JOURNAL OF BACTERIOLOGY, Jan. 2010, p. 560–567 Vol. 192, No. 2
0021-9193/10/$12.00 doi:10.1128/JB.01233-09

Short-Term Signatures of Evolutionary Change in the *Salmonella enterica* Serovar Typhimurium 14028 Genome▽†

Tyler Jarvik,[1] Chris Smillie,[1] Eduardo A. Groisman,[2] and Howard Ochman[1]*

Department of Chemistry and Biochemistry, University of Arizona, Tucson, Arizona,[1] and Howard Hughes Medical Institute, Department of Molecular Microbiology, Washington University School of Medicine, St. Louis, Missouri[2]

Received 11 September 2009/Accepted 1 November 2009

Salmonella enterica **serovar Typhimurium is a Gram-negative pathogen that causes gastroenteritis in humans and a typhoid-like disease in mice and is often used as a model for the disease promoted by the human-adapted *S. enterica* serovar Typhi. Despite its health importance, the only *S.* Typhimurium strain for which the complete genomic sequence has been determined is the avirulent LT2 strain, which is extensively used in genetic and physiologic studies. Here, we report the complete genomic sequence of the *S.* Typhimurium strain 14028s, as well as those of its progenitor and two additional derivatives. Comparison of these *S.* Typhimurium genomes revealed differences in the patterns of sequence evolution and the complete inventory of genetic alterations incurred in virulent and avirulent strains, as well as the sequence changes accumulated during laboratory passage of pathogenic organisms.**

The genomes of related bacteria can differ in three ways: (i) gene content, where one bacterial species or strain harbors genes absent from the other organism; (ii) nucleotide substitutions within largely conserved DNA sequences, which can result in amino acid changes in orthologous proteins, form pseudogenes, and promote distinct expression patterns of genes present in the two organisms; and (iii) changes in gene arrangement, caused by inversions and translocations. These differences have been observed not only across bacterial species but also among strains belonging to the same species. Recent genomic analyses have revealed that many bacterial pathogens of humans are virtually monomorphic (1) and exhibit very limited sequence diversity, raising questions about the nature of the genetic changes governing distinct behaviors. Furthermore, several bacterial pathogens that have been subjected to extensive passage in the laboratory display altered virulence characteristics, but the genetic basis for these alterations remains largely unknown. Here, we address both of these questions by determining and analyzing the genome sequences of closely related isolates of *Salmonella enterica* serovar Typhimurium, a Gram-negative pathogen that has been used as a preeminent model to investigate basic genetic mechanisms (2, 8, 46, 59), as well as the interaction between bacterial pathogens and mammalian hosts (11, 41).

The genus *Salmonella* is divided into two species: *Salmonella bongori* and *Salmonella enterica*, which together comprise over 2,300 serovars differing in host specificity and the disease conditions they promote in various hosts. For example, *S. enterica* serovar Typhi is human restricted and causes typhoid fever, whereas serovar Typhimurium is a broad-host-range organism that causes gastroenteritis in humans and a typhoid-like disease in mice. Although the complete genome sequences of 15 *Salmonella enterica* strains are available, there is only a single representative of *S.* Typhimurium—strain LT2 (31). Despite its wide application in genetic analysis, strain LT2 is highly attenuated for virulence in both *in vitro* and *in vivo* assays (52, 56), leading many investigators to use other *S.* Typhimurium isolates to examine the genetic basis for bacterial pathogenesis (11, 14, 16).

Over 300 virulence genes (3, 5, 47) have already been identified in *Salmonella enterica* serovar Typhimurium 14028 (now termed *S. enterica* subsp. *enterica* serovar Typhimurium ATCC 14028), which is a descendant of CDC 60-6516, a strain isolated in 1960 from pools of hearts and livers of 4-week-old chickens (P. Fields, personal communication). Whereas strain 14028 has been typed as LT2, a designation based on phage sensitivity (27), the two strains were isolated from distinct sources decades apart, which makes their genealogy and exact relationship obscure. A derivative of the original 14028 strain with a rough colony morphology (due to changes in O-antigen expression) was designated 14028r to distinguish it from the original smooth strain, renamed 14028s, and was used in a genetic screen for *Salmonella* virulence genes because it retained lethality for mice and the ability to survive within murine macrophages. Strain 14028 was also used for the identification of *Salmonella* genes that were specifically expressed during infection of a mammalian host (30). Both 14028 and LT2 possess a 90-kb virulence plasmid promoting intracellular replication and systemic disease (14), but they differ in their prophage contents, as is often the case among *S.* Typhimurium strains (12, 13).

To identify the individual changes that differentiate *S.* Typhimurium strains and to assess the nature of variation that arises during laboratory storage and passage, we determined the genome sequence of strain 14028s. This genome was then used as a reference for sequencing its progenitors, including the original source strain CDC 60-6516 and the earliest smooth and rough variants. Our analysis uncovered the genomic differences that arose during the past decades of laboratory cul-

* Corresponding author. Mailing address: Department of Chemistry and Biochemistry, University of Arizona, 233 Life Sciences South, 1007 E. Lowell St., Tucson, AZ 85718. Phone: (520) 626-8355. Fax: (520) 621-9288. E-mail: hochman@u.arizona.edu.

† Supplemental material for this article may be found at http://jb.asm.org.

▽ Published ahead of print on 6 November 2009.

tivation and showed that derivatives with different virulence potentials can follow distinct patterns of sequence evolution.

MATERIALS AND METHODS

Bacterial strains. Sequences were generated for the genomes of four strains of *S.* Typhimurium. First, a complete and annotated sequence was produced for the genome of a contemporary isolate of strain 14028s. Freezer stocks of this strain have been repeatedly passaged and subcultured and used in experimental settings for more than 2 decades. Using the completed 14028s sequence as a reference, we subsequently obtained the genomic sequences of three archival stocks of the following strains. (i) CDC 60-6516, a pathogenic strain, was initially isolated from samples of hearts and livers of 4-week-old chickens and served as the source from which the original 14028 strain was derived. This strain was stored in stab culture at room temperature from 1960 to 2003 and subsequently as a glycerol stock at −70°C from 2003 to 2009. (ii) ATCC 14028r is the original rough-colony strain described previously (11). The strain derives from a frozen glycerol stock stored on 18 February 1984. (iii) ATCC 14028s was obtained from a frozen glycerol stock stored on 21 February1985. To discriminate it from the contemporary reference strain, this strain is referred to here as 14028s-o. Note that although this strain was deposited later, it is not derived from ATCC 14028r, but rather represents the original lineage from which the rough strain evolved. The majority of *Salmonellae* form smooth colonies, but spontaneous mutations to rough colony morphology occur at a frequency of $>10^{-5}$. Comparisons of the allelic variants of these strains with those already typed by MLST assigned them to *Salmonella enterica* sequence type (ST) 19.

Genome sequencing and assembly. To derive the complete nucleotide sequence of strain 14028s, we employed a mixed strategy consisting of producing: (i) random reads that averaged 220 nucleotides (nt) and sampled each nucleotide, on average, 17 times (i.e., 17× coverage), obtained with a Roche 454-FLX pyrosequencer; (ii) 30× coverage of reads averaging 35 nt obtained with an Illumina-Solexa sequencer; (iii) 100× mate pair coverage by ABI SOLiD, in which 22-nt reads are derived from both ends of randomly sheared 500-bp fragments; and (iv) 2× mate pair coverage by ABI-Sanger sequencing, in which >900-nt reads are derived from both ends of clones produced from 2- to 3-kb cloned fragments.

Reads obtained using the 454 and Sanger methods were loaded together into Roche's GSassembler using default parameters, and all contigs corresponding to a single read were discarded. All remaining contigs were aligned with MUMmer and then compared to the published sequence of *S.* Typhimurium (strain LT2; accession no. NC003197) using BLASTN. Contigs aligning with more than one location on the reference genome were tagged as repeat elements. Mate pair data produced by ABI-SOLiD were used to predict the order and orientation of these 454/Sanger contigs, thereby confirming that no inversions or translocations were present relative to the LT2 genome. However, because most of the breaks in the assembly occurred in repeat regions, ABI-SOLiD reads were not informative for closing any gaps.

Genome finishing. Because GSassembler (and other sequence assemblers) can collapse repeated elements by underestimating the diversity or number of tandemly repeated units, we initially treated all repetitive sequences as gaps. Each of these gaps in the 14028s genome was filled by designing PCR primers to regions of unique sequence adjacent to the gap, amplifying across the unknown region, and sequencing the PCR products via conventional methods. Whenever the resulting PCR products exceeded the length of a single Sanger sequencing read, complete double-stranded sequences were obtained by primer walking. In cases where the gaps were too large to amplify from unique flanking sequences (e.g., rRNA operons), we designed forward and reverse primers based on repeated sequences believed to reside in the middle of the gap. These primers were used in conjunction with unique flanking primers to generate two PCR products that together spanned the gap. Only one repeated element, corresponding to an ~20-kb region common to two prophages, was too large to resolve by these methods, and in this case, the consensus sequence derived from GSassembler was included in the finished genome in both locations (1304985 to 1325378 and 2789731 to 2810125).

Error correction. Initial comparisons of the complete, gap-free 14028s genome to the published LT2 genome revealed a number of 1-bp indels that resulted in frameshifts in annotated genes. Many of these indels were situated in homopolymeric sequences, which are known to be prone to errors in 454 pyrosequencing technology. To verify indels in stretches of homopolymers, we manually examined the Sanger reads spanning each homopolymeric region that contained an indel and corrected the sequence to match the Sanger read if it did not match the consensus sequence. Although data obtained by Illumina/Solexa were not integrated in our genome assembly, these reads were also used to confirm or correct homopolymeric indels. For 15 homopolymeric indels for which Sanger or Solexa reads were either absent or inconsistent, the DNA sequences were checked by PCR amplification and sequencing.

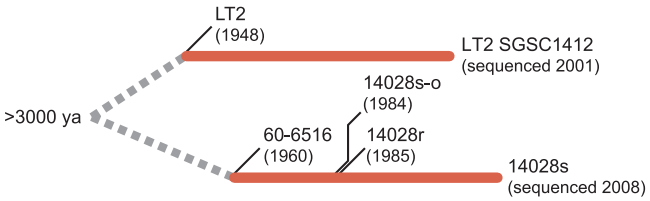

FIG. 1. Genealogy of *S.* Typhimurium strains (see Materials and Methods for source and storage conditions). Wherever possible, isolation dates are noted, and the divergence time between the LT2 and the 14028 lineages is estimated based on sequence data generated in the current study. ya, years ago.

Annotation. Wherever possible, the 14028s genome annotation followed that of the *S.* Typhimurium LT2 genome. For sequences not present in that genome, candidate genes were identified in Prodigal (http://compbio.ornl.gov/prodigal) and searched with BLASTP against the nonredundant protein sequence database. Because numerous other *Salmonella* genome sequences have become available since LT2 was first published (31), BLASTX searches of all remaining intergenic regions were also performed against the nonredundant database. All potential coding regions ≥30 amino acids in length and annotated in any other *Salmonella* genome were added to the 14028s annotation. Due to the different start codons listed for a number of genes, we also updated our annotation to reflect the longer gene product whenever the reading frame was conserved.

Genome resequencing. After completing the 14028s genome, we obtained ~400-nt reads via Roche 454-Titanium pyrosequencing of strains 14028s-o, 14028r, and 60-6516 to coverage depths of 21×, 20×, and 32×, respectively. Each genome was assembled with GSassembler using default parameters, and the resulting contigs were aligned with the finished 14028s genome using BLASTN. Apparent differences between the resequenced genomes and the 14028s genome were filtered to exclude the unreliable set of 1-bp indels in homopolymers of >4 bp, as well as those 1-bp indels and base substitutions marked as low confidence by GSassembler. All remaining base substitutions and indels, including all homopolymeric indels observed in more than one genome, were checked by Sanger sequencing of a PCR corresponding to the region.

Nucleotide sequence accession numbers. The complete *S.* Typhimurium 14028s genome has been deposited in GenBank under accession numbers CP001363 for the bacterial chromosome and CP001362 for the plasmid.

RESULTS

Sequencing strategies. Using four sequencing platforms (ABI-Sanger, Roche-454, Illumina/Solexa, and ABI-SOLiD), we assembled the complete sequence of a contemporary isolate of strain 14028s, which was, in turn, used as a reference for sequencing the genomes of three progenitor strains (Fig. 1). Due to the very short read lengths and/or high error frequencies, data generated by ABI-SOLiD and Illumina/Solexa were not included in the initial genome assembly of strain 14028s. The 454 run yielded a total of 372,340 reads (82,619,725 bases; 17× coverage), and the Sanger runs yielded 10,752 reads (9,958,592 bases; 2× coverage). The combined assembly based on these two data sets formed 259 contigs, 106 of which corresponded to individual reads and were removed from subsequent analyses. The remaining 153 contigs served as the starting point for genome assembly and were aligned with the published LT2 genome. Several of the 14028s contigs represented repeat elements not present in the published LT2 genome, and alternatively, several regions of the published genome had no counterpart in 14028s.

Divergence between the LT2 and 14028s genomes. (i) Indels. After all gaps in the 14028s genome were closed, comparison

with the LT2 genome uncovered 55 single-base indels leading to frameshifts in annotated genes. Although individual *Salmonella* strains are each expected to contain some unique pseudogenes, 454 sequencing technologies are prone to errors in long hompolymeric runs, thereby generating indels. By reexamining each of these sites using sequencing technologies other than 454, we found that 25 of the 55 single-base indels were attributable to sequencing errors. Not all such errors occurred in long homopolymers: six were in runs of three or fewer mononucleotides, and two occurred within runs of nonidentical nucleotides. Additionally, data from the three resequenced genomes revealed 25 candidate errors (primarily intergenic homopolymers) in our original 14028s assembly that were each checked by PCR and Sanger sequencing.

The 14028s and LT2 genomes are over 98% identical in sequence, with the greatest difference between the two strains attributable to the distribution of four prophages (Fig. 2). Two intact LT2 prophages (Fels-1 and Fels-2) are not present in 14028s; however, there are remnants of several Fels-2 genes at the corresponding positions in the 14028s genome, indicating that the prophage is ancestral to both strains but was subsequently eliminated from 14028s. In contrast, strain 14028s contains two prophages not present in LT2: one, integrated within $tRNA^{Ser}$, is 40,146 bp in length and >99% identical to the previously characterized *S. enterica* phage ST64B, and the other, inserted near the 3′ end of the *icdA* gene, spans 51,101 bp. Most of this 51-kb prophage is virtually identical to a prophage in *S. enterica* serovar Newport; however, the remaining portions, constituting about 30% of its length, display significant identity to prophages in other bacteria only at the amino acid level. This phage, referred to as Gifsy-3, contains two known virulence genes, *sspH1* and *pagJ* (12).

In addition to the prophage insertions and deletions, numerous other classes of insertion/deletion events differentiate the 14028s and LT2 genomes (Table 1). Even though four IS*200* elements in 14028s and LT2 reside in identical locations in the two strains, there are eight others (six in 14028s and two in LT2) that are confined to only one of the strains. One of the IS*200* insertions unique to 14028s disrupts STM1228, encoding a putative periplasmic protein, whereas the other five occur within intergenic regions. In addition to changes in genome contents attributable to phage and transposable elements, there were 11 other indels >100 bp in length, three of which occurred within protein-coding regions, including a 5.1-kb deletion that removed at least four genes (STM3256 to STM3259) from the 14028s genome. Aside from genes encoded within prophages, there are no other gene-size insertions into 14028s or large deletions in LT2.

When the 14028s and LT2 genomes were compared, there were a total of 142 indels under 100 bp in length, mostly due to single-base insertions or deletions (see below). Although a few of these single-base indels could have arisen from sequencing errors (only those indels that occurred within coding regions were initially confirmed by PCR and Sanger sequencing), the vast majority were subsequently verified by ABI-Solid and/or Solexa/Illumina reads. Considering only unambiguous indels that were verified by PCR and/or alternate sequencing technologies, there were approximately equal numbers of insertions and deletions in the two strains.

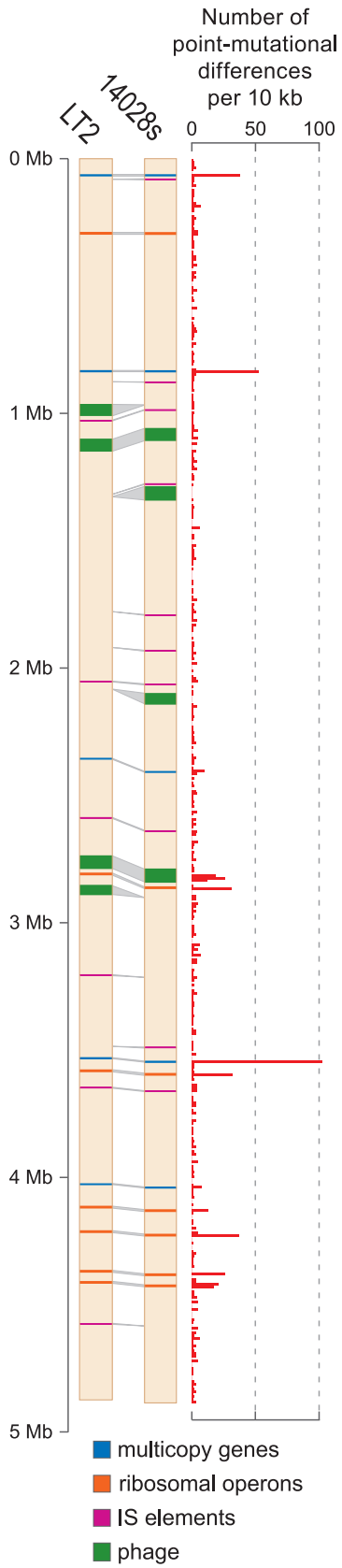

FIG. 2. Alignment of the *S.* Typhimurium 14028s and LT2 genomes. Differences in genome organization and gene contents and the distribution of substitutions (point mutations and 1-bp indels) are shown.

TABLE 1. Events causing strain-specific changes in DNA content

Event	No. in:	
	14028s	LT2
Genome size	4,870,265	4,857,432
1-bp events	66	37
Intergenic regions	38	19
Frameshifts in ORFs	26	15
Structural RNAs	2	3
2- to 99-bp events	17	22
Intergenic regions	6	12
Frameshifts in ORFs	1	4
In-frame indels	9	1
Structural RNAs	1	5
>100-bp events	12	12
Total strain-specific DNA	103,192	90,359

(ii) Base substitutions. We detected a total of 962 base substitutions between LT2 and 14028s. Due to our postassembly finishing of all repeated elements in the 14028s genome, base substitutions (and small indels) were elevated by nearly an order of magnitude in multicopy regions (red bars in Fig. 2). In all likelihood, the sequences of these regions in the LT2 genome reflect overcollapsed contigs in which merged sequences from nonidentical repeats were assembled in duplicated regions of the published genome. Due to the potential that these multicopy regions of the published LT2 genome (which include insertion sequence [IS] elements, rRNA operons, and prophages) are subject to sequencing artifacts, we excluded all such regions from the subsequent analyses, leaving a total of 540 base substitutions separating LT2 and 14028s (Table 2).

Overall, substitutions in intergenic regions and synonymous sites occur at about the same frequency (1.58×10^{-4} per intergenic site versus 1.61×10^{-4} per synonymous site), as expected if these sites are neutral or nearly so or if both types contain the same proportion of neutral sites. The K_a/K_s ratio (i.e., the ratio of nonsynonymous to synonymous site substitutions) based on substitution differences within annotated genes is 0.531, which is over 10 times that observed in comparisons of homologous genes from *Escherichia coli* and *S. enterica* ($K_a/K_s = 0.036$). The inflated K_a/K_s ratio observed between the two closely related *Salmonella* strains is due to the shorter duration of purifying selection acting on slightly deleterious mutations.

The availability of the genome sequences of several outside reference strains (32, 40, 55) allowed us to infer the ancestral states of most of the observed base substitutions and to establish in which of the two strains each substitution occurred. In all, 272 base substitutions (and 50 indels) could be unequivocally assigned to the 14028s lineage versus 244 base substitutions (and 42 indels) assigned to the LT2 lineage (Fig. 3; absolute numbers of each class of substitution are given in Table S1 in the supplemental material). For three classes of substitutions, there are large differences between the two strains. (i) There are more TA-to-CG substitutions in intergenic regions of 14028s. (ii) The complementary substitution, CG-to-TA, is prevalent at nonsynonymous sites of 14028s (72 versus 48) and was the most common type of substitution detected. (iii) In contrast to the asymmetry between the strains in the frequency of TA/CG transitions, CG-to-GC transversions in LT2 significantly ($P = 0.001$) outnumbered those in 14028s. The ability to recognize the direction of each substitution also allowed us to identify specific trends in base-compositional bias. For those sites that could be assigned, there was a decrease in overall G+C content in both strains, which was most pronounced at synonymous sites.

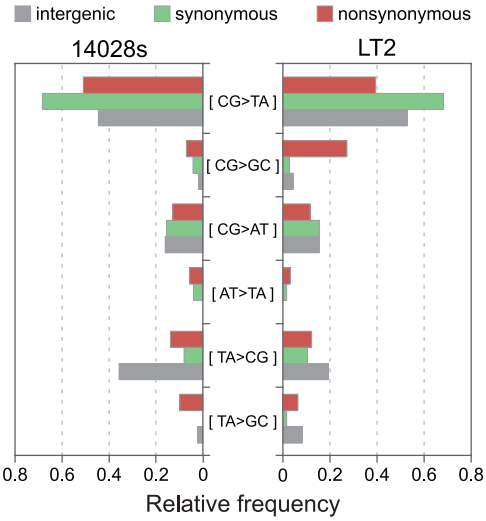

FIG. 3. Relative frequencies of each type of base substitution detected at synonymous and nonsynonymous sites and in intergenic regions. The frequencies are adjusted for the base composition of the particular class of sites.

TABLE 2. Mutations in single-copy regions of the *S.* Typhimurium genomes

Class of site	Total no. of sites	No. of SNPs[a]	SNP rate[b]	No. of indels[a]	Indel rate[b]	SNP/indel ratio
Total	4,641,529	540	1.16	112	0.24	4.82
Intergenic DNA	606,052	107	1.77	66	1.09	1.62
Coding DNA	4,035,477	433	1.07	46	0.11	9.41
Synonymous sites	976,178	161	1.65			
Nonsynonymous sites	3,059,299	272	0.89			

[a] Totals do not include 446 single-nucleotide polymorphisms (SNPs) and 51 indels in repeated elements and multicopy genes.
[b] Number per 10 kb since the divergence of strains 14028s and LT2.

(iii) Gene-inactivating mutations and pseudogenes. When all single-base substitutions and small indels were considered (while those few coding genes that occurred in multiple copies were ignored), 13 of the 14028s genes contained potentially inactivating mutations that resulted in considerably shortened coding regions, which, when combined with those genes disrupted by large insertions or deletions, yielded a total of 19 pseudogenes in 14028s that were intact in LT2 (Table 3). In addition to fimbrial protein (*lpfD*) and virulence factor (*ratB*) genes, which are at least 20% shorter than their LT2 counterparts, there are several other slightly truncated genes in 14028s, including fumarase (*fumA*), ribonulease Z (*elaC*), and secreted effector protein (*avrA*) genes that have been shown to be distributed variably among *Salmonella* strains (see Table S2

TABLE 3. Strain-specific pseudogenes

Strain	Mutation type	Locus tag	Gene	Function	Length (%)[a]
14028s	Frameshift	STM14_5527		Putative phospholipase D	60.1
		STM14_0252		Putative inner membrane protein	78.7
		STM14_2018		Putative regulatory protein	72.0
		STM14_2083	*yciO*	Hypothetical protein	76.3
		STM14_2307		Putative cytoplasmic protein	72.6
		STM14_2636		Putative cytoplasmic protein	52.1
		STM14_3082	*ratB*	Putative outer membrane protein	79.9
		STM14_4384	*lpfD*	Long polar fimbrial protein	48.6
		STM14_3788		Putative amidohydrolase	52.6
		STM14_3185		DNA packaging-like protein	75.8
		STM14_0189	*yacH*	Putative outer membrane protein	52.7
	Stop codon	STM14_3323		Putative phosphotransferase system IIC component	30.4
		STM14_5043		Putative Na^+-dependent transporter	74.6
	Large internal deletion:	STM14_0711		Molybdopterin-containing oxidoreductase iron-sulfur subunit	64.3
		STM14_3760	*nupG*	Nucleoside transport	30.9
	5′ truncation	STM14_3221		Hypothetical protein	60.9
		STM14_3942		PTS family galactitol-specific enzyme IIC	8.1
	3′ truncation	STM14_3941		PTS system fructose-specific EIIBC component	47.9
	Phage insertion	STM14_2427	*serU*	$tRNA^{Ser}$	
LT2	Frameshift	STM0342		Putative periplasmic protein	59.9
		STM1048/STM1048.1N		Host specificity protein J	73/26
		STM3191/STM3192		Putative arylsulfate sulfotransferase	54/44
	Stop codon	STM3980/STM3981	*assT*	Arylsulfotransferase	39/58

[a] ORF lengths are given as proportions of their full-length counterpart in the other strain. Entries for which two locus tags and two ORF lengths are listed represent cases in which a single gene was annotated as two separate genes in LT2.

in the supplemental material). Moreover, as discussed above, the *serU* tRNA gene contains a phage insertion in 14028s.

Of the 39 pseudogenes that were originally annotated in LT2, 38 are present in 14028s (and the other is located in a prophage absent from 14028s), and in no case is it evident that any of the shared pseudogenes were inactivated independently in the two strains. In addition, there are four genes in LT2 that were originally considered to be functional but whose homologs, due to a base substitution or frameshift, were found to encode longer proteins in 14028s. In three of these cases (host specificity protein J and two putative arylsulfate sulfotransferases), the open reading frame (ORF) in 14028s encompassed two adjacent genes in the LT2 genome, representing examples of recent LT2 pseudogenes (or potential sequencing errors) that are full length in other sequenced *Salmonella* genomes.

In addition to those genes truncated by the presence of frameshifting or nonsense mutations, there are numerous genes that have incurred nonsynonymous substitutions that could potentially affect protein function, as well as numerous frameshifts that have altered the sequence at one end of the ORF, even if they do not substantially shorten it (see Table S2 in the supplemental material). With respect to mutations that might be responsible for the difference in virulence potential of 14028s and LT2, there are amino acid substitutions in BigA, a surface-exposed virulence protein, in the putative virulence factors MviM and SrfB, as well as in several regulatory proteins, most notably RpoS and SlyA, which are both known to modulate *Salmonella* virulence (9, 26). For example, the start codon for the *rpoS* gene is ATG in 14028s but TTG in LT2, which results in lower RpoS levels in the latter strain. In the case of the SlyA protein, 14028s and LT2 differ at two positions, one at which Asp is changed conservatively to Glu and one where the LT2 Ala98 located within a predicted alpha helix adjacent to the DNA binding domain is replaced by a helix breaker Pro in 14028s (38).

Resequencing 14028s progenitors. We applied 454-Titanium technology to recover the genomic sequences of three strains that form the lineage leading to our contemporary isolate of 14028s. We obtained coverage depths of 32×, 20×, and 21× for strains 60-6516, 14028r, and 14028s-o, respectively, and in each case, automated assembly yielded approximately the same number of contigs (~200) as in the original 14028s assembly. These alignments revealed which contigs in the assemblies of the three resequenced strains matched multiple locations on the genome, reflecting the presence of repetitive elements. The repeated regions included prophages, IS elements, rRNA operons, and several other multicopy genes (notably the *oadAB*, *dcoAB*, and *ccmGH* operons), which together constitute ~3% of each genome (Fig. 2). Because the sequences of these potentially overcollapsed repeats were not individually resolved, as was done for the 14028s genome sequence, they were excluded from subsequent analyses.

Two structural changes in genome architecture were evident in the nonrepetitive regions of the three resequenced genomes. In the 14028r genome, the *hin* invertase region, responsible for flagellar phase variation, is in an orientation opposite to that in the other genomes. Strain 60-6516, the putative ancestor of all 14028 strains, lacks a 2,351-bp region including four genes: *dcuA* (an anaerobic C_4-dicarboxylate transporter), *aspA* (aspartate ammonia lyase), *fxsA* (F exclusion of bacteriophage T7), and STM14_5203, encoding a hypothetical protein of unknown function. The presence of this region in the three 14028 strains, and in all other *S. enterica* strains sequenced to date, points to a deletion that occurred in culture or during recent cultivation of strain 60-6516. Aside from these changes, we did not detect any cases of gene amplification or translocation, as

observed in some archival strains of LT2 that were maintained in stab culture (42).

After all ambiguous polymorphisms (i.e., low-confidence substitutions and indels and those proved false by PCR and Sanger sequencing) were removed and each position was compared to its inferred ancestral state based on the LT2 sequence, there were seven changes in 60-6516, two changes in 14028r, one change in 14028s-o, and three changes in 14028s. The last represent the changes that have accumulated during laboratory passage and consist of three nonsynonymous substitutions, one in *nuoL*, one in the transcriptional regulator gene *cytR*, and one in locus STM14_1964 (encoding a putative cytoplasmic protein). Of the mutations found in the three progenitor strains, three occur in genes of known function: there is a frameshift *rpoS* mutation, coding for the starvation-induced sigma factor (10, 17) that differentiates 60-6516 from the other strains; there is a frameshift mutation in the *rfbJ* gene of 14028r responsible for the rough colony phenotype because the *rfbJ* gene codes for CDP-abequose synthase (22); and there is a 12-bp deletion in the *rbsR* gene of 14028s-o, coding for the repressor of the ribose regulon (24).

DISCUSSION

By determining the complete nucleotide sequence of *S.* Typhimurium 14028s, we were able to (i) identify the specific alterations that are responsible for the phenotype traits, particularly the virulence attributes, specific to this strain; (ii) search for genomic signatures of strain domestication; and (iii) distinguish lineage-specific changes in the rates and patterns of mutations. We should first note that the initial comparison of our sequence to the published LT2 genome revealed that base substitutions and indels were much more frequent in regions representing repeated DNA, including rRNA operons, multicopy elements, and duplicated genes. The type, magnitude, and distribution of these changes suggest that amalgamation and overcollapsing of sequences in the repeated regions of the LT2 genome are responsible for the prevalence of substitution in repeated sequences. The phenomenon of contig overcollapse is fairly common in genomic sequences reported for eukaryotes with high densities of repeated elements. Because it affected the assembly of ~3% of the LT2 genome, we were forced to exclude all polymorphisms detected in repeated regions. As a result, we removed nearly half of the base substitutions and indels from our data set in order to ascertain unambiguously the differences that exist between the compared *Salmonella* strains.

When the base substitutions that separate strains LT2 and 14028s were considered, CG-to-TA transitions (likely originating from deamination events) appeared at the highest frequency and constituted over 50% of the point-mutational differences between the two strains. They were followed by the complementary TA-to-CG mutation, which occurred at less than one-third of the frequency. Several studies have experimentally assessed the spectrum of mutations in enteric bacteria (6, 15, 28, 29, 49, 51). However, the relative frequencies and rank order of base substitutions in these analyses (even in those not employing mutator strains) do not mirror those detected in our genome comparisons (see Fig. S1 in the supplemental material), indicating that the rates and patterns of mutations can vary greatly under different natural and laboratory conditions (35, 45). This difference was also observed in the resequenced, laboratory-stored strains of 14028s: those few base substitutions that we detected were almost evenly split between CG-to-TA and TA-to-AT mutations, the latter of which was the rarest mutation observed between 14028s and LT2.

The disparities between the spectra of mutations observed under natural and laboratory-based conditions might not only be caused by differences in the occurrence of each class of mutation but could also result from selective processes. Nonsynonymous sites constitute the largest target in bacterial genomes, and certain mutations at these sites could potentially be deleterious and subsequently be removed from the genome. The K_a/K_s values (i.e., the ratio of the rate of nonsynonymous substitutions to the rate of synonymous substitutions) for the 14028s-LT2 comparisons are less than 1.0, indicating that some proportion of mutations at nonsynonymous sites have been removed. Overall, the per-site substitution rate at nonsynonymous sites was about half that at synonymous and intergenic sites (4.3×10^{-5} versus 8.0×10^{-5}), and in particular, the number of CG-to-GC mutations was significantly higher ($P < 0.0005$) at nonsynonymous sites. Naturally, most indels within genes cause frameshifts and are highly deleterious, but the rate of single-base indels to single-base substitutions in noncoding spacer regions was 1:3, identical to that reported for the aphid endosymbiont *Buchnera aphidicola* (34).

Several methods can be used to estimate the date of divergence between 14028s and LT2. Applying the experimentally observed mutation rate of 0.9×10^{-10} per base per replication calculated for *Salmonella* (21) and an estimate of 100 to 200 generations per year in the wild yields a divergence date of 3,000 to 6,000 years ago. Alternatively, given a substitution frequency of 8.0×10^{-5} at synonymous sites and the previously calibrated rate of 0.9% per million years based on the estimated *E. coli-Salmonella* divergence (36, 37), the split between 14028s and LT2 is estimated to have occurred approximately 9,000 years ago. Finally, our resequencing data, which recognized nine substitutions in the 50 years separating the 14028s and 60-6516 lineages, suggest that the divergence between 14028 and LT2 occurred about 3,000 years ago. It is common for divergence times based on laboratory-derived mutation rates and on sequence comparisons to differ by an order of magnitude (35); however, it appears that over the short term the estimates are fairly similar.

Since its divergence from the attenuated LT2 strain, the highly virulent 14028s accumulated 10% more base substitutions, primarily at nonsynonymous sites. Examination of the genomic locations of these nonsynonymous substitutions indicates that they occur in genes with varied functions (as opposed to affecting a particular class of genes, such as those involved in host interactions). This genome-wide distribution suggests that these substitutions have been fixed through a nonadaptive process, i.e., genetic drift, which occurs in lineages that experience population bottlenecks (4). Since 14028s is pathogenic and LT2 is not and the 14028s population sizes are more apt to fluctuate, it is not surprising that, like other host-associated bacteria, the 14028s lineage has accumulated a high number of mildly deleterious mutations, as reflected in an increased number of nonsynonymous substitutions (23).

Comparisons of the genome contents of multiple *Salmonella* serovars have shown that that several virulence genes with sporadic distributions are phage associated and that host-restricted serovars experience significant gene loss (32, 43, 44, 53). Although strains 14028s and LT2 are more closely related than any pair of *Salmonella* strains examined in these studies, there are differences in their genome contents associated with the insertion and deletion of prophages. As previously reported (12), LT2, but not 14028s, harbors the Fels-1 and Fels-2 phages, whereas 14028s uniquely carries Gifsy-3 and an ST64B-like phage. In addition, we have detected several phage-related genes that are restricted to only one of the strains (Fig. 2). Although genes encoded on Gifsy-2 are essential for virulence in mice, this is not the case for Gifsy-3, suggesting that the virulence effects of the Gifsy-3 genes (*sspH1* and *pagJ*) are limited to the enteric phase of the disease (33). Because the genome contents of strains LT2 and 14028s are very similar, differences in the functional status of genes and/or activities of the encoded gene products shared by the strains are most likely the source of the observed difference in virulence properties. Few, if any, of the pseudogenes unique to LT2 are likely candidates to explain its reduced virulence, suggesting that polymorphisms in any of several genes with known roles in virulence, such as *bigA*, *shdA*, *avrA*, *sseI*, *mviM*, *srfB*, *sipA*, *slyA*, or *rpoS*, may be responsible for their distinct pathogenic properties.

Numerous studies have examined the accumulation of substitutions in laboratory-propagated strains of bacteria, in which organisms are typically kept under continuous-growth conditions (in either serial culture or chemostats) and the mutations conferring a selective advantage are identified (18, 25, 57). In such situations, not only do mutations and genomic rearrangements accumulate, but salient characteristics that are not under selection are also often lost over time. For example, it is common for cultivated bacteria to become less virulent (20, 50), lose flagella (48), change colony morphology (7), develop auxotrophies, and eliminate extrachromosomal elements. Distinguishing all of the genetic differences among evolved or closely related bacterial strains has recently become a relatively straightforward process (18, 19, 39, 54, 58). Through comparative genomic analyses, we have shown that strains obtained from natural sources and subsequently used for experimental studies display characteristic rates and patterns of mutations, and we have identified the potential source of the phenotypic differences among strains.

ACKNOWLEDGMENTS

This research was supported in part by grants from the National Institutes of Health (GM56120 and GM74738 to H.O. and AI49561 and AI42236 to E.A.G.). E.A.G. is an investigator of the Howard Hughes Medical Institute.

We thank Becky Nankivell for preparing the figures and Leigh Riley at NCBI for her queries and assistance in submitting the sequence and in assigning accession numbers.

REFERENCES

1. **Achtman, M.** 2008. Evolution, population structure, and phylogeography of genetically monomorphic bacterial pathogens. Annu. Rev. Microbiol. **62:**53–70.
2. **Ames, B. N., B. Garry, and L. A. Herzenberg.** 1960. The genetic control of the enzymes of histidine biosynthesis in *Salmonella typhimurium*. J. Gen. Microbiol. **22:**369–378.
3. **Bowe, F., C. J. Lipps, R. M. Tsolis, E. Groisman, F. Heffron, and J. G. Kusters.** 1998. At least four percent of the *Salmonella typhimurium* genome is required for fatal infection of mice. Infect. Immun. **66:**3372–3377.
4. **Charlesworth, B.** 2009. Fundamental concepts in genetics: effective population size and patterns of molecular evolution and variation. Nat. Rev. Genet. **10:**195–205.
5. **Chaudhuri, R. R., S. E. Peters, S. J. Pleasance, H. Northen, C. Willers, G. K. Paterson, D. B. Cone, A. G. Allen, P. J. Owen, G. Shalom, D. J. Stekel, I. G. Charles, and D. J. Maskell.** 2009. Comprehensive identification of *Salmonella enterica* serovar Typhimurium genes required for infection of BALB/c mice. PLoS Pathog. **5:**e1000529.
6. **Cupples, C. G., and J. H. Miller.** 1989. A set of *lacZ* mutations in *Escherichia coli* that allow rapid detection of each of the six base substitutions. Proc. Natl. Acad. Sci. U. S. A. **86:**5345–5349.
7. **Davidson, C. J., A. P. White, and M. G. Surette.** 2008. Evolutionary loss of the *rdar* morphotype in *Salmonella* as a result of high mutation rates during laboratory passage. ISME J. **2:**293–307.
8. **Demerec, M., I. Blomstrand, and Z. E. Demerec.** 1955. Evidence of complex loci in *Salmonella*. Proc. Natl. Acad. Sci. U. S. A. **41:**359–364.
9. **Fang, F. C., S. J. Libby, N. A. Buchmaier, P. C. Loewen, J. Switala, J. Harwood, and D. G. Harwood.** 1992. The alternative sigma factor *katF* (*rpoS*) regulates *Salmonella* virulence. Proc. Natl. Acad. Sci. U. S. A. **89:**11978–11982.
10. **Ferenci, T.** 2008. The spread of a beneficial mutation in experimental bacterial populations: the influence of the environment and genotype on the fixation of *rpoS* mutations. Heredity **100:**446–452.
11. **Fields, P. I., R. V. Swanson, C. G. Haidaris, and F. Heffron.** 1986. Mutants of *Salmonella typhimurium* that cannot survive within the macrophage are avirulent. Proc. Natl. Acad. Sci. U. S. A. **83:**5189–5193.
12. **Figueroa-Bossi, N., S. Uzzau, D. Maloriol, and L. Bossi.** 2001. Variable assortment of prophages provides a transferable repertoire of pathogenic determinants in *Salmonella*. Mol. Microbiol. **39:**260–271.
13. **Figueroa-Bossi, N., and L. Bossi.** 2004. Resuscitation of a defective prophage in *Salmonella* cocultures. J. Bacteriol. **186:**4038–4041.
14. **Gulig, P. A., and R. Curtiss III.** 1987. Plasmid-associated virulence of *Salmonella typhimurium*. Infect. Immun. **55:**2891–2901.
15. **Hall, B. G.** 1991. Spectrum of mutations that occur under selective and non-selective conditions in *E. coli*. Genetica **84:**73–76.
16. **Harrington, K. A., and C. E. Hormaeche.** 1986. Expression of the innate resistance gene Ity in mouse Kupffer cells infected with *Salmonella typhimurium in vitro*. Microbial Pathog. **1:**269–274.
17. **Hengge, R.** 2008. The two-component network and the general stress sigma factor RpoS (sigma S) in *Escherichia coli*. Adv. Exp. Med. Biol. **631:**40–53.
18. **Herring, C. D., A. Raghunathan, C. Honisc, T. Patel, M. K. Applebee, A. R. Joyce, T. J. Albert, F. R. Blattner, D. van den Boom, C. R. Cantor, and B. Ø. Palsson.** 2006. Comparative genome sequencing of *Escherichia coli* allows observation of bacterial evolution on a laboratory timescale. Nat. Genet. **38:**1406–1412.
19. **Holt, K. E., J. Parkhill, C. J. Mazzoni, P. Roumagnac, F. X. Weill, I. Goodhead, R. Rance, S. Baker, D. J. Maskell, J. Wain, C. Dolecek, M. Achtman, and G. Dougan.** 2008. High-throughput sequencing provides insights into genome variation and evolution in *Salmonella* Typhi. Nat. Genet. **40:**987–993.
20. **Hopkins, R. J., J. G. Morris, Jr., J. C. Papadimitriou, C. Drachenberg, D. T. Smoot, S. P. James, and P. Panigrahi.** 1996. Loss of *Helicobacter pylori* hemagglutination with serial laboratory passage and correlation of hemagglutination with gastric epithelial cell adherence. Pathobiology **64:**247–254.
21. **Hudson, R. E., U. Bergthorsson, J. R. Roth, and H. Ochman.** 2002. Effect of chromosome location on bacterial mutation rates. Mol. Biol. Evol. **19:**85–92.
22. **Jiang, X. M., B. Neal, F. Santiago, S. J. Lee, L. K. Romana, and P. R. Reeves.** 1991. Structure and sequence of the *rfb* (O antigen) gene cluster of *Salmonella* serovar *typhimurium* (strain LT2). Mol. Microbiol. **5:**695–713.
23. **Kuo, C.-H., N. A. Moran, and H. Ochman.** 2009. The consequences of genetic drift for bacterial genome complexity. Genome Res. **19:**1450–1454.
24. **Laikova, O. N., A. A. Mironov, and M. S. Gelfand.** 2001. Computational analysis of the transcriptional regulation of pentose utilization systems in the gamma subdivision of Proteobacteria. FEMS Microbiol. Lett. **205:**315–322.
25. **Lenski, R. E., C. L. Winkworth, and M. A. Riley.** 2003. Rates of DNA sequence evolution in experimental populations of *Escherichia coli* during 20,000 generations. J. Mol. Evol. **56:**498–508.
26. **Libby, S. J., W. Goebel, A. Ludwig, N. Buchmeier, F. Bowe, F. C. Fang, D. G. Guiney, J. G. Songer, and F. Heffron.** 1994. A cytolysin encoded by *Salmonella* is required for survival within macrophages. Proc. Natl. Acad. Sci. U. S. A. **91:**489–493.
27. **Lilleengen, K.** 1948. Typing *Salmonella* by means of bacteriophage. Acta Pathol. Microbiol. Scand. Suppl. **77:**11–125.
28. **Lind, P. A., and D. I. Andersson.** 2008. Whole-genome mutational biases in bacteria. Proc. Natl. Acad. Sci. U. S. A. **105:**17878–17883.
29. **MacKay, W. J., S. Han, and L. D. Samson.** 1994. DNA alkylation repair limits spontaneous base substitution mutations in *Escherichia coli*. J. Bacteriol. **176:**3224–3230.
30. **Mahan, M. J., J. M. Slauch, and J. J. Mekalanos.** 1993. Selection of bacterial

virulence genes that are specifically induced in host tissues. Science **259:**686–688.
31. **McClelland, M., K. E. Sanderson, J. Spieth, S. W. Clifton, P. Latreille, L. Courtney. S. Porwollik, J. Ali, M. Dante, F. Du, et al.** 2001. Complete genome sequence of *Salmonella enterica* serovar Typhimurium LT2. Nature **413:**852–856.
32. **McClelland, M., K. E. Sanderson, S. W. Clifton, P. Latreille, S. Porwollik, A. Sabo, R. Meyer, T. Bieri, P. Ozersky, M. McLellan, et al.** 2004. Comparison of genome degradation in Paratyphi A and Typhi, human-restricted serovars of *Salmonella enterica* that cause typhoid. Nat. Genet. **36:**1268–1274.
33. **Miao, E. A., and S. I. Miller.** 1999. Bacteriophages in the evolution of pathogen-host interactions. Proc. Natl. Acad. Sci. U. S. A. **96:**9452–9454.
34. **Moran, N. A., H. J. McLaughlin, and R. Sorek.** 2009. The dynamics and time scale of ongoing genomic erosion in symbiotic bacteria. Science **323:**379–382.
35. **Ochman, H.** 2003. Neutral mutations and neutral substitutions in bacterial genomes. Mol. Biol. Evol. **20:**2091–2096.
36. **Ochman, H., and A. C. Wilson.** 1987. Evolution in bacteria: evidence for a universal substitution rate in cellular genomes. J. Mol. Evol. **26:**74–86.
37. **Ochman, H., S. Elwyn, and N. A. Moran.** 1999. Calibrating bacterial evolution. Proc. Natl. Acad. Sci. U. S. A. **96:**12638–12643.
38. **Okada, N., Y. Oi, M. Takeda-Shitaka, K. Kanou, H. Umeyama, T. Haneda, T. Miki, S. Hosoya, and H. Danbara.** 2007. Identification of amino acid residues of *Salmonella* SlyA that are critical for transcriptional regulation. Microbiology **153:**548–560.
39. **Orsi, R. H., M. L. Borowsky, P. Lauer, S. K. Young, C. Nusbaum, J. E. Galagan, B. W. Birren, R. A. Ivy, Q. Sun, L. M. Graves, et al.** 2008. Short-term genome evolution of *Listeria monocytogenes* in a non-controlled environment. BMC Genomics **9:**539.
40. **Parkhill, J., G. Dougan, K. D. James, N. R. Thomson, D. Pickard, J. Wain, C. Churcher, K. L. Mungall, S. D. Bentley, M. T. G. Holden, et al.** 2001. Complete genome sequence of a multiple drug resistant *Salmonella enterica* serovar Typhi CT18. Nature **413:**848–852.
41. **Pike, R. M., and G. M. Mackenzie.** 1940. Virulence of *Salmonella typhimurium*. I. Analysis of experimental infection in mice with strains of high and low virulence. J. Bacteriol. **40:**171–195.
42. **Porwollik, S., R. M. Wong, R. A. Helm, K. K. Edwards, M. Calcutt, A. Eisenstark, and M. McClelland.** 2004. DNA amplification and rearrangements in archival *Salmonella enterica* serovar Typhimurium LT2 cultures. J. Bacteriol. **186:**1678–1682.
43. **Porwollik, S., R. M. Wong, and M. McClelland.** 2002. Evolutionary genomics of *Salmonella*: gene acquisitions revealed by microarray analysis. Proc. Natl. Acad. Sci. U. S. A. **99:**13807–13812.
44. **Porwollik, S., C. A. Santiviago, P. Cheng, L. Florea, S. Jackson, and M. McClelland.** 2005. Differences in gene content between *Salmonella enterica* serovar Enteritidis isolates and comparison to closely related serovars Gallinarum and Dublin. J. Bacteriol. **187:**6545–6555.
45. **Saint-Ruf, C., and I. Matic.** 2006. Environmental tuning of mutation rates. Environ. Microbiol. **8:**193–199.
46. **Sanderson, K. E., A. Hessel, and B. A. D. Stocker.** 1996. Strains of Salmonella and other Salmonella species used in genetic analysis, p. 2496–2503. *In* F. C. Neidhardt, R. Curtiss III, J. L. Ingraham, E. C. C. Lin, K. B. Low, B. Magasanik, W. S. Reznikoff, M. Riley, M. Schaechter, and H. E. Umbarger (ed.), *Escherichia coli* and *Salmonella typhimurium*: molecular and cellular biology, 2nd ed. ASM Presss, Washington, DC.
47. **Santiviago, C. A., M. M. Reynolds, S. Porwollik, S. H. Choi, F. Long, H. L. Andrews-Polymenis, and M. McClelland.** 2009. Analysis of pools of targeted *Salmonella* deletion mutants identifies novel genes affecting fitness during competitive infection in mice. PLoS Pathog. **5:**e1000477.
48. **Sellek, R. E., R. Escudero, H. Gil, I. Rodríguez, E. Chaparro, E. Pérez-Pastrana, A. Vivo, and P. Anda.** 2002. *In vitro* culture of *Borrelia garinii* results in loss of flagella and decreased invasiveness. Infect. Immun. **70:**4851–4858.
49. **Shaaper, R. M., and R. L. Dunn.** 1991. Spontaneous mutation in the *Escherichia coli lacI* gene. Genetics **129:**317–326.
50. **Somerville, G. A., S. B. Beres, J. R. Fitzgerald, F. R. DeLeo, R. L. Cole, J. S. Hoff, and J. M. Musser.** 2002. *In vitro* serial passage of *Staphylococcus aureus*: changes in physiology, virulence factor production, and *agr* nucleotide sequence. J. Bacteriol. **184:**1430–1437.
51. **Suzuki, T., K. Suzuki, Y. Tashiro, K. Saito, and D. Umeno.** 2007. Probing the mutation spectrum in *E. coli*. Nucleic Acids Symp. Ser. **51:**289–290.
52. **Swords, W. E., B. M. Cannon, and W. H. Benjamin, Jr.** 1997. Avirulence of LT2 strains of *Salmonella typhimurium* results from a defective *rpoS* gene. Infect. Immun. **65:**2451–2453.
53. **Thomson, N. R., D. J. Clayton, D. Windhorst, G. Vernikos, S. Davidson. C. Churcher, M. A. Quail, M. Stevens, M. A. Jones, M. Watson, et al.** 2008. Comparative genome analysis of *Salmonella* Enteritidis PT4 and *Salmonella* Gallinarum 287/91 provides insights into evolutionary and host adaptation pathways. Genome Res. **18:**1624–1637.
54. **Velicer, G. J., G. Raddatz, H. Keller, S. Deiss, C. Lanz, I. Dinkelackei, and S. C. Schuster.** 2006. Comprehensive mutation identification in an evolved bacterial cooperator and its cheating ancestor. Proc. Natl. Acad. Sci. U. S. A. **103:**8107–8112.
55. **Vernikos, G. S., N. R. Thomson, and J. Parkhill.** 2007. Genetic flux over time in the *Salmonella* lineage. Genome Biol. **8:**R100.
56. **Wilmes-Riesenberg, M. R., J. W. Foster, and R. Curtiss III.** 1997. An altered *rpoS* allele contributes to the avirulence of *Salmonella typhimurium* LT2. Infect. Immun. **65:**203–210.
57. **Woods, R., D. Schneider, C. L. Winkworth, M. A. Riley, and R. E. Lenski.** 2006. Tests of parallel molecular evolution in a long-term experiment with *Escherichia coli*. Proc. Natl. Acad. Sci. U. S. A. **103:**9107–9112.
58. **Zeigler, D. R., Z. Prágai, S. Rodriguez, B. Chevreux, A. Muffler, T. Albert, R. Bai, M. Wyss, and J. B. Perkins.** 2008. The origins of 168, W23, and other *Bacillus subtilis* legacy strains. J. Bacteriol. **190:**6983–6995.
59. **Zinder, N. D., and J. Lederberg.** 1952. Genetic exchange in *Salmonella*. J. Bacteriol. **64:**679–699.