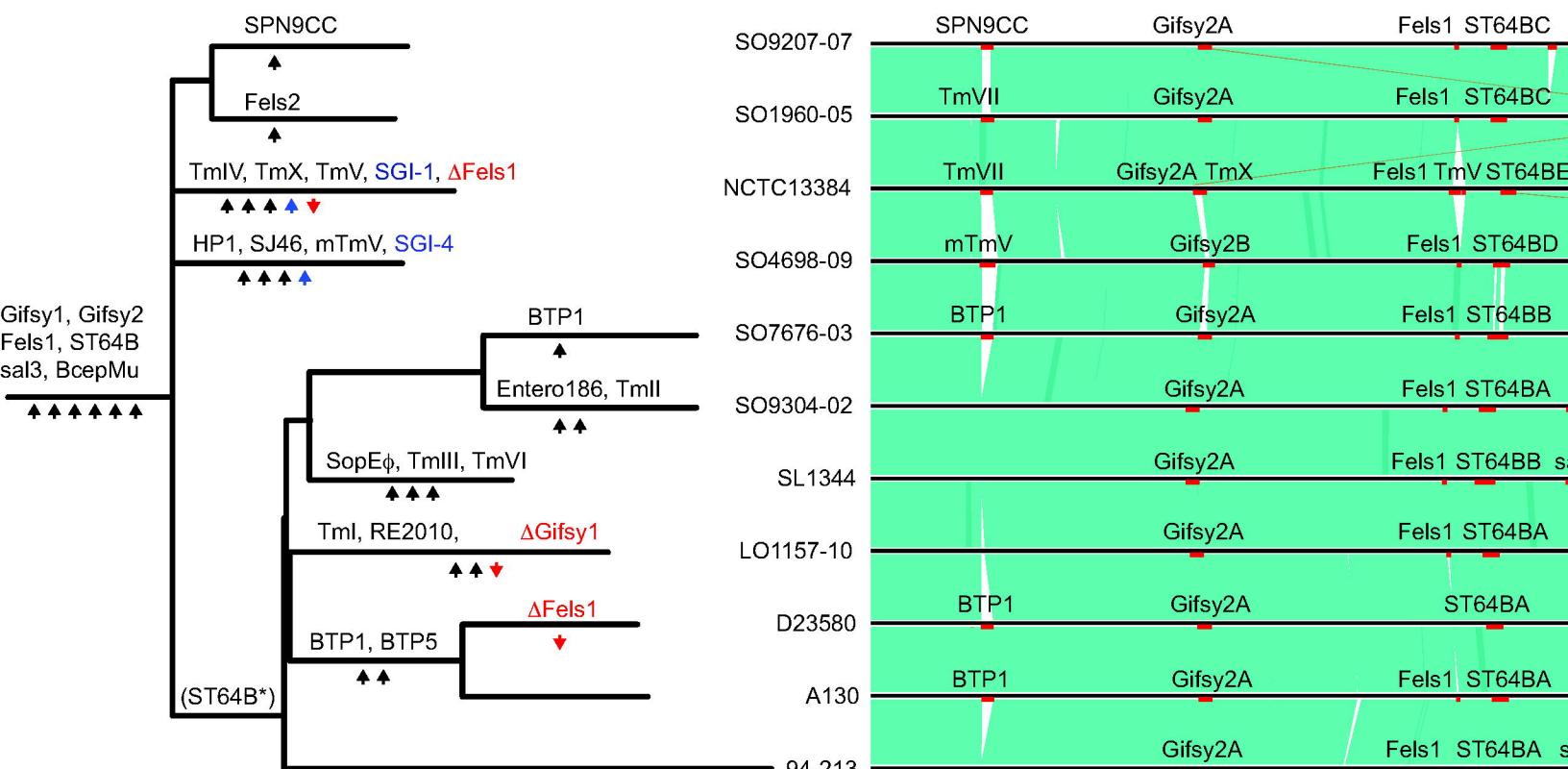


Figure 2

