**Recombinational basis of serovar diversity in Salmonella enterica**

(Jia Li*, Kimberly Nelson*, Alma C. McWhorter†, Thomas S. Whittam*, and Robert K. Selander‡)

*Molecular Evolutionary Genetics, Pennsylvania State University, University Park, PA 16802; †Centers for Disease Control, Atlanta, GA 30333

**ABSTRACT**

The fliC gene, which encodes phase 1 flagellin, was sequenced in strains of 15 Salmonella enterica serovars expressing flagellar antigenic factors of the g series. The occurrence of each of the flagellin serotypes g,m, m,t, and g,zsl in distantly related strains is the result of horizontal exchange of DNA, as indicated by identity or close similarity in nucleotide sequence of all or parts of the antigenic factor-determining central region of fliC. The flagellin genes of some serovars are complex mosaic structures composed of diverse segments derived through multiple recombination events. Thus, recombination of horizontally transferred segments (intragenic) or entire genes (assortative) within and among subspecies is identified as a major evolutionary mechanism generating both allelic variation at the fliC locus and serovar diversity in natural populations. Evidence that flagellar serological diversity is promoted by diversifying selection in adaptation to host immune defense systems or flagellotropic phage is discussed.

The primary basis for classification of strains of the pathogenic bacterium Salmonella enterica is a serotyping scheme in which 2324 serovars have been recognized on the basis of the antigenic properties of the phase 1 and phase 2 flagellar proteins (H1 and H2 antigens), which are encoded by the fliC and fliB genes, respectively, and the cell-surface lipopolysaccharide (O antigen), which is synthesized by enzymes specified by genes of the rfb region (1, 2). For the phase 1 flagellin, 52 antigenic factors and 61 serotypes (single factors or combinations of factors) have been distinguished.

The demonstration by multilocus enzyme electrophoresis (MLEE) that the same polysaccharide and flagellin serotypes may occur in distantly related strains suggested that horizontal transfer and recombination events involving the fliC, fliB, and rfb genes are relatively frequent (3–6). For fliC, this hypothesis subsequently was supported by partial sequencing of the gene in strains of several serovars (7) and identification of a plasmid-borne fliC-like gene (8). But the recombination events could not be classified as intragenic or assortative (entire gene) (9), and the generality of these findings is unknown.

We here report the results of a comparative sequence analysis of fliC in strains of 15 S. enterica serovars for which overall genomic relatedness has been estimated by MLEE. (10) Our findings demonstrate that recombination is a major evolutionary mechanism generating both allelic variation at the fliC locus and serovar diversity in natural populations.

**MATERIALS AND METHODS**

**Bacterial Isolates.** Fifteen strains of S. enterica subspecies I, II, IV, and VII that express combinations of phase 1 flagellar antigenic factors of the g series (f, g, m, s, t, and zsl) (1, 10, 11) (Table 1) were analyzed.

<table>
<thead>
<tr>
<th>Serovar*</th>
<th>ET</th>
<th>Strain no.</th>
<th>Antigenic factors†</th>
</tr>
</thead>
<tbody>
<tr>
<td>Banana</td>
<td>Ba1</td>
<td>S5332</td>
<td>4,12</td>
</tr>
<tr>
<td>Berta</td>
<td>Be1</td>
<td>S5321</td>
<td>1,9,12</td>
</tr>
<tr>
<td>Derby</td>
<td>De13</td>
<td>S241</td>
<td>1,9,12</td>
</tr>
<tr>
<td>Enteritis</td>
<td>En1</td>
<td>S53</td>
<td>1,9,12</td>
</tr>
<tr>
<td>Newmexico</td>
<td>Nm1</td>
<td>S5323</td>
<td>9,12</td>
</tr>
<tr>
<td>Oranienburg</td>
<td>Or1</td>
<td>S5331</td>
<td>6,7,14</td>
</tr>
<tr>
<td>Othmarschen</td>
<td>Otl</td>
<td>S5334</td>
<td>6,7,14</td>
</tr>
<tr>
<td>Pensacola</td>
<td>Pe1</td>
<td>S5325</td>
<td>1,9,12</td>
</tr>
<tr>
<td>11 6,7,g,m,s,t</td>
<td>II1</td>
<td>S5333</td>
<td>6,7</td>
</tr>
<tr>
<td>11 6,7,g,m,s,t</td>
<td>II2</td>
<td>S5336</td>
<td>6,7</td>
</tr>
<tr>
<td>11 42,g,t,13</td>
<td>II3</td>
<td>S2993</td>
<td>6,7</td>
</tr>
<tr>
<td>11 48,g,m,t</td>
<td>II4</td>
<td>S5335</td>
<td>48</td>
</tr>
<tr>
<td>11 6,7,m</td>
<td>II5</td>
<td>S5337</td>
<td>6,7</td>
</tr>
<tr>
<td>11 45,g,zsl</td>
<td>IV1</td>
<td>S3015</td>
<td>45</td>
</tr>
<tr>
<td>11 7,45</td>
<td>VII1</td>
<td>S3013</td>
<td>1,40</td>
</tr>
</tbody>
</table>

*ET, electrophoretic type. †Except as noted by Roman numerals, the serovars are of subspecies I. ‡O, somatic lipopolysaccharide; H1, phase 1 flagella.

MLEE. The strains were assayed for allelic variation at 27 enzyme loci by standard methods (12, 13).

**RESULTS**

**Sequence Variation in Relation to Flagellin Structure.** Sequence variation in the fliC gene and the flanking regions among 15 strains of 15 serovars of S. enterica is shown in Fig. 1. In S. enterica, as in other bacteria (14), the N- and C-terminal regions of the flagellin molecule are strongly conserved in both amino acid sequence and length, whereas the central region is highly variable (11, 13, 15–19). The terminal regions are involved in secretion and polymerization (20), and amino acid substitutions in the central region, which forms a knoblike protrusion from the flagellar core (21, 22), produce the antigenic variation assayed in serotyping (8, 13, 18, 19, 23–25).

For purposes of the present analysis, three regions of fliC are distinguished (Figs. 1 and 2): C1 (conserved region 1)

Abbreviations: MLEE, multilocus enzyme electrophoresis; ET, electrophoretic type.

‡To whom reprint requests should be addressed.

§The sequences reported in this paper have been deposited in the GenBank database (accession nos. M84974, U06197–U06206, U06225–U06227, U06455).

2552
includes codons 1–181; V (variable region) is the central segment composed of codons 182–370; and C2 (conserved region 2) consists of codons 371–505. Among the 15 strains, C1 and C2 are invariant in size, but in the V region, ET II 2 has a deletion of codon 319, and ETs Nm 1, IV 1, and VII 1 share a deletion of codon 216 (Fig. 1). Six other ETs have
three additional codons between codons 222 and 223. As shown in Fig. 2, both nonsynonymous and synonymous substitutions occur much more frequently in V than in C1 and C2. Almost all amino acid polymorphism in C1 and C2 reflects differences between strains of serotype g,z51 and those of other serotypes.

Evidence of Recombination. It is instructive to compare evolutionary trees based on (i) MLEE analysis, indexing overall genetic relatedness (Fig. 3A), (ii) the nucleotide sequence of the combined C1 and C2 regions of fliC (Fig. 3B), and (iii) the sequence of the central gene of the fliC gene (Fig. 3C). If evolution of fliC has involved little or no recombination, trees for the V region and the C1 and C2 regions should be topologically similar, although rates of substitution may differ, and both gene-segment trees would be expected to resemble the MLEE tree. The occurrence of flagellin of the same serotype in distantly related strains might reflect either mutational convergence in amino acid sequence or the retention of an amino acid sequence from a common ancestor, but, in either case, because of silent nucleotide substitutions, strains of like flagellin should not cluster closely together on the basis of their fliC sequences. In contrast, recombination of DNA acquired through horizontal transfer among strains will be indicated by a clustering of fliC sequences specifying the same flagellin serotype, regardless of the overall genetic relatedness of the strains in which they occur.

Serotype g,m. Strains of both serovars Enteritidis (En 1) and Ochtmarschen (Ot 1) express serotype g,m (Table 1). They are divergent for chromosomal genetic character (Fig. 3A), yet their C1 + C2 fliC sequences are identical (Figs. 1 and 3B), and their V regions differ by only a single nonsynonymous substitution (bp 946) (Figs. 1 and 3C).

Serotype g,z51. Strains of ETs Nm 1, IV 1, and VII 1 are of serotype g,z51. Although they represent three subspecies (Fig. 3A), they cluster together in the C1 + C2 sequence tree (Fig. 3B). Note that the 5' half of C1 in VII 1 is similar to that of the subspecies I ET De 13 and very different from that of IV 1 or Nm 1 (Fig. 1). In the sequence of the V region, IV 1 and VII 1 are identical, and Nm 1 is more similar to these two strains than it is in the case of C1 + C2.

Serotype m(t), ETs Pe 1, Ba 1, Or 1, and II 5 are of serotype m(t). The three subspecies I ETs are not closely allied, and all three are very distantly related to II 5 (Fig. 3A). On the basis of the C1 + C2 sequence, their relationships are closer, but II 5 is still relatively distant; in the V region tree (Fig. 3C), the four ETs form a tight cluster that is strongly divergent from other clusters and lineages. From examination of Fig. 1, it is apparent that the V region and part of C2 have been exchanged among the four ETs, and as noted below, part of this very distinctive sequence has also been transferred to ETs Be 1 and II 4.

Other Serotypes. Most of the fliC gene of De 13 (f,g) is similar in sequence to that of En 1, but a part of region V is shared with ETs of subspecies II. The fliC gene of ET II 3 (f,g,t) is also a mosaic of segments: C1 has both subspecies I- and subspecies II-like segments; the 5' half of V is like that of En 1; and the 3' half is subspecies II sequence. Unusually complex mosaic fliC structures are shown by Be 1 (f,g,t) and II 4 (g,m,t). In Be 1, C1 is a mixture of small
segments of subspecies I and II, with a piece of subspecies II sequence in the middle of C1. The 3' part of V is the distinctive m,t serotype sequence until just after the position of the three extra codons; then it becomes subspecies I-like, with a short segment of subspecies II in the middle; and the 3' part of V is subspecies II sequence. ET II 4 also has the m,t sequence in the 5' part of V, and the middle part is somewhat like the m,t sequence, but the 3' part of the region is subspecies I-like.

The two ETs of serotype g,m,s,t (II 1 and II 2) cluster together in all three trees. In C1 there are seven base pair differences, and in C2 there is one nucleotide difference, but their V region sequences are identical, with the notable exception of a segment between bp 940 and 957, in which they differ at 11 sites and a codon has been deleted in II 2. This nonrandom clustering of polymorphic sites, which was identified by Stephens' (28) statistical test, strongly suggests a recombination event involving II 2. Application of Stephens' test also confirmed other examples of significant phylogenetic partitions among the flic sequences of the 15 strains noted above.

**DISCUSSION**

**Amino Acid Sequence Variation and Tertiary Flagellin Structure.** According to Vonderviszt et al. (29), the central region of the flagellin polypeptide consists of two domains that differ in physicochemical properties. Whereas the terminal regions are α-helical strands, the central region is composed of β-strands and β-turns, with a small α-helical segment centering on residue 300 and corresponding to the point where the two domains join. Our study revealed that there has been little sequence change in the segment of the flic gene centering on codon 300, in contrast to the numerous substitutions that have occurred in adjacent segments (Fig. 2).

**Variation in Gene Size.** Some of the diversity in flic arises from alterations in length of the central region. Frankel et al. (23, 24) showed that deletion of a 261-bp segment changes the serotype of serovar Typhi from d to j, and Newton et al. (30) observed antibody-selected spontaneous deletions that affected the antigenic properties of the flagellin in serovar Muenchen (d). In the central region, we found two single-codon deletions in four strains and a three-codon addition in the six strains that have acquired all or part of the distinctive m,t sequence. In other work, we have found that the flic sequence of serovar Antarctica (g,z3) is identical to that of Enteritidis En 1 (g,m), except for the duplication of codons 309-314 (J.L., unpublished data).

**Point Mutation as a Source of Serotypic Variation.** The flagellin sequences studied here are too heterogeneous for us to deduce the molecular genetic basis of the antigenic differences, but studies of other strains have demonstrated that serotype modification may be produced by replacement of only one or a few amino acids (see also ref. 30). For example, the flagellins of Enteritidis (g,m) and Bledgam (g,m,q) differ by a single amino acid (J.L., unpublished data), and Dublin (g,p) differs from Enteritidis (g,m) by three substitutions (18). [A recent report (11) of eight amino acid differences between the sequences of these serovars was based on an erroneous sequence for a strain of Dublin (J.L., unpublished data).]

**Horizontal Transfer and Recombination.** The occurrence of each of the three flagellin serotypes g,m, m,t, and g,z3 in distantly related strains is clearly attributable to horizontal exchange and recombination rather than to convergence in amino acid sequence or retention of ancestral sequences. Recombination of segments of the central region has been especially extensive among ETs of subspecies II, and it has generated great diversity in flagellin and lipopolysaccharide serotype combinations, as well as new flagellin serotypes, of which f,g,t of II 3 and g,m,t of II 4 may be examples. The genes of some ETs—notably Be 1 (f,g,t) and II 4 (g,m,t)—are complex mosaics as a consequence of multiple intragenic recombination events, and it is not unlikely that most S. enterica flic genes are to some degree mosaic in structure. Similar mosaic organization has been reported for the highly variable IgA protease genes of Neisseria spp. (31) and Haemophilus influenzae (32).

Notwithstanding the considerable sequence diversity among flic alleles of strains expressing g-series antigenic factors, all of them are much more similar to one another than any of them is to the serotype a, c, d, i, and r alleles (J.L., unpublished data). The occurrence of markedly divergent families of alleles in S. enterica suggests the possibility that sequences are from time to time recruited from the flagellin genes of other species, or even from other types of genes. It is perhaps noteworthy that the central region of the flagellin of serovar Paratyphi A shows some similarity in amino acid sequence and structure to comparable regions of certain serine proteases (33). Gene conversion, which has been demonstrated experimentally for S. enterica flagellin genes by Okazaki et al. (19), presumably explains why some antigenic factors that are normally intrinsic to the phase 2 flagellin gene (fliB) (for example, e,n,x) may also be encoded by flic genes (2).

The evidence of extensive recombination in flic within and between subspecies of S. enterica revealed by the present study contrasts with the picture emerging from comparable analyses of genes encoding most metabolic enzymes and other housekeeping proteins, including gapA (34), putP (35), and mdh (36), in which recombination events are rare.

**Adaptive Function of Flagellar Polymorphism.** Opinions regarding the biological significance of the extensive flagellar antigenic polymorphism in S. enterica and other bacteria vary widely. One view (37) is that for populations of pathogenic bacteria allelic polymorphism in the dominant surface antigens, such as those of the flagella, is adaptive in permitting reinfection of hosts. And Reeves (38) has suggested that antigenic variation in flagellin and the cell-surface lipopolysaccharide is subject to "niche-specific selection" in S. enterica and Escherichia coli. But for the flagellins of S. enterica, Joys (39) advocated a model of "unconstrained evolution" in which random amino acid substitutions accumulate by mutation in the central region in the absence of functional constraint and resultant counter-selection. Even if the central region is under some constraint, since β-sheet structure and net electrostatic charge are conserved (14), in this interpretation antigenic diversity has no adaptive function per se. A corollary of this neutral mutation hypothesis is that effective (realized) rate of recombination may also be relatively high.

The problem is to determine to what extent flagellar antigenic diversity is due to neutral amino acid changes and to what degree it is promoted by frequency-dependent or other types of diversifying selection of mutants and recombinants. That diversity has some adaptive function in S. enterica is indicated by the existence of phase variation (diphasic condition), which is clearly an evolved mechanism by which individual cell lineages may alternately express flagella of different serotypes (1, 40). But, unfortunately, there is little available information on the role of antigenic variation among serovars in immune system evasion (14, 39, 41). The observation that sensitivity to flagellotropic phage may be serotype dependent (42) suggests another possible adaptive basis for flagellin polymorphism.

One finding of the present study may be relevant. The rate of synonymous substitution per synonymous site (dS), as well as that of nonsynonymous substitution per nonsynonymous site (dN), is significantly higher in the central region of the flic gene than in the conserved terminal regions. The simplest
explanation for the elevation of $d_S$ is the recombinational incorporation of segments of divergent sequence derived from other bacteria (see discussion in ref. 43). But the fact that the ratio $d_S/d_N$ is much smaller in the central region (0.544/0.059 = 9.2) than in the terminal regions of fIC (0.227/0.008 = 28.3) and in several other genes of S. enterica we have analyzed (34–36) is suggestive of the action of positive selection for amino acid replacement substitutions in the highly variable central region.

We thank P. Ho, S. A. Ploock, and S. L. Spigelmyer for technical assistance. This research was supported by Public Health Service Grants AI-22144 (to R.K.S.) and AI-24566 (to T.S.W.) from the National Institutes of Health.