Evolution of multi-gene segments in the *mutS*–*rpoS* intergenic region of *Salmonella enterica* serovar Typhimurium LT2

Michael L. Kotewicz, Baoguang Li, Dan D. Levy, J. Eugene LeClerc, Andrew W. Shifflet and Thomas A. Cebula

Author for correspondence: Thomas A. Cebula. Tel: +1 301 827 8281. Fax: +1 301 827 8260.
e-mail: tcebula@cfsan.fda.gov

Division of Molecular
Biology, Center for Food
Safety and Applied
Nutrition, Food and Drug
Administration,
8301 Muirkirk Road,
Laurel MD 20708, USA

The nucleotide sequence of the 12.6 kb region between the *mutS* and *rpoS* genes of *Salmonella enterica* serovar Typhimurium LT2 (*S. typhimurium*) was compared to other enteric bacterial *mutS*–*rpoS* intergenic regions. The *mutS*–*rpoS* region is composed of three distinct segments, designated HK, O and S, as defined by sequence similarities to contiguous ORFs in other bacteria. Inverted chromosomal orientations of each of these segments are found between the *mutS* and *rpoS* genes in related *Enterobacteriaceae*. The HK segment is distantly related to a cluster of seven ORFs found in *Haemophilus influenzae* and a cluster of five ORFs found between the *mutS* and *rpoS* genes in *Escherichia coli* K-12. The O segment is related to the *mutS*–*rpoS* intergenic region found in *E. coli* O157:H7 and *Shigella dysenteriae* type 1. The third segment, S, is common to diverse *Salmonella* species, but is absent from *E. coli*. Despite the extensive collinearity and conservation of the overall genetic maps of *S. typhimurium* and *E. coli* K-12, the insertions, deletions and inversions in the *mutS*–*rpoS* region provide evidence that this region of the chromosome is an active site for horizontal gene transfer and rearrangement.

Keywords: *mutS*, *rpoS*, horizontal gene transfer, *Escherichia coli*

INTRODUCTION

Genome-scale comparisons of microbial sequences are being made possible by the increasing availability of completely sequenced bacterial genomes. These comparisons offer the opportunity of exploring the basis for specific traits and tracing their likely evolutionary origin, such as the virulence traits in related pathogens or in particular pathogen and commensal relationships. This approach will be most fruitful when applied to areas of the genome that are genetically variable, for instance, where gene acquisitions by horizontal transfer have given otherwise conserved regions a polymorphic character. Regions of the chromosome that contain pathogenicity islands, as well as smaller blocks of virulence genes, represent such areas. Pathogenicity islands have most likely been acquired by horizontal gene exchange

(Hacker & Kaper, 2000); the virulence traits they engender allow the pathogen, in one genetic step, to adapt to a new host or to a new pathogenic niche. Thus, the comparison of such regions of the chromosome in related bacteria has the potential for revealing determinants that confer particular phenotypes characteristic of a pathogen.

Several lines of evidence demonstrate the polymorphic nature of the *mutS* genomic region and suggest that it has been a region active in horizontal transfer during its evolution. The *mutS* chromosomal region is the location for the insertion of blocks of virulence genes in the case of two widely diverged pathogens, *Salmonella enterica* serovar Typhimurium (*S. typhimurium*) and *Haemophilus influenzae*. In *S. typhimurium*, a 40 kb pathogenicity island (SPI-1) is inserted 5' to the *mutS* gene (Mills *et al.*, 1995). In *H. influenzae*, a 3.1 kb tryptophanase gene cluster (*tna*) is inserted on the 3' side of the *mutS* gene in strains that cause spinal meningitis in infants. This insertion allows the utilization of tryptophan and, thus, provides a growth advantage for the pathogen, particularly in the tryptophan-rich environment of cerebro-spinal fluid. Notably, the insertion is

Abbreviations: EPEC, enteropathogenic *Escherichia coli*; EHEC, enterohaemorrhagic *Escherichia coli*; SARC, *Salmonella* Reference Collection C; *S. typhi*, *Salmonella enterica* serovar Typhi; *S. typhimurium*, *Salmonella enterica* serovar Typhimurium.

The GenBank accession number for the sequence reported in this paper is AY050714.

absent from many non-pathogenic *H. influenzae* strains (Martin *et al.*, 1998).

Previously, the *mutS* region of the *Escherichia coli* chromosome was identified as having extensive genetic variability and a region that was subject to genetic exchange during the evolution of pathogenic lineages (LeClerc *et al.*, 1999). A nearly identical 2.9 kb segment of DNA between the *mutS* and *rpoS* genes is found in *E. coli* O157:H7 and *Shigella dysenteriae* type 1 strains, but it is absent in *E. coli* K-12 and related lineages. Abutting the novel sequence in *Shigella dysenteriae*, an IS1 element replaces the *prpB* gene found in both *E. coli* K-12 and O157:H7, suggesting that the newly emerged O157:H7 pathogen acquired the sequence from a *Shigella*-like ancestor (LeClerc *et al.*, 1999). Recent reports have further characterized genetic polymorphisms in this region within *E. coli* species. An insertion of 2.1 kb, in place of the 2.9 kb insert, is found in strains of uropathogenic *E. coli* (Culham & Wood, 2000), and larger intergenic regions exist in strains of enteropathogenic *E. coli* (EPEC) and enterohaemorrhagic *E. coli* (EHEC) (Herbelin *et al.*, 2000). In addition to the findings of genetic variability, phylogenetic analysis of EHEC and EPEC pathogens, as well as strains of the *E. coli* Reference (ECOR) collection, revealed that an unexpected level of recombination between *mutS* genes, and between surrounding sequences, has occurred during the evolution of these strains (Denamur *et al.*, 2000; Brown *et al.*, 2001a, b). Moreover, analysis of uropathogenic *E. coli* strains provides further evidence for both the mosaic character of the *mutS*–*rpoS* intergenic region and the horizontal transfer of *mutS* alleles (E. W. Brown, M. L. Kotewicz, D. J. Schu, J. Unowsky, B. Li, W. L. Payne, J. E. LeClerc and T. A. Cebula, unpublished results).

The close genetic relationship between *S. typhimurium* and *E. coli*, together with the identification of natural *mutS* mutators in a diverse collection of *Salmonella* outbreak strains, prompted us to examine the *mutS*–*rpoS* region in *S. typhimurium*, the prototypical laboratory *Salmonella* strain. This study presents a comparison of the *S. typhimurium* *mutS*–*rpoS* intergenic region to similar sequences identified in other species. Variation in the *mutS*–*rpoS* region found among natural isolates of *Salmonella* is assessed. During the comparison of these *mutS*–*rpoS* intergenic regions, the complete genome sequences of *S. typhimurium* (McClelland *et al.*, 2001) and *Salmonella enterica* serovar Typhi (*S. typhi*) (Parkhill *et al.*, 2001) have been published.

METHODS

Bacterial strains. The *Salmonella* Reference Collection C (SARC) was generously provided by F. Boyd (Boyd *et al.*, 1996). Other bacterial strains were from the Food and Drug Administration strain collection. The *S. typhimurium* LT2 strain used in these studies was *hisG46* (Hartman *et al.*, 1971).

Cloning. Cosmid clones were generated from *S. typhimurium*, as described by Sambrook *et al.* (1989). High-molecular-mass

genomic DNA was isolated either by using gentle SDS and phenol/chloroform extractions or by using a PUREGENE DNA Isolation Kit (Gentra) according to the manufacturer's directions. Partially digested *Sau3AI* genomic fragments in the 20–40 kb size range were generated using a twofold enzyme dilution series. Restriction fragments were ligated into the *Bam*HI site of cosmid pSCOS (kindly provided by S. D. Thomas and Dr G. Evans, Baylor University, Texas). After packaging the ligation mixture into bacteriophage λ particles using Gigapack Gold packaging extracts (Stratagene) and infection of *E. coli* DH5 α , kanamycin-resistant cells were selected. A library containing several thousand clones was generated.

Individual colonies of the library were grown in liquid culture in 96-well microtitre plates and spotted onto Brain Heart Infusion (BHI) agar plates containing 50 μ g ampicillin ml⁻¹. After overnight growth, the colonies were lifted onto Whatman no. 541 filter paper and processed as described previously (Kupchella *et al.*, 1994). Oligonucleotide probes were labelled using T4 polynucleotide kinase and [³²P]ATP (Sambrook *et al.*, 1989) and were used to hybridize bacterial DNA affixed to the filters. After autoradiography, positive colonies were selected from master plates. Positive colonies were purified by single-colony isolation, grown and then alkaline-lysed to prepare cosmid DNA using Midi-Maxi prep columns according to the manufacturer's directions (Qiagen).

After the identification of several cosmid clones spanning the *mutS*–*rpoS* region by probe and PCR analyses, cosmid cMK4 was subjected to more detailed restriction mapping, using *Pst*I, *Kpn*I and *Not*I. Four restriction fragments, encompassing the entire region between *mutS* and *rpoS*, were extracted from agarose gels and subcloned into appropriately digested pUC19 or pBluescriptKS (Stratagene) vectors. The set of subclones generated were: pBL1, a pUC19 plasmid containing a 4.2 kb *Kpn*I–*Kpn*I fragment; pBL2, a pBluescriptKS plasmid containing a 3.5 kb *Kpn*I–*Not*I fragment; pBL3, a pBluescriptKS plasmid containing a 3.5 *Not*I–*Pst*I fragment; and pBL4, a pUC19 plasmid containing a 1.8 kb *Pst*I–*Pst*I fragment.

Long PCR. The PCR primers used to identify and confirm positive cosmid colonies were positioned at the 3' end of *mutS* (5'-GTCTGGTGTAACGTCAG-3') and at the 3' end of *rpoS* (5'-CAATATGTCATCGGCACAG-3'). Long PCR was performed in a Robocycler (Stratagene) using the Perkin Elmer XL Kit. The following cycling conditions were used, 94 °C for 1.5 min followed by 30 cycles at 94 °C for 1.0 min, 55 °C for 1.5 min and 68 °C for 4.5 min. As expected, cosmid clones contained approximately 35 kb of insert DNA.

Sequencing. Plasmid inserts in pBL1 and pBL2 were sequenced commercially (Lark Technologies). The pBL3 and pBL4 inserts were sequenced in our laboratory by primer-walking sequencing of plasmid DNA and by sequencing *Sau3A1* subclones. An additional sequence determination was made on pBL3 and pBL4 plasmid DNA (Amplicon Express).

Sequence analysis. GenBank entries used for sequence comparisons included: U29579, *E. coli* K-12 strain MG1665 *mutS*–*rpoS* intergenic region; HIU32753, *H. influenzae* *nlpD* and *mutS*; HIU32781 and HIU32782, *H. influenzae* ORFs 1–7; AF054420, *E. coli* O157:H7 *mutS*–*rpoS* intergenic region; AE008832 and AE008833, *S. typhimurium* LT2 sections 140 and 141 of 224 of the complete genome; AL627276, *S. typhi* strain CT18 complete chromosome segment 12/20; Y13230, *Enterobacter cloacae* *rpoS*; AJ422108, *Enterobacter cloacae* *rpoS*, *slyA*, *pad1* and *yclC*; AJ422107, *Kluyvera cryocrescens* *rpoS* (partial), *slyA*, *pad1* and *yclC* (partial);

Fig. 1. The HK segment, a contiguous set of ORFs found in *H. influenzae*, *S. typhimurium* and *E. coli* K-12. In addition to the HK segment (grey arrows), the S segment (open arrow) and the O segment (solid arrow) are indicated in the *mutS-rpoS* intergenic region of *S. typhimurium*. Arrows indicate an arbitrarily defined polarity for each of the segments. Refer to Table 1 for the ORF nomenclature used.

Z99105, *Bacillus subtilis* genome segment 2 of 21. GenBank accession no. AF242211 contains the partial sequence for the *mutS-rpoS* intergenic region of *E. coli* E2348/69, an EPEC 1 strain, and is representative of the EPEC 1, EPEC 2 and EHEC 2 groups of enteric pathogens. We refer to them as EPEC/EHEC strains and distinguish them from the EHEC 1 group, which includes *E. coli* O157:H7 (Herbelin *et al.*, 2000). DNA sequence analysis was performed using the Wisconsin Package (version 10.2) of the Genetics Computer Group (GCG). BLAST and Clusters of Orthologous Groups of proteins (COG) analyses were performed at the National Center for Biotechnology (<http://www.ncbi.nlm.nih.gov>).

PCR and hybridization analyses of SARC strains. Based on the *S. typhimurium* sequence (AY050714), oligonucleotides were synthesized to generate PCR products from specific 2–4 kb targets within the *mutS-rpoS* intergenic region: S segment (2.0 kb), 5'-CGATAAGCAACCTCAAGCGAC-3' and 5'-AACGGCATAAATGCCCTGAC-3'; HK segment ORFs 1–2 (2.2 kb of the segment not present in *E. coli* K-12), 5'-TGACATCAGAGACACCCATC-3' and 5'-CGCCTTATTAACTCATCGC-3'; HK segment ORFs 3–7 (4.2 kb of the segment common to *E. coli* K-12 and *H. influenzae*), 5'-AGTTGATCTCGCCATCATCC-3' and 5'-ACCGTGACTTTATCTTCTGCC-3'; O segment (2.1 kb), 5'-GTTTTC-TCGTTTTACCAGCC-3' and 5'-TCACCATCTTCACAT-ATCCC-3'. The PCR oligonucleotide primers used for sizing the HK ORFs 5/6 deletion were ORF5for (5'-TCCAGG-TGCCGCTCATTGAGC-3') and ORF6rev (5'-GCCAGCA-ACGTTTATCGCATC-3'). The PCR conditions used were 94 °C for 1.5 min followed by 30 cycles at 94 °C for 1.0 min, 55 °C for 1.5 min and 68 °C for 1.5 min.

Probes were made by random priming of PCR products using [α -³²P]dATP (NEB BLOT Kit). Probe hybridization and wash conditions were as described above for oligonucleotide probes, except that strains were grown on BHI plates without the antibiotic; hybridization and washing took place at 60 °C using 6× SSC (saline sodium citrate).

RESULTS

The *mutS-rpoS* region of the *Salmonella* chromosome – gene positions and orientations

The genomes of *S. typhimurium* and *E. coli* K-12 are highly similar. Despite over 100 million years of evolutionary separation, the two organisms show a remarkable conservation of genome size and organization. The *mutS* genes of *S. typhimurium* and *E. coli* K-12 are located at 63.9 and 61.5 min, respectively, basically at the same position on their respective chromosomes (see Neidhardt *et al.*, 1996). Although the gene order in the two organisms often appears to be conserved (Liu *et al.*, 1993), the position of *mutS* relative to its neighbouring genes on the *Salmonella* genetic map is shown to be different from that in *E. coli* K-12 (Neidhardt *et al.*, 1996). The sequence of this region of the *S. typhimurium* chromosome was examined to determine the orientations of *mutS* and *rpoS*.

Fragments of approximately 35 kb in size were generated by partial *Sau*3A1 digestion of chromosomal DNA from *S. typhimurium*, and these were cloned into cosmid vector pSCOS. Several cosmids containing *mutS* and *rpoS* were identified by colony hybridization using *Salmonella*-specific probes. Four subclones were generated from one cosmid, cMK4, and sequenced. The sequence data spanning *mutS* to *rpoS* were assembled and analysed (GenBank accession no. AY050714). During the course of this work, the genome sequence for *S. typhimurium* was completed (McClelland *et al.*, 2001).

The *mutS-rpoS* intergenic sequence data demonstrate that the *mutS* and *rpoS* genes of *S. typhimurium* are separated by 12.6 kb of DNA, rather than by the 6.9 kb

Table 1. Similarity of *S. typhimurium* HK ORF sequences to *H. influenzae* and *E. coli* K-12 ORF sequences

Comparison of the HK segment ORFs of <i>S. typhimurium</i> and <i>H. influenzae</i>					
ORF name (this work)	GenBank ORF		Size of <i>S. typhimurium</i> ORF (aa)	Similarity (%)	
	<i>H. influenzae</i>	<i>S. typhimurium</i>		nt	aa
<i>mutS</i>	HI1707	STM2909	855	64	75
HK1	HI1015	STM2913	498	63	79
HK2	HI1014	STM2914	316	55	66
HK3	HI1013	STM2915	258	57	70
HK4	HI1012	STM2916	212	54	66
HK5	HI1011	STM2917	420	57	65
HK6	HI1010	STM2918	307	61	71
HK7	HI1009	STM2919	254	53	60
<i>rpoS</i>	NP	STM2924	330	–	–
Comparison of the HK segment ORFs of <i>S. typhimurium</i> and <i>E. coli</i> K-12					
ORF name (this work)	GenBank ORF		Size of <i>S. typhimurium</i> ORF (aa)	Similarity (%)	
	<i>E. coli</i> K-12	<i>S. typhimurium</i>		nt	aa
<i>mutS</i>	<i>mutS</i>	STM2909	855	83	91
HK1	NP	STM2913	498	–	–
HK2	NP	STM2914	316	–	–
HK3	o258	STM2915	258	73	84
HK4	o212	STM2916	212	77	88
HK5	o388	STM2917	420	73	84
HK6	o302	STM2918	307	73	80
HK7	o265	STM2919	254	77	89
<i>rpoS</i>	<i>rpoS</i>	STM2924	330	92	93

NP, Not present.

of DNA that separates these two genes in *E. coli* K-12. These data, in addition to the sequences of *fhlA*, SPI1 and *mutS* (Mills *et al.*, 1995), demonstrated the gene order *fhlA*–SPI1–*mutS*–*rpoS* in *S. typhimurium*. These results show that the orientation and position of the *Salmonella mutS* and *rpoS* genes relative to the flanking genes are the same as found for *E. coli* K-12 (Blattner *et al.*, 1997). The integrity of the cosmid clone and the size of the *mutS*–*rpoS* intergenic region were verified by long-PCR analysis of genomic *S. typhimurium* DNA (data not shown). Publication of the complete genome sequence of *S. typhimurium* (including corresponding GenBank accession nos AE008832 and AE008833) has confirmed our sequence and gene assignments. The complete genome data differed from our 13 163 bp sequence at two bases. Furthermore, the complete genome sequence of *S. typhi* (GenBank accession no. AL627276) reveals a sequence and gene arrangement in the *mutS*–*rpoS* region that is highly similar to that of *S. typhimurium*.

Multi-gene segments in the *S. typhimurium mutS*–*rpoS* intergenic region

Analysis of the 12.6 kb *mutS*–*rpoS* intergenic region suggested that this region was organized into three distinct segments. Each segment is composed of a group of contiguous genes or ORFs that are also found as contiguous genes in other species. For reference, the 7.0, 2.9 and 2.7 kb *Salmonella* segments have been designated HK, O and S, respectively, to denote their similarities to DNA sequences in *H. influenzae* and *E. coli* K-12 (HK), in *E. coli* O157:H7 and *Shigella dysenteriae* type 1 (O), and in *Salmonella* species (S) (Fig. 1).

The HK segment

The 12.6 kb DNA sequence between the *S. typhimurium mutS* and *rpoS* genes contains a 7.0 kb sequence that is similar to a 7 kb sequence found in *H. influenzae* (Fig. 1). The mean nucleotide similarity between these sequences is 58% and is relatively uniform across the HK

Fig. 2. Inverted orientations of the S (open arrow), HK (grey arrows) and O (solid arrows) segments found in *S. typhimurium* and related enteric bacteria. The approximate number of nucleotide base pairs between the end of *mutS* and the end of *rpoS* for each species is indicated at the right of the figure. Refer to Table 1 for the ORF nomenclature used. The area enclosed by the square brackets present on the EPEC 1, EPEC 2 and EHEC 1 taxa diagram indicates the portion of the region that has not been sequenced; however, the entire region has been mapped by RFLP analysis (Herbelin *et al.*, 2000).

segment. The amino-acid-coding similarities of seven contiguous ORFs provide stronger evidence for an ancestral relationship between these sequences. Because of multiple designations for ORFs found in the annotated *H. influenzae*, *E. coli* K-12 and *Salmonella* genome sequences, the ORFs in the HK segment have been labelled 1–7 for convenience of comparison across these species. Their amino-acid similarities range from 60 to 79% (Table 1), and are higher than nucleotide similarities. It should be noted that the *H. influenzae* HK segment is not located near *mutS* on the chromosome.

Comparisons of the *Salmonella* HK segment with the *mutS*–*rpoS* intergenic region of *E. coli* K-12 revealed the presence of similar coding sequences for ORFs 3–7. The nucleotide similarity between these *Salmonella* and *E. coli* K-12 ORF sequences ranges from 73 to 77% (Table 1). In place of HK ORFs 1 and 2 found in *Salmonella*, a single ORF, $\alpha 454$, is found in *E. coli* K-12 (Fig. 1). As depicted in Fig. 1, the orientation of the HK segments in *Salmonella* and *E. coli* is inverted relative to the *mutS* and *rpoS* genes in the two species. In addition to the HK segments in *H. influenzae*, *Salmonella* and *E. coli*, analysis for Clusters of Orthologous Groups of proteins (COG) showed one other set of contiguous HK ORFs in *Pasteurella multocida* (GenBank accession nos AE006173 and AE006174). This segment of contiguous

HK ORFs is not located near *mutS* on the *P. multocida* chromosome.

The O segment

A 2.9 kb segment of DNA adjacent to the *rpoS* gene of *S. typhimurium* is 81% similar to a 2.9 kb sequence in the *mutS*–*rpoS* intergenic region of pathogenic *E. coli* O157:H7. The sequence was identified as a novel sequence in *E. coli* O157:H7 because of its absence in *E. coli* K-12 (LeClerc *et al.*, 1996). A sequence identical to that in *E. coli* O157:H7 abuts an IS1 insertion element in *Shigella dysenteriae* type 1 strains (LeClerc *et al.*, 1999). As is the case for the HK segment, the O segment lies in inverted orientations in the *Salmonella* and *E. coli* lineages (Fig. 2).

The *Salmonella* O segment and the segment found in *E. coli* O157:H7 and *Shigella dysenteriae* type 1 have three common ORFs, named according to their, albeit weak, similarities to the *slyA*, *pad1* (*yclB*) and *yclC* genes (Carter *et al.*, 1999). *yclA*, *yclB*, *yclC* and *yclD* are four contiguous ORFs in the *Bacillus subtilis* genome (GenBank accession no. Z99105). Sequences similar to *yclB*, *yclC* and truncated *yclD* ORFs are found within the O segments of *Salmonella* and *E. coli* O157:H7. Despite the inverted orientations of the O segment in *Salmonella*

Table 2. Hybridization-probe analysis and long-PCR analysis of *Salmonella* SARC strains

SARC group	SARC no.	Strain	Segment probes				PCR product size (kb) [†]
			S (2 kb)	K (4 kb)*	H (2 kb)*	O (2 kb)	
I	SARC1	S4194	+	+	+	+	13·8
	SARC2	S3333	+	+	+	+	13·3
II	SARC3	S2985	—	+	+	+	11·7
	SARC4	S2993	+	+	+	+	13·8
IIIa	SARC5	S2980	+	+	—	+	8·5
	SARC6	S2983	+	+	—	+	8·5
IIIb	SARC7	S2978	+	+	+	+	13·8
	SARC8	S2979	+	+	+	+	13·8
IV	SARC9	S3015	+	+	+	+	13·8
	SARC10	S3027	+	+	+	+	13·8
V	SARC11	S3041	+	+	+	+	13·8
	SARC12	S3044	+	+	+	+	13·8
VI	SARC13	S2995	+	+	+	+	13·8
	SARC14	S3057	+	+	+	+	13·8
VII	SARC15	S3013	+	+	+	+	13·8
	SARC16	S3014	+	+	+	+	13·8
Probe controls							
	<i>S. typhimurium</i>	LT2	+	+	+	+	
	<i>Salmonella arizonae</i>	ATCC 13314	+	+	—	+	
	<i>Shigella dysenteriae</i>	SD567	—	—	—	+	
	<i>E. coli</i> O157:H7	EC536	—	—	—	+	
	<i>E. coli</i> K-12	W3110	—	+	—	—	

* Refer to Results for an explanation of these probes.

[†] See Fig. 3.

and *E. coli*, both groups have retained the truncated distal *yclD* ORF.

The S segment

A 2·7 kb sequence that abuts the end of *mutS* in *S. typhimurium* (Fig. 1) was designated as the S segment because it was characteristic for *Salmonella* species and was absent in *E. coli* strains. No similar sequence is present in the complete genomes of *E. coli* K-12 or *E. coli* O157:H7, nor in the other *E. coli* sequences in GenBank. The same position in *E. coli* strains is occupied by o218 (Fig. 1), which encodes a protein phosphatase (*prpB*, Missiakas & Raina, 1997; also designated *pphB*, Blattner *et al.*, 1997); o218 has no sequence similarity to ORFs found in the S segment.

The 2·7 kb S segment of *S. typhimurium* shows 65% nucleotide similarity to 2·7 kb of DNA found between the *slyA* and *rpoS* genes of a clinical isolate of *Enterobacter cloacae* (GenBank accession no. AJ422108). The O segment in *E. cloacae* is inverted relative to its orientation in *S. typhimurium* (Fig. 2). The same arrangement and orientation of the S and O segments found in *E. cloacae*, relative to the *mutS* and *rpoS* genes, is seen in *Klebsiella pneumoniae* (Washington University Genome Sequencing Center; <http://genome.wustl.edu/gsc>). Upon the examination of additional genome sequence data, Clusters of Orthologous

Groups of proteins (COG) analysis showed that the S segment ORFs are found in *Pseudomonas aeruginosa* (GenBank accession no. AE004560). In *P. aeruginosa*, the contiguous S segment ORFs are not located near the *mutS* and *rpoS* genes on the chromosome.

Segments in strains of the SARC collection

To examine the presence of the S, HK and O segments across a broad spectrum of *Salmonella* strains, colony hybridization using PCR probes generated from each of the segments in *S. typhimurium* was performed on strains from the SARC collection. The SARC collection contains two representative strains from each of eight characterized *Salmonella* groups, chosen to be broadly representative of the genetic diversity within *Salmonella* species (Boyd *et al.*, 1996). Random primer labelling of PCR probes spanning 2–4 kb and non-stringent hybridization conditions were used to assure the recognition of target sequences. Two probes were used for the analysis of the HK segment, since it comprises sequence similar to *H. influenzae* (ORFs 1–7) and *E. coli* K-12 (ORFs 3–7). In Table 2 the probe designated H hybridizes to ORFs 1 and 2, and the probe designated K hybridizes to ORFs 3–7.

The hybridization pattern generated for *S. typhimurium* was representative of most strains in the SARC collection (Table 2). Of the 16 SARC strains, 13 showed

Fig. 3. Agarose-gel electrophoresis of long-PCR products generated from the *mutS-rpoS* intergenic region of the SARC collection of *Salmonella* strains. Lanes 1–16, SARC strains nos 1–16 (refer to Table 2), respectively. The sizes of the molecular mass markers (M, in kb) are indicated at the left and right of the image.

hybridization signals for the S, HK and O segments. This is especially notable in the case of SARC group V, which contains two *Salmonella bongori* strains. Despite being the most evolutionarily distant group among the salmonellae (Reeves *et al.*, 1989), the group V strains SARC11 and SARC12 showed hybridization signals for all three segments.

Three strains did not show the patterns expected for hybridization to the S, HK and O segments (Table 2). The S segment probe did not hybridize with SARC3, a SARC group II strain, and the H probe for ORFs 1 and 2 failed to hybridize with SARC5 and SARC6, both *Salmonella arizonae* strains of SARC group IIIA.

Long-PCR analysis was performed on chromosomal DNA from each of the 16 SARC strains, to further characterize their *mutS-rpoS* intergenic regions. Using PCR primers from sites within the conserved *mutS* and *rpoS* genes, a 13.8 kb PCR product characteristic of *S. typhimurium* was produced in 12 of these strains (Fig. 3). SARC5 and SARC6, the two *S. arizonae* strains of group IIIA, yielded 8.5 kb products, which indicated a deletion of about 5 kb relative to the 13.8 kb PCR product. This result is consistent with the lack of hybridization of the H probe in these strains. The SARC3 strain showed a PCR product of 11.7 kb, indicating a deletion of about 2 kb, consistent with the probing data, which showed the lack of hybridization within the S segment.

Within SARC group 1, long-PCR analysis of SARC2, a *S. typhi* strain, produced a slightly smaller PCR product than SARC1, a *S. typhimurium* strain that yielded the expected 13.8 kb PCR product (Fig. 3). A comparison of the *S. typhimurium* sequence with the *S. typhi* genome sequence (AL627276) revealed a deletion of 859 bp in the *S. typhi* sequence, which corresponded to the sequence in *S. typhimurium* encoding the end of ORF 6 and the beginning of ORF 5. Use of PCR primers flanking the deletion indicated by the *S. typhi* sequence yielded a

PCR product of 0.2 kb from the chromosomal DNA of the SARC2 strain, whereas both SARC1 and our *S. typhimurium* strain yielded the expected 1.1 kb PCR product (data not shown). This deletion provides a convenient marker for distinguishing *S. typhi* strains from *S. typhimurium* strains using a simple PCR product length polymorphism.

DISCUSSION

Conserved ORFs, chromosomal linkage groups and selfish segments

Comparative analysis of the 12.6 kb *mutS-rpoS* intergenic sequence from *S. typhimurium* revealed three distinct segments, which have been designated HK, O and S on the basis of their sequence similarity to related species. Each segment is composed of protein-coding sequences that appear to be evolutionarily maintained as units. Although the orientation of each segment was found to be inverted relative to the *mutS* and/or *rpoS* genes of other species, the termini of the ORFs have been conserved. This suggests that the segments and their ORFs may be ‘selfish segments’, akin to selfish operons (Lawrence & Roth, 1966). Lawrence & Roth (1966) theorized that a natural symbiosis has permitted horizontal transfer to create and shape clusters of genes within the chromosome, reasoning that selective advantages order genes involved in the same biochemical pathway, making them contiguous. In a similar fashion, the evolution of ‘selfish segments’ may have been framed by selection pressures on traits within these sequences. The recombinational shuffling of sets of ORFs (segments) as functional units would then be favoured by selection; interrupted units would not be favoured. It is intriguing that each of the segments (i.e. HK, O and S) has an ORF at one terminus which has similarity to a transcriptional regulator in the LysR, DeoR and MarR families (Perez-Rueda & Collado-Vides, 2001).

In the HK segment, the ordered contiguous ORFs 1–7 relate DNA sequences across species as broadly divergent as *Salmonella* and *H. influenzae*, and they thus reiterate the chromosomal linkage within their last common ancestor (de Rosa & Labedan, 1998). Similar initiation and termination codon positions for the seven HK ORFs in *S. typhimurium* and *H. influenzae* reinforce this relationship. Furthermore, three pairs of ORFs in the HK segment show overlapping initiation and termination codons, such as within the sequence ATGA. This overlap, or translational coupling, has been shown to link the expression of genes whose co-expression or protein product ratios are critical (Andre *et al.*, 2000). In the complete genome sequence of *E. coli* K-12, 9% of ORFs show translational coupling (Blattner *et al.*, 1997). In the HK segment of *S. typhimurium*, translational coupling occurs at the sequence ATGA for ORF pair 1 and 2 and for ORF pair 4 and 5; ORF pair 5 and 6 is also translationally coupled within an ATGTTAA sequence. In *H. influenzae*, ORFs 4 and 5 are translationally coupled within an ATGA sequence. In the O segment in *S. typhimurium*, o120 and o402 are translationally

coupled within the sequence ATGA. The long-term retention of contiguous ORFs and these translational overlaps suggest functional linkages within the HK segment, as in the retention of functionally related members of operons.

Evolution of *mutS*–*rpoS* intergenic regions: chromosomal inversions and segment insertions and deletions

Sequences with similarity to each of the three segments (HK, O and S) present in the *mutS*–*rpoS* intergenic region from *S. typhimurium* were found in an inverted orientation in *E. coli* K-12, *E. coli* O157:H7 or *Enterobacter cloacae* strains (Fig. 2). The nucleotide and amino-acid similarities, as well as the conservation of ORFs, among the HK, O and S segments of *Salmonella* and *Escherichia coli* strains might be anticipated from the overall similarities between the genomes of these organisms, yet finding multiple inverted short segments within a limited region (10 kb) of the chromosome was unexpected.

Some inversions within chromosomes, characteristically centred about the axis of replication, seem to reflect ancient changes within bacteria [see reviews by Hughes (2000) and Eisen *et al.* (2000)]. Previous comparisons of the *Salmonella* and *E. coli* genetic maps demonstrated a large inversion with end points near map positions 26 and 36 min (Smith *et al.*, 1990; Sanderson *et al.*, 1996). Naturally occurring large inversions have been noted in the chromosome of *S. typhi*, where the chromosome appears to have been shuffled by recombination between homologous rRNA genes (Liu & Sanderson, 1996, 1998). Recently, three large inversions have been characterized for *Salmonella enterica* serovar Pullorum, a fowl-adapted pathogen (Liu *et al.*, 2002). In comparison to the above large chromosomal inversions, the inversions of short segments located between the *mutS* and *rpoS* genes in the *Salmonella* and *E. coli* genomes represent a new type of polymorphism between closely related bacterial species.

There is a contrast between the orderly accumulation of point mutations and the punctuated changes caused by such inversions or the insertion or deletion of segments in a chromosome. We have attempted to understand the evolution of the *mutS*–*rpoS* intergenic region, including segment orientations and the insertion or deletion of segments, by comparing the sequences of closely related strains/species. It is useful to examine the structural changes that have taken place in this chromosomal region in the context of enterobacterial evolution. The phylogenetic history of the enteric species has been constructed from highly conserved 16S rRNA sequences (see Dauga, 2002). Of note is the observation that *Klebsiella* and *Enterobacter* appear to have diverged earlier in enterobacterial evolution (i.e. they are more ancestral in the family tree), while *Salmonella* and *E. coli* emerged later. *Klebsiella* and *Enterobacter* both have a *mutS*–*rpoS* intergenic region containing the S and O segments (Fig. 2), suggesting that the ancestral enterobac-

terial genome also contained these two segments. The addition of the HK segment and the inverted orientations of segments between *Salmonella* and *Escherichia coli* suggest that these changes occurred later in the evolution of enteric bacteria and represent more recently acquired configurations in the *mutS*–*rpoS* intergenic regions of *Salmonella* and *E. coli* species.

Elements of the large intergenic region of *Salmonella* are most similar to those found in EPEC and EHEC strains (Herbelin *et al.*, 2000, and Fig. 2). These pathogenic *E. coli* strains contain homologues of both the *Salmonella* HK segment and the O segment. The simplest model to account for these strain differences is that two inversions, one of the HK segment and another of the O segment, have occurred since *Salmonella* and *E. coli* shared a common ancestor that contained the large intergenic region (Herbelin *et al.*, 2000).

Simple inversions, however, fail to explain the different ORFs found at inversion junctions in *Salmonella* and *E. coli*. In one example, ORFs 1 and 2 in the HK segment in *Salmonella* are absent in known *E. coli* strains. *Salmonella* ORFs 1 and 2 are replaced by o454 in *E. coli* strains. Although superficially resembling ORF 1 in its position and size, o454 shows only 46% nucleotide similarity with ORF 1, compared to similarities of 73–77% with ORFs 3–7 (Table 1). Furthermore, the amino-acid similarity between *E. coli* o454 and *Salmonella* ORF 1 is 49%, far lower than the mean 85% similarity exhibited by the other HK cohorts (Table 1). Protein BLAST analysis indicates that both ORF 1 and o454 belong to the permease families of transporters with strong similarity to gluconate-transport proteins. The nucleotide and amino-acid comparisons, however, as well as the absence of the contiguous ORF 2 in *E. coli*, argue that o454 of *E. coli* and ORF 1 of *Salmonella* are not homologues and that o454 has been substituted for *Salmonella* ORFs 1 and 2.

In another example, o218 (protein phosphatase *prpB*) is found next to the 3' end of *mutS* in *E. coli* at the junction of the inverted versions of the HK segment. The S segment occurs at this position in *Salmonella*. The absence of o218 next to *mutS* in *Salmonella*, and in the evolutionarily older genus *Klebsiella*, suggests that its acquisition in the *E. coli* lineage was more recent. This contradicts the hypothesis that o218 is present in the common ancestor of both *Salmonella* and *E. coli* (Herbelin *et al.*, 2000). *Salmonella* species do contain o216, which is 61% similar to o218 (see Fig. 2). o216 has been referenced as *pigE* (pathogenicity island gene, Pancetti & Galan, 2001) and as *pphB* (protein phosphatase, McClelland *et al.*, 2001; also see Shi *et al.*, 2001). o216 is found on the 5' side of *mutS* next to or as part of the SPI pathogenicity island insertion in *Salmonella*, whereas o218 is found on the 3' side of *mutS* in the *E. coli* lineage. We, therefore, propose that o216 and o218 have been independently acquired by the *Salmonella* and *E. coli* lineages, respectively.

The large – and probably ancestral-like – *mutS*–*rpoS* intergenic region in *S. typhimurium*, containing the HK,

O and S segments identified here, is present in much of the genetic spectrum of *Salmonella* as represented by the SARC collection of 16 strains. The exceptions to this are the apparent absence of HK ORFs 1 and 2 in SARC group IIIA, the loss of the S segment in SARC3 and the deletion at the ORF 5/6 junction in SARC2 and other *S. typhi* strains. These changes are likely to represent recombinational shuffling in the *mutS-rpoS* region, whose polymorphic character is more apparent among *E. coli* strains when similar comparisons are made. Two points may be made here. First, recombination in the region may lead to more subtle changes, such as intragenic recombinations, which will be best uncovered by sequence analysis. In this light, the examination of *mutS* sequence evolution among SARC strains has implicated a role for intragenic recombination in the *S. typhi* SARC2 strain mentioned above, at a site within the *mutS* gene upstream of the characterized deletion (Brown *et al.*, 2002). Second, comparisons of *Salmonella* strains within a group – i.e. strains that share a pathological niche – may show exchange on a greater scale than the broadly representative strains of the SARC collection. The surprisingly high incidence (> 1%) of *mutS* mutators among natural isolates of *Salmonella* (LeClerc *et al.*, 1996; LeClerc & Cebula, 2000) and the heightened ability of these mutators to recombine foreign DNA with host sequences (Matic *et al.*, 1995) make such exchanges likely.

SUMMARY

Comparative genome analysis demonstrated multiple inversions and deletions in the *mutS-rpoS* intergenic region of the chromosomes of enteric bacteria. The collinear and highly conserved chromosomes of *Salmonella* and *E. coli* species show a high degree of polymorphism in this intergenic region. However, segments comprising contiguous ORFs from distant lineages, notably from *H. influenzae*, have been conserved.

ACKNOWLEDGEMENTS

We thank William L. Payne for his technical assistance and A. Tormo, J. M. Navarro and E. Martinez for communication of results prior to publication. We acknowledge Eric Brown's contributions to our thoughts on segment changes and phylogeny, and thank Billie Barnett and Marcia L. Meltzer for their help with figures.

REFERENCES

- Andre, A., Puca, A., Sansone, F., Brandi, A., Antico, G. & Calogero, R. A. (2000). Reinitiation of protein synthesis in *Escherichia coli* can be induced by mRNA *cis*-elements unrelated to canonical translation initiation signals. *FEBS Lett* **468**, 73–78.
- Blattner, F. R., Plunkett, G., 3rd, Bloch, C. A. & 14 other authors (1997). The complete genome sequence of *Escherichia coli* K-12. *Science* **277**, 1453–1474.
- Boyd, E. F., Wang, F. S., Whittam, T. S. & Selander, R. K. (1996). Molecular genetic relationships of the salmonellae. *Appl Environ Microbiol* **62**, 804–808.
- Brown, E. W., LeClerc, J. E., Li, B., Payne, W. L. & Cebula, T. A. (2001a). Phylogenetic evidence for horizontal transfer of *mutS* alleles among naturally occurring *Escherichia coli* strains. *J Bacteriol* **183**, 1631–1644.
- Brown, E. W., LeClerc, J. E., Kotewicz, M. L. & Cebula, T. A. (2001b). Three R's of bacterial evolution: how replication, repair, and recombination frame the origin of species. *Environ Mol Mutagen* **38**, 248–260.
- Brown, E. W., Kotewicz, M. L. & Cebula, T. A. (2002). Detection of recombination in *Salmonella enterica* using the incongruence length difference test. *Mol Phylogenet Evol* **24**, 102–120.
- Carter, P. E., Butler, L., Booth, I. R. & Thomson-Carter, F. M. (1999). Characterization of the *mutS-rpoS* region from STEC and non-STEC. In *Abstracts of the 99th General Meeting of the American Society for Microbiology*, p. 237. Washington, DC: American Society for Microbiology.
- Culham, D. E. & Wood, J. M. (2000). An *Escherichia coli* reference collection group B2- and uropathogen-associated polymorphism in the *rpoS-mutS* region of the *E. coli* chromosome. *J Bacteriol* **182**, 6272–6276.
- Dauga, C. (2002). Evolution of the *gyrB* gene and the molecular phylogeny of *Enterobacteriaceae*: a model molecule for molecular systematic studies. *Int J Syst Evol Microbiol* **52**, 531–547.
- Denamur, E., Lecomte, G., Darlu, P. & 9 other authors (2000). Evolutionary implications of the frequent horizontal transfer of mismatch repair genes. *Cell* **103**, 711–721.
- de Rosa, R. & Labedan, B. (1998). The evolutionary relationships between the two bacteria *Escherichia coli* and *Haemophilus influenzae* and their putative last common ancestor. *Mol Biol Evol* **15**, 17–27.
- Eisen, J. A., Heidelberg, J. F., White, O. & Salzberg, S. L. (2000). Evidence for symmetric chromosomal inversions around the replication origin in bacteria. *Genome Biol* **1**, RESEARCH0011.1–0011.9.
- Hacker, J. & Kaper, J. B. (2000). Pathogenicity islands and the evolution of microbes. *Annu Rev Microbiol* **54**, 641–679.
- Hartman, P. E., Hartman, Z. & Stahl, R. C. (1971). Classification and mapping of spontaneous and induced mutations in the histidine operon of *Salmonella*. *Adv Genet* **16**, 1–34.
- Herbelin, C. J., Chirillo, S. C., Melnick, K. A. & Whittam, T. S. (2000). Gene conservation and loss in the *mutS-rpoS* genomic region of pathogenic *Escherichia coli*. *J Bacteriol* **182**, 5381–5390.
- Hughes, D. (2000). Evaluating genome dynamics: the constraints on rearrangements within bacterial genomes. *Genome Biol* **1**, REVIEWS0006.1–0006.8.
- Kupchella, E., Koch, W. H. & Cebula, T. A. (1994). Mutant alleles of tRNA (*Thr*) genes suppress the *hisG46* missense mutation in *Salmonella typhimurium*. *Environ Mol Mutagen* **23**, 81–88.
- Lawrence, J. G. & Roth, J. R. (1996). Selfish operons: horizontal transfer may drive the evolution of gene clusters. *Genetics* **143**, 1843–1860.
- LeClerc, J. E. & Cebula, T. A. (2000). *Pseudomonas* survival strategies in cystic fibrosis. *Science* **289**, 391–392.
- LeClerc, J. E., Li, B., Payne, W. L. & Cebula, T. A. (1996). High mutation frequencies among *Escherichia coli* and *Salmonella* pathogens. *Science* **274**, 1208–1211.
- LeClerc, J. E., Li, B., Payne, W. L. & Cebula, T. A. (1999). Promiscuous origin of a chimeric sequence in the *Escherichia coli* O157:H7 genome. *J Bacteriol* **181**, 7614–7617.
- Liu, S. L. & Sanderson, K. E. (1996). Highly plastic chromosomal organization in *Salmonella typhi*. *Proc Natl Acad Sci USA* **93**, 10303–10308.
- Liu, S. L. & Sanderson, K. E. (1998). Homologous recombination

- between *rrn* operons rearranges the chromosome in host-specialized species of *Salmonella*. *FEMS Microbiol Lett* **164**, 275–281.
- Liu, S. L., Hessel, A. & Sanderson, K. E. (1993). Genomic mapping with I-Ceu I, an intron-encoded endonuclease specific for genes for ribosomal RNA, in *Salmonella* spp., *Escherichia coli*, and other bacteria. *Proc Natl Acad Sci USA* **90**, 6874–6878.
- Liu, G. R., Rahn, A., Liu, W. Q., Sanderson, K. E., Johnston, R. N. & Liu, S. L. (2002). The evolving genome of *Salmonella enterica* serovar Pullorum. *J Bacteriol* **184**, 2626–2633.
- Martin, K., Morlin, G., Smith, A., Nordyke, A., Eisenstark, A. & Golomb, M. (1998). The tryptophanase gene cluster of *Haemophilus influenzae* type b: evidence for horizontal gene transfer. *J Bacteriol* **180**, 107–118.
- Matic, I., Rayssiguier, C. & Radman, M. (1995). Interspecies gene exchange in bacteria: the role of SOS and mismatch repair systems in evolution of species. *Cell* **80**, 507–515.
- McClelland, M., Sanderson, K. E., Spieth, J. & 23 other authors (2001). Complete genome sequence of *Salmonella enterica* serovar Typhimurium LT2. *Nature* **413**, 852–856.
- Mills, D. M., Bajaj, V. & Lee, C. A. (1995). A 40 kb chromosomal fragment encoding *Salmonella typhimurium* invasion genes is absent from the corresponding region of the *Escherichia coli* K-12 chromosome. *Mol Microbiol* **15**, 749–759.
- Missiakas, D. & Raina, S. (1997). Signal transduction pathways in response to protein misfolding in the extracytoplasmic compartments of *E. coli*: role of two new phosphoprotein phosphatases PrpA and PrpB. *EMBO J* **16**, 1670–1685.
- Neidhardt, F. C., Curtiss, R., III, Ingraham, J. L. & 7 other editors (1996). *Escherichia coli and Salmonella: Cellular and Molecular Biology*, 2nd edn, vols I and II. Washington, DC: American Society for Microbiology.
- Pancetti, A. & Galan, J. E. (2001). Characterization of the *mutS*-proximal region of the *Salmonella typhimurium* SPI-1 identifies a group of pathogenicity island-associated genes. *FEMS Microbiol Lett* **197**, 203–208.
- Parkhill, J., Dougan, G., James, K. D. & 38 other authors (2001). Complete genome sequence of a multiple drug resistant *Salmonella enterica* serovar Typhi CT18. *Nature* **413**, 848–852.
- Perez-Rueda, E. & Collado-Vides, J. (2001). Common history at the origin of the position–function correlation in transcriptional regulators in archaea and bacteria. *J Mol Evol* **53**, 172–179.
- Reeves, M. W., Evins, G. M., Heiba, A. A., Plikaytis, B. D. & Farmer, J. J., 3rd (1989). Clonal nature of *Salmonella typhi* and its genetic relatedness to other salmonellae as shown by multilocus enzyme electrophoresis and proposal of *Salmonella bongori* comb. nov. *J Clin Microbiol* **27**, 313–320.
- Sambrook, J., Fritsch, E. F. & Maniatis, T. (1989). *Molecular Cloning: a Laboratory Manual*, 2nd edn. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory.
- Sanderson, K. E., Hessel, A., Shu-Lin, L. & Rudd, K. E. (1996). The genetic map of *Salmonella typhimurium*, Edition VIII. In *Escherichia coli and Salmonella: Cellular and Molecular Biology*, 2nd edn, pp. 1903–1999. Edited by F. C. Neidhardt and others. Washington, DC: American Society for Microbiology.
- Shi, L., Kehres, D. G. & Maguire, M. E. (2001). The PPP-family protein phosphatases PrpA and PrpB of *Salmonella enterica* serovar Typhimurium possess distinct biochemical properties. *J Bacteriol* **183**, 7053–7057.
- Smith, C. M., Koch, W. H., Franklin, S. B., Foster, P. L., Cebula, T. A. & Eisenstadt, E. (1990). Sequence analysis and mapping of the *Salmonella typhimurium* LT2 *umuDC* operon. *J Bacteriol* **172**, 964–978.

Received 8 January 2002; revised 14 May 2002; accepted 14 May 2002.