1 Genome archaeology of two laboratory *Salmonella enterica enterica* sv Typhimurium

2 Julie Zaworski[*†], Oyut Dagva[*§], Anthony W Kingston[*‡], Alexey Fomenkov[*], Richard D.

3 Morgan[*], Lionello Bossi[§], Elisabeth A. Raleigh[*] **orcid.org/0000-0003-0745-9925**

4 [*]New England Biolabs, Ipswich MA, 01938-2723

5 [§]University Paris-Saclay, CEA, CNRS, Institute for Integrative Biology of the Cell (I2BC),

6 91198, Gif-sur-Yvette, France.

7 [†]julie.zaworski@protonmail.com, Optilian, 92010 Montrouge, France; ORCID orcid.org/0000-

8 0003-4744-7208

9 [§]oyutdagva@gmail.com, Université de Lorraine, INRAE, DynAMic, F-54000 Nancy, France;

10 ORCID orcid.org/0000-0001-6337-7639

11 [‡]awkingston@gmail.com, Department of Biology and Mathematics, D'Youville College,

12 Buffalo, NY 14201

13

14 **Reference numbers:** data available in public repositories

15 ER3625: CP067091-CP067092

16 SUB8846059: ER3625 full genome

17 STK005: CP067397-CP067398

18 SUB8846115: STK005 (reisolate LB5000)

19

20 **Running Head:** *Salmonella* LB5000 and ER3625 genomes

21 **Keywords:** *Salmonella* LB5000, prophages, hybrid genome, transduction scars, H antigen, Gifsy

22 prophages, 6-methyladenine

23

24 Corresponding author:

25 Elisabeth A. Raleigh

26 Research Department

27 New England Biolabs

28 240 County Road

29 Ipswich, MA 01935

30 978 380 7238

31 raleigh@neb.com

Abstract

The Salmonella research community has used strains and bacteriophages over decades, exchanging useful new isolates among laboratories for study of cell surface antigens, metabolic pathways and restriction-modification studies. Here we present the sequences of two laboratory *Salmonella* strains (STK005, an isolate of LB5000; and its descendant ER3625). In the ancestry of LB5000, segments of ~15 and ~42 kb were introduced from *Salmonella enterica* sv Abony 103 into *Salmonella enterica* sv Typhimurium LT2, forming strain SD14; this strain is thus a hybrid of *S. enterica* isolates. Strains in the SD14 lineage were used to define flagellar antigens from the 1950s to the 1970s, and to define three restriction-modification systems from the 1960s to the 1980s. LB5000 was also used as host in phage typing systems used by epidemiologists. In the age of cheaper and easier sequencing, this resource will provide access to the sequence that underlies the extensive literature.

INTRODUCTION

Bacteria, phages and conjugal plasmids are in a constant arms race to develop respectively defense against DNA invasion and methods of evading defenses. Bacteria express variable surface features to thwart entry (e. g. flagellae, outer membrane proteins, O-antigen sugar chains) and internal limitation mechanisms to abort establishment (e. g. restriction systems and abortive-infection suicide factors) (DORON *et al.* 2018; AZAM AND TANJI 2019). Phages and plasmids may defeat host restriction-modification (RM) systems with antirestriction proteins, which frequently block type I DNA restriction-modification enzymes (e. g. (ZAVILGELSKY AND RASTORGUEV 2009; SERFIOTIS-MITSA *et al.* 2010; BALABANOV *et al.* 2012; ROBERTS *et al.* 2012; PIYA *et al.* 2017; GONZALEZ-MONTES *et al.* 2020). Low-specificity DNA modification enzymes

54 expressed only for the entering molecule are also found in prophages (MURRAY *et al.* 2018) and
55 plasmids (FOMENKOV *et al.* 2020).

56 Classical studies of horizontal exchange processes within and between genera frequently
57 employed strains in the ancestry of LB5000, particularly for study of cell surface antigens and
58 restriction systems. Early hybrid serovar ancestors were constructed by transduction from *S.*
59 *enterica* sv Abony into *S. enterica* sv Typhimurium. These studies defined flagellar antigen
60 types, the phenomenon of flagellar phase variation (JOYS AND STOCKER 1965; JOYS AND
61 STOCKER 1969) and the use of O-antigen as phage receptor (ORNELLAS AND STOCKER 1974;
62 STOCKER *et al.* 1980). Three restriction systems were identified using this hybrid lineage
63 (BULLAS AND COLSON 1975). In LB5000 itself, the three restriction systems of strain LT2 were
64 inactivated (BULLAS AND RYU 1983), allowing use for epidemiologic purposes as a host for
65 phages in the Anderson phage typing system (SCHMIEGER 1999). Intergeneric conjugal transfer
66 of chromosomal markers was studied using *Escherichia coli* K-12 derivatives and *Salmonella*
67 *enterica* sv Typhimurium lysotypes LT2 and LT7, using F' plasmids and integrated F factors
68 (SANDERSON 1987). LB5000 and relatives are particularly useful as an intermediate host in
69 molecular genetic applications, moving constructions between strains and genera (SANDERSON
70 AND ROTH 1988; SEGALL AND ROTH 1989; CHAUDHURI *et al.* 2009; LERCHE *et al.* 2017; YOON *et*
71 *al.* 2018; GRIEWISCH *et al.* 2020; JONES-CARSON *et al.* 2020).

72 Genome sequences are available for LT2 (MCCLELLAND *et al.* 2001) and LT7
73 (ZAWORSKI *et al.* 2021), but none within this hybrid lineage.

74 Here we present annotated sequences of two *S. enterica* sv Typhimurium variant
75 genomes: STK005 (our isolate of LB5000, from Anca Segall's G251), and ER3625, a derivative
76 of STK005 built in our lab for studies of HGT events. The LB5000 genome carries scars from

77 genetic steps in its descent from LT2, including the replacement of two flagellum-encoding
78 segments of the LT2 genome by segments of *S. enterica* sv Abony, with concomitant loss of
79 prophage Fels2. Recombination events involving two prophages document intralineage allele
80 shuffling. Mutations linked to the amino acid requirement of the strain were identified as well.

81 **MATERIAL AND METHODS**

82 ***Salmonella* growth**: For each strain, a single colony was grown in RB (10 g soy peptone, 5 g
83 yeast extract, 5 g NaCl per liter) with the appropriate antibiotics (kanamycin 40 µg/ml,
84 streptomycin 100 µg/ml) overnight at 37˚C with 250 rpm agitation.

85 **DNA preparation, sequencing and de novo assembly**

86 **DNA extraction:** STK005 and ER3625 were prepared for PacBio RSII or Sequel sequencing.
87 gDNA was extracted with the Monarch Genomic DNA purification kit (#T3010 New England
88 Biolabs; Ipswich, MA, USA). For Oxford Nanopore sequencing of STK005, gDNA was
89 extracted from 1 ml of overnight culture with the kit Monarch HMW DNA Extraction kit
90 (T#3060 New England Biolabs; Ipswich, MA, USA) following the manufacturer instructions for
91 High Molecular Weight DNA extraction from gram-negative bacteria. After the extraction, the
92 HMW DNA was homogenized following by heat and pipetting before QC. A clean peak > 60 kb
93 was found on analysis with the Genomic DNA ScreenTape® (Agilent Technology, Santa Clara,
94 CA).

95 **Library preparation:** For the PacBio RSII sequencing platform, STK005 SMRTbell libraries
96 were constructed from genomic DNA samples sheared to between 10 and 20 kb using the G-
97 tubes protocol (Covaris; Woburn, MA, USA), end repaired, and ligated to PacBio hairpin
98 adapters. Incompletely formed SMRTbell templates and linear DNAs were digested with a
99 combination of Exonuclease III and Exonuclease VII (New England Biolabs; Ipswich, MA,
100 USA). DNA qualification and quantification were performed using the Qubit fluorimeter
101 (Invitrogen, Eugene, OR) and 2100 Bioanalyzer (Agilent Technology, Santa Clara, CA). The
102 SMRTbell library was prepared according to PacBio sample preparation protocol sequenced with

103 C4-P6 chemistry with a 300 min collection time. A total of 751,460 polymerase reads with N50
104 length of 17,626 bp was obtained (approximately 360X coverage) before filtering.

105 ER3625 libraries were sequenced using the PacBio Sequel I platform. Briefly, SMRTbell
106 libraries were constructed from genomic DNA samples following the PacBio protocol for Sequel
107 using the kit 100-938-900. DNA qualification and quantification were performed using the Qubit
108 fluorimeter (Invitrogen, Eugene, OR) and 2100 Bioanalyzer (Agilent Technology, Santa Clara,
109 CA). The resulting library, of average size 11,640 bp, was prepared for binding following the
110 PacBio guidelines generated by SMRT Link and run on a Sequel I machine.

111 The STK005 Nanopore library was prepared with 1.7 µg of HMW DNA following the
112 Genomic DNA by ligation (SQK-LSK109 protocol) from Oxford Nanopore. The adapter ligation
113 was performed for 1 h instead of 10 min. The flow cell was primed and loaded with ~100 ng of
114 library following the same documentation. After a 21h run on a second-use FLO-MIN106 flow
115 cell with the GridIONxs apparatus, 449K reads were obtained, with 27.39 kb read length
116 basecalled N50. The basecalling was performed on the fly with the high-accuracy basecalling
117 model from Guppy (version 3.2.10)

118 **Sequence assembly:** The STK005 (PacBio RSII reads) were assembled using RS_HGAP3
119 program with parameters: minimum subread length 5,000 bp, minimum polymerase quality 0.8,
120 Minimum polymerase read length 10,000 bp, 5 Mbp genome size and other parameters as default
121 settings. The three resulting contigs, of size 4,802,499 bp, 103,465 bp and 8,138 bp, were
122 circularized with Circlator (HUNT *et al.* 2015) with CCS generated with parameters minimum
123 full passes 3, minimum predicted accuracy 95, minimum read length of insert 1000 bp and the
124 other parameters as default. The smallest contig was a subset of the chromosome contig.

125 PacBio polishing for STK005 was done by running the RS_Reseq program twice, with
126 Minimum subread length 500, minimum polymerase read quality 75 and then 80, and 1500 bp
127 for Minimum polymerase read length. The final assembly is composed of a 4,796,208 bp
128 chromosome and a 93,938 bp plasmid.

129 The STK005 Nanopore reads were corrected and filtered using Canu (KOREN *et al.*
130 2017), the output corrected reads were then trimmed and assembled with Canu (version 1.4). A
131 first correction step was run with parameters minReadLength = 10000, stopOnReadQuality=false
132 and other as default. Then trimming and assembly were run with genomeSize= 5m and other
133 parameters as default. Two contigs of length 162,546 and 4,863,042 bp corresponding to the
134 plasmid and chromosome.

135 In parallel, the reads were mapped on the PacBio de novo assembly using minimap2
136 (version 2.17) with default parameters (LI 2018). In both cases, no inversion in between the
137 Gifsy prophages was observed.

138 PacBio Sequel assembly of ER3625 used the RS_HGAP4 program with parameters: 5
139 Mb expected genome size, minimum mapped length 5 kb, minimum mapped concordance 70
140 and other parameters as default settings. The 2 contigs obtained were circularized with Circlator
141 (version 1.5.5). Polishing was done by running the RS_Reseq program with minimum mapped
142 length 500, yielded 19 corrections. The final assembly is composed of a 4,796,355 bp
143 chromosome (coverage 541x) and a 93,939 bp plasmid (2,247x coverage).

144 **Modification motifs:** DNA motifs and degree of modification were generated using InterPulse
145 Duration (IPD) Ratios analysed with RS_Modification_and_Motif_Analysis from PacBio as in
146 (FLUSBERG *et al.* 2010; CLARK *et al.* 2012).

147

148 **Genome annotation:** Genbank deposits STK005 (chromosome CP067397, plasmid CP067398)
149 and ER3625 (chromosome CP067091 and plasmid CP067092) genomes were annotated with the
150 Genbank PGAP pipeline (TATUSOVA *et al.* 2016; HAFT *et al.* 2018). In-process analysis
151 employed annotations in Geneious (2018 version in steps: Menu Align and Assemble; Align
152 Whole Genomes; Tab Alignment View; Save as editable copy; Menu Annotate and Predict;
153 Annotate from Database) using LT2 [NC_003197.2] and Abony [NZ_CP007534.1] and segment
154 extractions from them as reference genomes in a local database folder.

155 **Data availability**

156 Supplementary material is available at figshare DOI... Table S1 "Pedigree LB5000 STK005" traces
157 the descent of LB5000 and some relatives through the literature since 1959; with wild ancestor,
158 donor strain, recipient strain, renaming, construction reference and mode of construction. File S1
159 "STK005 variants" contains 5 Excel spreadsheets documenting positions at which STK005
160 sequence differs from the genome sequences representing two ancestral components. For three
161 genome segments derived from LT2, variant nucleotide positions of LT2 (NC_003197.2) are
162 listed, with consequences for amino acid changes; for two short segments derived *Salmonella*
163 *enterica* serovar Abony SW803 (H1, 15.4 kb and H2, 41.7 kb), all annotated loci are listed as well,
164 since few positions vary from the Abony reference. The reference segments H1 and H2 were
165 extracted from *Salmonella enterica subsp. enterica* serovar Abony str. 0014 - NZ_CP007534.1:
166 for transduction 1 of Table S1, H1--nt 3956473-3942278 (reversed); for transduction 2, H2--nt
167 4713943-4725447(catenated)1-18493 (the segment crosses the numerical origin, so is catenated).
168 File S2 "Analysis of 11 kb high-similarity region in the Gifsy prophages" describes and displays
169 alignments (referred to as "Embedded figs") that resolve allele configurations in Gifsy prophages.
170 File S3 "STK005 genotype inferences from sequence" describes the recorded genotype; explains

171 assignment of flagellar genotype; describes the protocol used to generate the variant list found in
172 File S1; lists protein changes that account for the genotype; and points out some additional possible
173 changes. File S4 "PCR survey for 42bp insertion in isolates of *S. enterica* sv Typhimurium str
174 LT2" demonstrates the insertion is present in Gifsy-1 and conserved in most LT2 strains.

175 All the sequencing data can be found in the BioProject PRJNA605961. The genome
176 sequences of STK005 and ER3625 are available from GenBank with the accession
177 numbers (CP067397-CP067398) and (CP067091-CP067092), respectively. All the raw data are
178 available under the SRA ID SRR13390529, (SRR13404095, SRR13404094, SRR13404093,
179 SRR13404092, SRR13404091) and SRR13787361 for ER3625, STK005 RSII and STK005
180 Nanopore respectively. The methylation data are available as supplementary files on NCBI under
181 the IDs SUPPF_0000003906 and SUPPF_0000003907. Both strains are available from New
182 England Biolabs.

183 **RESULTS AND DISCUSSION**

184 **Strain engineering scars: *Salmonella enterica* subsp. *enterica* serovar Abony DNA in**
185 **LB5000**

186 The engineering history of LB5000 (Figure 1) includes two transductions from *Salmonella*
187 serovar Abony strain 803 (SPICER AND DATTA 1959) (see details and references in Table S1).
188 These two transductions left scars in the form of high-divergence segments in the genome
189 surrounding the flagellar loci H1 and H2 (Figure 2A). In *Salmonella enterica*, both H1 and H2
190 flagellar antigens are highly variable in wild lineages in the population; in addition, each
191 individual lineage is capable of switching from expressing its H1 antigen to its H2 or the reverse,
192 a phenomenon known as phase variation (SIMON *et al.* 1980). This case of phase variation

193 involves reversible site-specific rearrangement of regulatory signals that enable expression from
194 only one of two distant loci. In LB5000, both of these loci derive from *S. enterica* sv Abony.
195

196 **H1, phase 1 flagellin region transduction patch**: The H1 phenotype depends on the allelic state
197 of the gene *fliC*, within the H1 box of Figure 2A. This patch is the result of a first transduction
198 mediated by P22: recipient strain LT2 SL270 replaced the H1-i 1.2 allele of LT2 with H1-b from
199 *S.* Abony SW803 (SPICER AND DATTA 1959; JOYS AND STOCKER 1969) (see Table S1). Allele
200 H1-b was bequeathed to LB5000 ancestor SD7. The highly divergent interval (LT2 locus_ID
201 STM1946 (*uvrC*) to STM1961 (*fliD*)) includes two unshared IS insertions (IS3 in STK005,
202 IS200 in LT2) and encompasses the *fliC* flagellin gene. In contrast, this region in LB5000
203 (JJB81_10090 to JJB81_10010; nt 2037077 -> 2052486) aligns with only three mismatches to *S.*
204 Abony (SEEA0014_RS19195 to SEEA0014_RS19270; nt 3941779 -> 3957191); see below and
205 File S1.

206 **H2, phase 2 flagellin region transduction patch**: The H2 phenotype is determined by the
207 allelic state of the gene *fljB*, a *fliC* homolog. When the adjacent regulatory gene *fljA* is expressed,
208 it mediates repression of distant H1 expression; when the local DNA segment is inverted, *fljA*
209 and *fljB* are not expressed, and repression of distant gene *fliC* is lifted. Wild type for this LT2
210 isolate is H2-1.2; *S.* Abony SW803 was H2-enx. H2-enx was bequeathed to LB5000 ancestor
211 SD14.

212 These genetic changes are reflected in a second divergent region of the alignment (H2
213 black rectangle in Figure 2A; closer view in Figure 2B). Genes present in LT2 but absent in
214 STK005 include the Fels2 prophage (STM2694 to STM2740) and the flanking genes STM2741

215 to STM2769. Present but divergent are genes JJB81_13790 to JJB81_13790 when aligned with
216 LT2 locus_ID STM2679 and STM3779. We infer that this divergent region represents genomic
217 replacement of 106 kb of *S.* sv Typhimurium LT2 by 41 kb of *S.* sv Abony 803 during the
218 second transduction (Figure 1, panel A). Confirmation of the inference is illustrated in Figure
219 2C: 41 kb of our sequence is very similar (3 insertions, 14 mismatches) to a 41 kb segment of
220 NZ_CP007534 (sequence of *S. enterica* sv Abony strain 0014 (NCTC 6017; ATCC BAA-2162)
221 (TIMME *et al.* 2013); see File S1 sheet 4). Both flanks of the long deletion show high divergence
222 from LT2, and the *fljB* region is 100% identical to the sequenced *S.* sv Abony genome. Prophage
223 Fels-2 is not present in Abony serovar isolates, consistent with the present configuration.

224 ***hin* gene, STM2772, JJB81_13890**: The Hin protein mediates inversion of a segment including
225 its own gene and a promoter of *fliC* (H2, phase 2) and the *fljA* (repressor of *fliC* (H1, phase 1))
226 (SIMON *et al.* 1980). The STK005 *hin* homolog (JJB81_13890) is within the H2 transduction
227 patch. It is similar to the Abony SEEA0014-RS00045 gene except for the 3 single nucleotide
228 insertions; these insertions are all shared among LT2, STK005 and ER3625. In addition to the
229 closer identity to *S. enterica* sv Abony, the *hin* gene has suffered an inversion during the descent
230 of STK005 from LT2 (also found in ER3625). Considering the expected protein activity, this
231 reorganization is evidence of expected function.

232 **ER3625 engineering and variation from STK005**: ER3625 was engineered in our laboratory to
233 create a non-pathogenic recipient (HOISETH AND STOCKER 1981) with no Mrr restriction, to
234 pursue work on R-M systems in HGT using the system previously used in our laboratory
235 (KINGSTON *et al.* 2015; KINGSTON *et al.* 2017). Strain engineering required several steps with a
236 variety of methods (Figure 1B).

237 For future work we wish to verify that no off-target changes had occurred during lamda-
238 Red-mediated genome engineering or accompanying transduction. The *mrr* engineered locus
239 carries the designed sequence (*npt* flanked by FRT sites replacing the gene). The *aroA* gene was
240 deleted, but the scar of the FRT site recombination is aberrant. This DNA segment is annotated
241 by NCBI pipeline as JJB80_05020 (DUF1317 domain-containing protein) and JJB80_05025
242 (DUF1382 protein family). No SNPs were identified in genes between STK005 and ER3625 for
243 either the chromosome or the plasmid.

244 **Structural complexity and intralineage prophage recombination**

245 Gifsy-1 and Gifsy-2 are two LT2 prophages located between 2,728,977 and 2,776,819
246 (STM2584 to STM2636) and between 1,098,231 and 1,143,702 (STM1005 to STM1056)
247 respectively. These two related prophages are patchily similar to each other; while overall DNA
248 identity is 44.1%, this ~42 kb region is interspersed with highly divergent segments and
249 unrelated insertions (Figure 3). Of concern are two long regions of high identity: a ~12 kb stretch
250 of > 98% identity in the regions expressed early in lytic infection: recombination, regulators
251 related to CI, CII and Cro of lambda, and replication; and an 8 kb patch of ~ 78% identity at the
252 other end of prophages.

253 These long regions with very high identity are of concern for both technical and
254 biological reasons. First, long repeats are an assembly challenge even with long-read
255 technologies. Second, recombination within the host may generate new gene combinations that
256 are then transmissible to new hosts on prophage induction. Extensive mosaicism among tailed
257 phages is well known (GROSE AND CASJENS 2014); and creation of a new mosaic by
258 recombinational marker rescue from defective prophages has been noted before (DE PAEPE *et al.*

259 2014). Evidence of recombination within a laboratory lineage was found in *E. coli* ER2796
260 (ANTON *et al.* 2015).

261 **Resolution of assembly errors:** De novo assembly of the STK005 genome initially yielded
262 (artifactual) chromosome inversion. On examination of reads aligned with the two Gifsy regions,
263 many poorly or partially mapped reads and an important number of SNPs were seen, suggesting
264 imperfect mapping. A reassembly using only long CCS (> 5 kb; 47.7x coverage) was then
265 created with HGAP3, then circularized with Circlator. This assembly did not show a genomic
266 inversion relative to LT2. Further polishing to resolve SNPs used RS_reseq with CCS > 500 bp
267 (220.9x coverage). This time, few reads were mismapped at the Gifsy regions, confirming the
268 absence of the inversion in STK005 genome.

269 In parallel, Oxford Nanopore was used to generate long reads that could cover 40-45 kb
270 regions and reinforce confidence in the structural organization of the assembly. Reads >10 kb
271 were used to generate a de novo assembly with Canu (version 1.4). This assembly agreed with
272 the final PacBio assembly, i.e. without inversion of the genome. Mapping of the Nanopore reads
273 on the PacBio assembly confirmed the structure.

274 **Prophage recombination:** This 12 kb region of near identity codes for proteins involved in
275 homologous recombination, as well as regulation and replication. The recombination proteins are
276 related to those of the *E. coli* Rac prophage RecET recombination cluster; also present are
277 homologs of Lambda Orf and Rap, which also affect recombination. RecE and RecT provide
278 homologous recombination systems that can substitute for the host RecABCD system (LEMIRE *et*
279 *al.* 2008; MURPHY 2016). Rap (originally named NinG) facilitates recombination near linear
280 ends via the Lambda Red pathway (TARKOWSKI *et al.* 2002), acting as a structure-specific
281 nuclease (SHARPLES *et al.* 2004).

282 Three segments within this long region posed interpretation problems: assembly error or
283 recombination during the lineage leading to STK005. Our interpretation is that the reference LT2
284 sequence is in error in one case, and both reciprocal recombination and gene conversion have
285 affected this region (see Figure 3 and Files S2 and S4).

286 First is the appearance of a novel allele of Gifsy-1 *recET*, including a reading-frame-
287 preserving insertion of 42 nt (14 aa) in the middle of the RecE CDS and 9 SNPs flanking it (File
288 S2 Embedded fig2). The Gifsy-1 allele of *recET* found here appears to be ancestral: it is also
289 found in the Gifsy-1 sequence of *S. enterica* sv Typhimurium strain 14028S (Genbank
290 CP034479.1; File S2 Embedded fig2). (FIGUEROA-BOSSI *et al.* 2001; LEMIRE *et al.* 2007). Strain
291 14028S was used extensively in laboratory studies due to much more virulent properties in mice
292 (e.g. (FIGUEROA-BOSSI *et al.* 1997)). PCR tests of 19 LT2 isolates or derivatives confirm the
293 presence of the insertion in all of them except strains specifically cured of Gifsy-1 (File S4
294 Embedded fig1 and Embedded table1). A recent, independently-determined LT2 sequence
295 (CP014051) was obtained using PacBio and Illumina technology. This deposit agrees with our
296 sequence of STK005 across this region.

297 Second, there has been a reciprocal exchange of the sequence between Gifsy-1 and Gifsy-
298 2 prophages (File S2 Embedded fig3A, 3B). This exchange includes the position of the 18-amino
299 acid (51 bp) indel variation in CDSs encoding a DNA break and rejoin activity illustrated in
300 Figure 3B above.

301 Third, gene conversion has resulted in loss of the original Gifsy-2 allele of Rap (NinG),
302 replaced by Gifsy-1 information without change in the original Gifsy-1 (File S2 Embedded
303 fig4A, 4B).

304 Finding evidence of recombination activity within the high-identity region is interesting.
305 Though normally silent as prophages, both Gifsy-borne LT2 homologs of RecET can be
306 activated by mutation to mediate recombination in vivo (LEMIRE *et al.* 2008). Action of these
307 phage recombination systems may promote the mosaic genomic relationships among phages that
308 has been found (MARTINSOHN *et al.* 2008). The two events found here are compatible with that
309 role.

310 **Recorded genotype and point mutations**

311 **Metabolic markers:** LB5000 was originally described as auxotrophic for methionine,
312 tryptophan and leucine with the genotype *metA22 metE551 trpC2 ilv-452*. We identified
313 mutations likely responsible for the reported phenotypes; we also found mutations in other genes
314 involved in amino acid synthesis, see File S3.

315 Concerning methionine production, changes are found in *metA* (*metA22*) and *metE*
316 (*metE551*). In *met*A (STM4182 in LT2 and JJB80_21020 in STK005), a 10 nt deletion at the
317 protein C-terminus truncates it by 9 amino acids. A SNP in *metE* (G877A) leads to G293R
318 protein substitution. The *trpC2* allele carries nt change G545A leading to G->D amino acid
319 substitution. For isoleucine and leucine auxotrophies, SNPs were found in several genes. Coding
320 changes were a G > A mutation at position 418 in *ilvC* (G140S amino acid change) and a C to T
321 transition in *leuA* at position 424 (D142N).

322 Considering other amino acid pathways, *proB* and *proC* of proline synthesis have
323 mutations respectively C > T at position 535 (P179S) and C > T at position 317 (A106V).
324 However, our auxotrophy tests found that the strain requires methionine, valine/isoleucine and
325 tryptophan/phenylalanine/tyrosine but can grow without addition of proline in the media (data
326 not shown). The synthesis of proline may be mediated via *putA*, which has two silent mutations.

327 **Genotype and phenotype of methyltransferases and their associates**: DNA methylation
328 motifs and level of methylation mediated by R-M systems and orphan MTases were analyzed
329 with the PacBio RSII. At least six methylation motifs are known in ancestral strain *Salmonella*
330 *enterica* serovar Typhimurium LT2 ((ROBERTS *et al.* 2015); REBASE Organism number 18099).
331 Three motifs result from enterobacterial "orphan" methyltransferases (M; M.SenLT2 Dam,
332 M.SenLT2 Dcm and M.SenLT2IV; (ROBERTS *et al.* 2015)), and three are associated with
333 restriction-modification (RM) phenomena (M.SenLT2I, M.SenLT2II and M.SenLT2III). Two of
334 these RM systems have been well characterized. Characterization began with a transfer isolate
335 SD14 (Stocker lab strain SL1027). M.SenLT2I (known in the literature as the StyLT or StyLTI
336 RM system (COLSON AND COLSON 1971)) is a Type III enzyme, and M.SenLT2II (StySB or
337 StyLTII (FULLER-PACE *et al.* 1984)) is Type I. The third system, RM.SenLT2III (StySA), is
338 known to confer m6A modification (HATTMAN *et al.* 1976) but its protection/restriction
339 mechanism of action remains unclear.

340 The LB5000 RM genotype is listed as *hsdL6 hsdSA29 hsdSB121* for StyLT, StySA and
341 StySB. We identified appropriate mutations consistent with the R^-M^+ phenotype (Table 1 for M
342 activity; File S3 for genotype discussion). For StyLT and StySB, sense mutations leading to
343 amino acid changes were identified in the endonuclease genes, as well as silent mutations. Two
344 G->A changes (position 393 and 2,749 of STM0358) are present in the *res* gene of StyLT, the
345 second one leading to a D->N amino acid change in the C-terminal domain. For StySB, two G-
346 >A changes (position 2,769 and 2,473 of STM4526) are present in *hsdR*, one resulting in an R-
347 >C amino acid change in the C-terminal domain.

348 The LB5000 StySA R^-M^+ phenotype was reported to be unstable (TSAI *et al.* 1989), and
349 we found our STK005 isolate to be phenotypically StySA R^-M^-. The genetic position of StySA

350 was mapped close to and counterclockwise of the *hsdSB* locus (COLSON AND VAN PEL 1974); we
351 identified a methyltransferase homolog (JJB81_22590) in an appropriate location, but found no
352 coding change relative to STM4495. The genetic requirements for action of this system are under
353 study.

354 In Table 1, we also find partial methylation for the conserved orphan methyltransferase
355 M.SenLT2IV, as previously observed in three *Salmonella* species (PIRONE-DAVIES *et al.* 2015).
356 Methylation of m5C by M.Dcm was not determined here, due to technical limitations of Pacific
357 Bioscience long-read sequencing in detection of m5C (FLUSBERG *et al.* 2010).

358 **ACKNOWLEGDGMENTS**

359 We thank Stanley Maloy and Anca Segall (San Diego State University) for *Salmonella* strains,
360 phages and advice and training in their use. We thank Chloé Baum et Giron Koetsier for
361 technical assistance with the Oxford Nanopore sequencing and high molecular weight DNA
362 extraction.

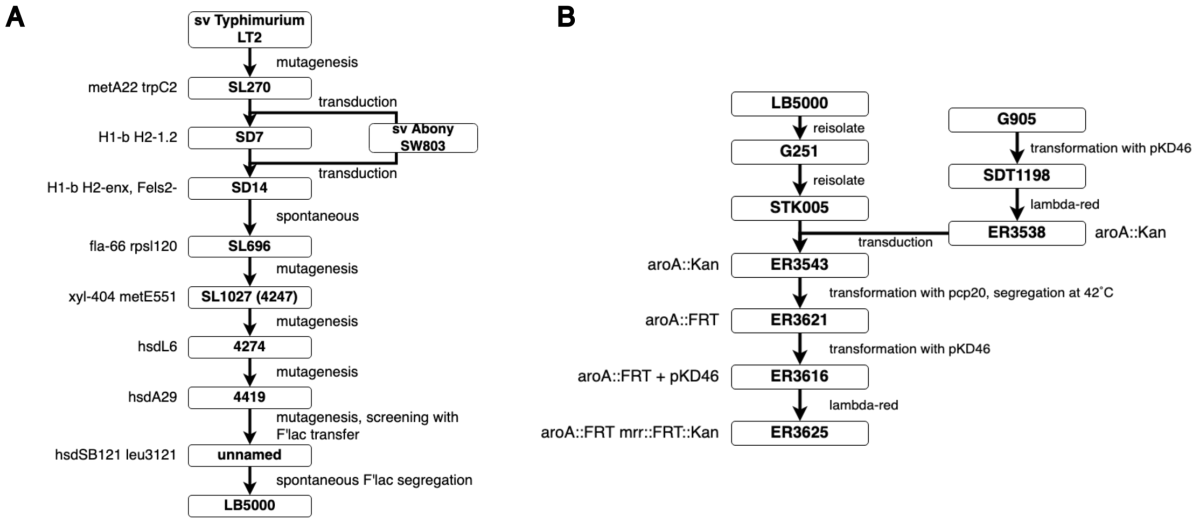
363 New England Biolabs provided support for JZ, OD, TK, AF, RM and EAR. The authors declare
364 no conflicts of interest.

365

366

367

Figures and Legends



368

369 Figure 1: Strain pedigree charts. A. LB5000 B. ER3625. Strain names are boxed. To the right of each construction
370 step (arrows) are the methods used; to the left of each box is the genotype designation assigned as a result of the
371 step. In some cases, a genotype marker was discovered during later investigation. The different strain
372 modifications were done by mutagenesis (random by use of NG), transduction (use of phages) or transformation
373 (plasmids with induction). For more details and references on these steps, refer to Table S1.

374

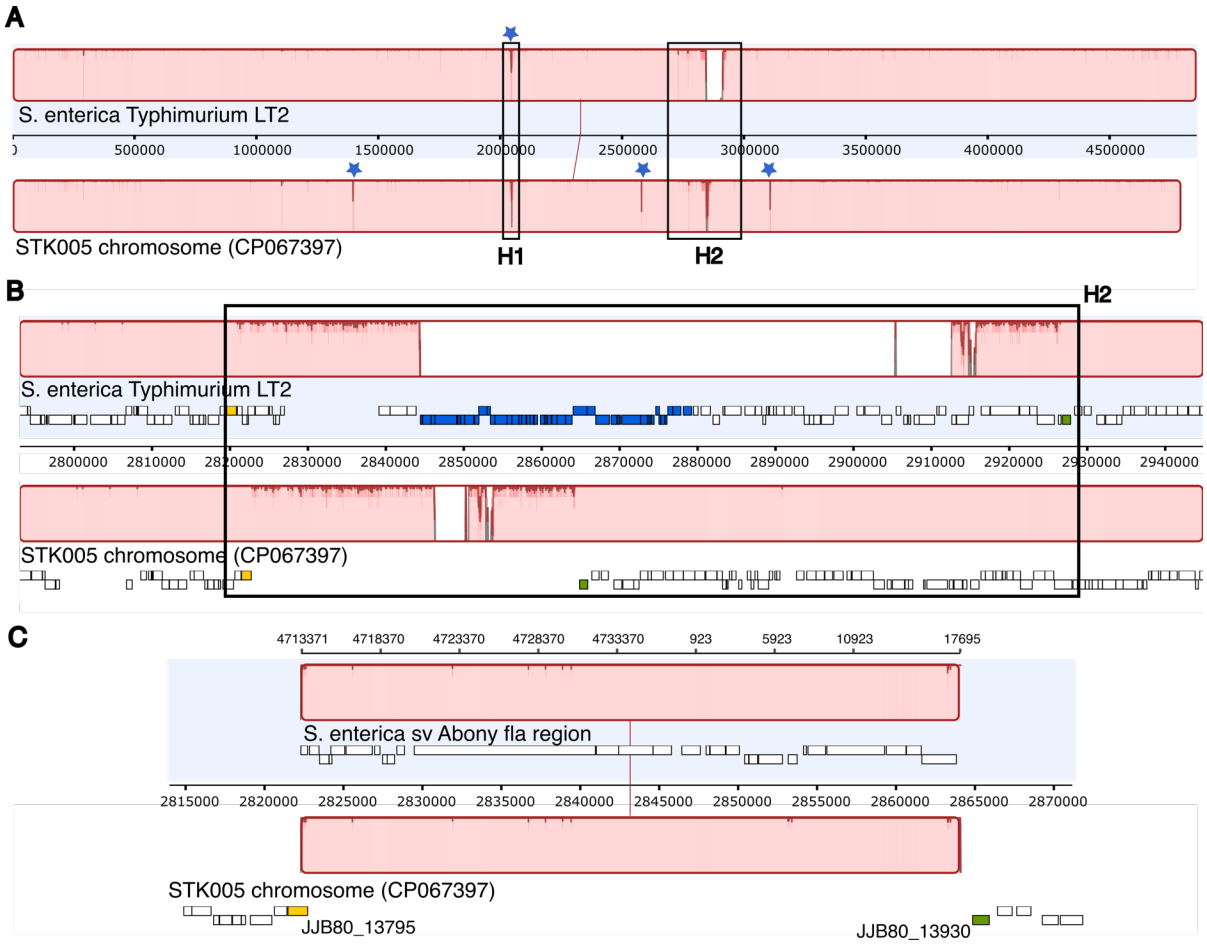


Figure 2: Recombination scars. Panel A is a Mauve alignment of the STK005 and LT2 chromosomes. Blue stars represent IS sequences not shared between the sequences. The flagellar regions H1 and H2 are boxed. Light pink color is for high similarity areas; the white area is a missing segment; and dark red lines dip where mismatches occur. Panel B is a close-up of the H2 region. Panel C is a Mauve alignment of S. Abony H2 (fla) region with STK005. In panel B and C, the yellow and green rectangles below the alignment are landmarks, respectively gene STM2679 (JJB80-13795) and STM2780 (JJB80-13930).

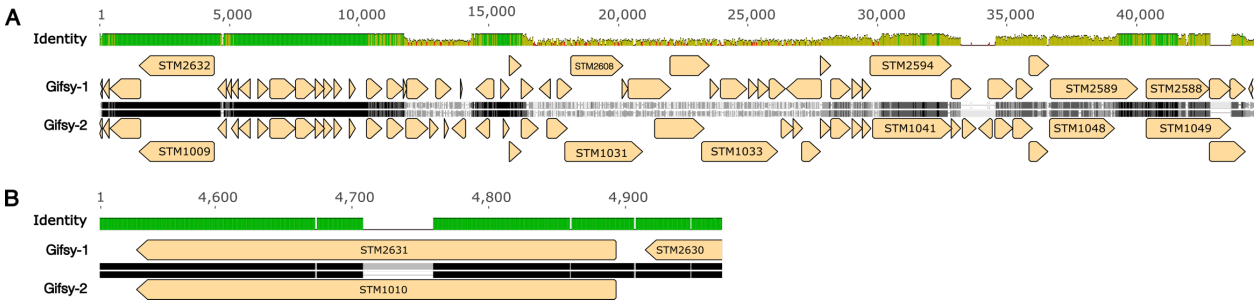


Figure 3: Pairwise alignment of the Gifsy-1 and Gifsy-2 prophages of LT2. Panel A: alignment of the whole prophages. Panel B: close-up of STM2631 and STM1010, revealing a 51 bp indel variation. The identity rows: green=identical bases; yellow=lower identity; red=very low local identity. The sequences Gifsy-1 and Gifsy-2 are in black, white and gray: black for identical bases, white for SNPs, gray for insertion/deletion (indel) variation (corresponding gaps shown with a thin line).

LITERATURE CITED

Anton, B. P., E. F. Mongodin, S. Agrawal, A. Fomenkov, D. R. Byrd *et al.*, 2015 Complete Genome Sequence of ER2796, a DNA Methyltransferase-Deficient Strain of *Escherichia coli* K-12. PLoS One 10**:** e0127446.

Azam, A. H., and Y. Tanji, 2019 Bacteriophage-host arm race: an update on the mechanism of phage resistance in bacteria and revenge of the phage with the perspective for phage therapy. Applied Microbiology and Biotechnology 5**:** 6694-6611.

Balabanov, V. P., V. Y. Kotova, G. Y. Kholodii, S. Z. Mindlin and G. B. Zavilgelsky, 2012 A novel gene, *ardD*, determines antirestriction activity of the non-conjugative transposon Tn5053 and is located antisense within the *tniA* gene. FEMS Microbiology Letters 337**:** 55-60.

402 Bullas, L. R., and C. Colson, 1975 DNA restriction and modification systems in *Salmonella*. III.
403 SP, a *Salmonella potsdam* system allelic to the SB system in S*almonella typhimurium*.
404 Mol Gen Genet 139**:** 177-188.
405 Bullas, L. R., and J. I. Ryu, 1983 *Salmonella typhimurium* LT2 strains which are r^- m^+ for all
406 three chromosomally located systems of DNA restriction and modification. J Bacteriol
407 156**:** 471-474.
408 Chaudhuri, R. R., S. E. Peters, S. J. Pleasance, H. Northen, C. Willers *et al.*, 2009
409 Comprehensive Identification of *Salmonella enterica* Serovar Typhimurium Genes
410 Required for Infection of BALB/c Mice. PLoS Pathogens 5**:** e1000529.
411 Clark, T. A., I. A. Murray, R. D. Morgan, A. O. Kislyuk, K. E. Spittle *et al.*, 2012
412 Characterization of DNA methyltransferase specificities using single-molecule, real-time
413 DNA sequencing. Nucleic Acids Research 40**:** e29-e29.
414 Colson, C., and A. M. Colson, 1971 A new *Salmonella typhimurium* DNA host specificity. J Gen
415 Microbiol 69**:** 345-351.
416 Colson, C. C., and A. A. Van Pel, 1974 DNA restriction and modification systems in Salmonella.
417 I. SA and SB, two *Salmonella typhimurium* systems determined by genes with a
418 chromosomal location comparable to that of the Escherichia coli *hsd* genes. Molecular
419 Genetics and Genomics 129**:** 325-337.
420 De Paepe, M., G. Hutinet, O. Son, J. Amarir-Bouhram, S. Schbath *et al.*, 2014 Temperate phages
421 acquire DNA from defective prophages by relaxed homologous recombination: the role
422 of Rad52-like recombinases. PLoS genetics 10**:** e1004181.
423 Doron, S., S. Melamed, G. Ofir, A. Leavitt, A. Lopatina *et al.*, 2018 Systematic discovery of
424 antiphage defense systems in the microbial pangenome. Science 359.

425 Figueroa-Bossi, N., S. Uzzau, D. Maloriol and L. Bossi, 2001 Variable assortment of prophages
426 provides a transferable repertoire of pathogenic determinants in Salmonella. Molecular
427 Microbiology 39**:** 260-271.
428 Figueroa-Bossi, N. N., E. E. Coissac, P. P. Netter and L. L. Bossi, 1997 Unsuspected prophage-
429 like elements in Salmonella typhimurium. Molecular Microbiology 25**:** 161-173.
430 Flusberg, B. A., D. R. Webster, J. H. Lee, K. J. Travers, E. C. Olivares *et al.*, 2010 Direct
431 detection of DNA methylation during single-molecule, real-time sequencing. Nat
432 Methods 7**:** 461-465.
433 Fomenkov, A., Z. Sun, I. A. Murray, C. Ruse, C. McClung *et al.*, 2020 Plasmid replication-
434 associated single-strand-specific methyltransferases. Nucleic Acids Research 48**:** 12858-
435 12873.
436 Fuller-Pace, F. V., L. R. Bullas, H. Delius and N. E. Murray, 1984 Genetic recombination can
437 generate altered restriction specificity. Proc. Natl. Acad. Sci. USA 81**:** 6095-6099.
438 González-Montes, L., I. del Campo, M. P. Garcillán-Barcia, F. de la Cruz and G. Moncalian,
439 2020 ArdC, a ssDNA-binding protein with a metalloprotease domain, overpasses the
440 recipient *hsdRMS* restriction system broadening conjugation host range. PLoS genetics
441 16**:** e1008750-1008722.
442 Griewisch, K. F., J. G. Pierce and J. R. Elfenbein, 2020 Genetic Determinants of *Salmonella*
443 Resistance to the Biofilm-Inhibitory Effects of a Synthetic 4-Oxazolidinone Analog.
444 Applied and Environmental Microbiology 86.
445 Grose, J. H., and S. R. Casjens, 2014 Understanding the enormous diversity of bacteriophages:
446 the tailed phages that infect the bacterial family Enterobacteriaceae. Virology 468-470**:**
447 421-443.

448 Haft, D. H., M. DiCuccio, A. Badretdin, V. Brover, V. Chetvernin *et al.*, 2018 RefSeq: an update
449 on prokaryotic genome annotation and curation. Nucleic Acids Res 46**:** D851-D860.
450 Hattman, S., S. Schlagman, L. Goldstein and M. Frohlich, 1976 *Salmonella typhimurium* SA
451 host specificity system is based on deoxyribonucleic acid-adenine methylation. J
452 Bacteriol 127**:** 211-217.
453 Hoiseth, S. K., and B. A. Stocker, 1981 Aromatic-dependent S*almonella typhimurium* are non-
454 virulent and effective as live vaccines. Nature 291**:** 238-239.
455 Hunt, M., N. D. Silva, T. D. Otto, J. Parkhill, J. A. Keane *et al.*, 2015 Circlator: Automated
456 circularization of genome assemblies using long sequencing reads. Genome Biology 16**:**
457 2601.
458 Jones-Carson, J., A. Yahashiri, J.-S. Kim, L. Liu, L. F. Fitzsimmons *et al.*, 2020 Nitric oxide
459 disrupts bacterial cytokinesis by poisoning purine metabolism. Science Advances 6**:**
460 eaaz0260.
461 Joys, T. M., and B. A. Stocker, 1965 Complementation of non-flagellate *Salmonella* mutants. J
462 Gen Microbiol 41**:** 47-55.
463 Joys, T. M., and B. A. Stocker, 1969 Recombination in H1, the gene determining the flagellar
464 antigen-i of *Salmonella typhimurium*; mapping of H1 and fla mutations. J Gen Microbiol
465 58**:** 267-275.
466 Kingston, A. W., C. Ponkratz and E. A. Raleigh, 2017 Rpn (YhgA-Like) Proteins of *Escherichia*
467 *coli* K-12 and Their Contribution to RecA-Independent Horizontal Transfer. J Bacteriol
468 199**:** e00787-00716-00717.
469 Kingston, A. W., C. Roussel-Rossin, C. Dupont and E. A. Raleigh, 2015 Novel *recA*-
470 Independent Horizontal Gene Transfer in Escherichia coli K-12. PloS one 10**:** e0130813.

471 Koren, S., B. P. Walenz, K. Berlin, J. R. Miller, N. H. Bergman *et al.*, 2017 Canu: scalable and
472 accurate long-read assembly via adaptive k-mer weighting and repeat separation.
473 Genome Res 27**:** 722-736.
474 Lemire, S., N. Figueroa-Bossi and L. Bossi, 2007 Prophage contribution to *Salmonella* virulence
475 and diversity, pp. 159-191 in *Horizontal Gene Transfer in the Evolution of Bacterial*
476 *Pathogenesis*, edited by H. S. M. Hensel. Cambridge University Press, Cambridge,
477 England.
478 Lemire, S., N. Figueroa-Bossi and L. Bossi, 2008 A singular case of prophage complementation
479 in mutational activation of *recET* orthologs in *Salmonella enterica* serovar Typhimurium.
480 J Bacteriol 190**:** 6857-6866.
481 Lerche, M., H. Sandhu, L. Flöckner, M. Högbom and M. Rapp, 2017 Structure and Cooperativity
482 of the Cytosolic Domain of the CorA Mg2+ Channel from *Escherichia coli.* Structure 25**:**
483 1175-1186.e1174.
484 Li, H., 2018 Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34**:** 3094-
485 3100.
486 Martinsohn, J. T., M. Radman and M. A. Petit, 2008 The λ Red proteins promote efficient
487 recombination between diverged sequences: implications for bacteriophage genome
488 mosaicism. PLoS Genet 4**:** e1000065.
489 McClelland, M., K. E. Sanderson, J. Spieth, S. W. Clifton, P. Latreille *et al.*, 2001 Complete
490 genome sequence of *Salmonella enterica* serovar Typhimurium LT2. Nature 413**:** 852-
491 856.
492 Murphy, K. C., 2016 λ Recombination and Recombineering. EcoSal Plus 7.

493 Murray, I. A., R. D. Morgan, Y. Luyten, A. Fomenkov, I. R. Correa, Jr. *et al.*, 2018 The non-
494 specific adenine DNA methyltransferase M.EcoGII. Nucleic Acids Res 46**:** 840-848.
495 Ornellas, E. P., and B. A. D. Stocker, 1974 Relation of lipopolysaccharide character to P1
496 sensitivity in *Salmonella typhimurium*. Virology 60**:** 491-502.
497 Pirone-Davies, C., M. Hoffmann, R. J. Roberts, T. Muruvanda, R. E. Timme *et al.*, 2015
498 Genome-wide methylation patterns in *Salmonella enterica* Subsp. *enterica* Serovars.
499 PLoS One 10**:** e0123639.
500 Piya, D., L. Vara, W. K. Russell, R. Young and J. J. Gill, 2017 The multicomponent
501 antirestriction system of phage P1 is linked to capsid morphogenesis. Mol Microbiol 105**:**
502 399-412.
503 Roberts, G. A., A. S. Stephanou, N. Kanwar, A. Dawson, L. P. Cooper *et al.*, 2012 Exploring the
504 DNA mimicry of the Ocr protein of phage T7. Nucleic Acids Res 40**:** 8129-8143.
505 Roberts, R. J., T. Vincze, J. Posfai and D. Macelis, 2015 REBASE--a database for DNA
506 restriction and modification: enzymes, genes and genomes. Nucleic Acids Res 43**:** D298-
507 299.
508 Sanderson, K. E., 1987 F-mediated conjugation, F strains, and Hfr strains of *Salmonella*
509 *typhimurium* and *Salmonella abony*, pp. 2406-2412, edited by F. C. Neidhardt, R. Curtiss
510 III, J. L. Ingraham, E. C. Lin, K. B. Low *et al.* pdfs.semanticscholar.org, Washington,
511 DC.
512 Sanderson, K. E., and J. R. Roth, 1988 Linkage map of S*almonella typhimurium*, edition VII.
513 Microbiological reviews 52**:** 485-532.
514 Schmieger, H., 1999 Molecular survey of the *Salmonella* phage typing system of Anderson. J
515 Bacteriol 181**:** 1630-1635.

516 Segall, A. M., and J. R. Roth, 1989 Recombination between homologies in direct and inverse
517 orientation in the chromosome of *Salmonella*: intervals which are nonpermissive for
518 inversion formation. Genetics 122**:** 737-747.
519 Serfiotis-Mitsa, D., A. P. Herbert, G. A. Roberts, D. C. Soares, J. H. White *et al.*, 2010 The
520 structure of the KlcA and ArdB proteins reveals a novel fold and antirestriction activity
521 against Type I DNA restriction systems in vivo but not in vitro. Nucleic Acids Research
522 38**:** 1723-1737.
523 Sharples, G. J., F. A. Curtis, P. McGlynn and E. L. Bolt, 2004 Holliday junction binding and
524 resolution by the Rap structure-specific endonuclease of phage λ. J Mol Biol 340**:** 739-
525 751.
526 Simon, M., J. Zieg, M. Silverman, G. Mandel and R. Doolittle, 1980 Phase variation: evolution
527 of a controlling element. Science 209**:** 1370-1374.
528 Spicer, C. C., and N. Datta, 1959 Reversion of transduced antigenic characters in *Salmonella*
529 *typhimurium*. J Gen Microbiol 20**:** 136-143.
530 Stocker, B., B. M. Males and W. Takano, 1980 *Salmonella typhimurium* mutants of RfaH-
531 phenotype: genetics and antibiotic sensitivities. Microbiology 116**:** 17-24.
532 Tarkowski, T. A., D. Mooney, L. C. Thomason and F. W. Stahl, 2002 Gene products encoded in
533 the *ninR* region of phage lambda participate in Red-mediated recombination. Genes Cells
534 7**:** 351-363.
535 Tatusova, T., M. DiCuccio, A. Badretdin, V. Chetvernin, E. P. Nawrocki *et al.*, 2016 NCBI
536 prokaryotic genome annotation pipeline. Nucleic Acids Research 44**:** 6614-6624.
537 Timme, R. E., J. B. Pettengill, M. W. Allard, E. Strain, R. Barrangou *et al.*, 2013 Phylogenetic
538 diversity of the enteric pathogen *Salmonella enterica* subsp. *enterica* inferred from

539 genome-wide reference-free SNP characters. Genome biology and evolution 5**:** 2109-
540 2123.
541 Tsai, S. P., R. J. Hartin and J. Ryu, 1989 Transformation in restriction-deficient *Salmonella*
542 *typhimurium* LT2. J Gen Microbiol 135**:** 2561-2567.
543 Yoon, W., Y. Yoo, Y. S. Chae, S. H. Kee and B. M. Kim, 2018 Therapeutic advantage of
544 genetically engineered *Salmonella typhimurium* carrying short hairpin RNA against
545 inhibin alpha subunit in cancer treatment. Annals of Oncology 29**:** 2010-2017.
546 Zavilgelsky, G. B., and S. M. Rastorguev, 2009 Antirestriction proteins ArdA and Ocr as
547 efficient inhibitors of Type I restriction-modification enzymes. Molecular Biology 43**:**
548 241-248.
549 Zaworski, J., C. McClung, C. Ruse, P. R. Weigele, R. W. Hendrix *et al.*, 2021 Genome analysis
550 of *Salmonella enterica* serovar Typhimurium bacteriophage L, indicator for StySA
551 (StyLT2III) restriction-modification system action. G3 Genes|Genomes|Genetics 11.
552
553

554 **Table 1. DNA modification systems, motifs and degree of modification**

SYSTEM[A]	MOTIF[B]	MOD TYPE[C]	% M[D] STK005	% M[D] ER3625
M.SenLBI (StyLT, StyLTI)	5' CAG**A**G 3'	m6A	99.9	99.9
M.SenLBII (StySB)	5′ G**A**GNNNNNNRTAYG 3′	m6A	99.1	98.3
	3′ CTCNNNNNNY**A**TRC 5′		100	100
SenLBORF22590P (StySA)	5' GATC**A**G 3'	m6A	0.3	0
M.SenLBIV	5' ATGC**A**T 3'	m6A	24.9	38.2
M.SenLBDam	5' G**A**TC 3'	m6A	99.9	99.9
M.SenLBDcm	5' C**C**WGG 3'	(m5C)	(1.0)	(0.3)
SenLBORF23430P	ND			

555 [a]System: name used in REBASE for MTases of LB5000 (REBASE organism 45877) followed by (names most
556 encountered in the RM literature). Numeral order and modification motifs for ER3625 (REBASE organism 45876)
557 are the same, except M.SenER3625 rather than M.SenLB. These systems are descendants of the LT2 systems
558 (REBASE organism 18099). SenLBORF22590P and SenER3625ORF22595 are present but inactive. They correspond
559 to M.SenLT2III. Plasmid-borne SenLBORF23430P and Sen362523455 descends from M.SenLT2ORF505, a
560 homolog of nonspecific MTase M.EcoGIX (Fomenkov supplementary Table 7).

561 [b]Motif: recognition sequence with modified base in bold. Sites shown once are symmetric, or asymmetric sites
562 methylated only on one strand; sites shown twice are asymmetric sites methylated on both strands.

563 [c] Mod Type: m6A=N6-methyladenine; m5C=5 methylcytosine. Dcm (m5C) motif is inferred from the presence

564 of the *dcm* gene; m5C is not efficiently detected by this approach.

565 [d]%M: (# motif sites modified (IPD >2, 25X coverage)/ motif occurrences)x100.

566

567