Molecular Analysis of a *Salmonella enterica* Group E1 *rfb* Gene Cluster: O Antigen and the Genetic Basis of the Major Polymorphism

Lei Wang, Lajwant K. Romana and Peter R. Reeves

*Department of Microbiology, The University of Sydney, New South Wales 2006, Australia*

Manuscript received May 3, 1991
Accepted for publication November 22, 1991

ABSTRACT

*Salmonella enterica* is highly polymorphic for the O antigen, a surface polysaccharide that is subject to intense selection by the host immune system. This polymorphism is used for serotyping *Salmonella* isolates. The genes encoding O antigen biosynthesis are located in the *rfb* gene cluster. We report here the cloning and sequence of the 19-kb *rfb* region from strain M32 (serovar anatum, group E1) and compare it with that of strain LT2 (serovar *typhimurium*, group B). Genes for biosynthetic pathways common to both strains are conserved and have very similar sequences. In contrast, the five genes for CDP-aborque synthesis, present in strain LT2, are absent in strain M32; three open reading frames (ORFs) of strain LT2, thought to include genes for transferases, are not present in strain M32 but are replaced by three different ORFs with little or low level of similarity. Both *rfb* gene clusters are low in G+C content, indicating that they were transferred from a common ancestral species with low G+C content to *S. enterica* relatively recently (in the evolutionary sense). We discuss the recombination and lateral transfer events which may have been involved in the evolution of the polymorphism.

We are studying a major polymorphism in the O antigen of *Salmonella enterica*, with over 60 different forms. We have analyzed several of them and there is good evidence that much of this extensive variation arose by lateral transfer of genes (Brahmbhatt et al. 1988; Verma, Quigley and Reeves 1988; Wyk and Reeves 1989; Verma and Reeves 1989; Jiang et al. 1991; Liu et al. 1991; Brown, Romana and Reeves 1991). The O antigen is on the surface and appears to be a target of intense selection by the immune system and this probably accounts for the origin and maintenance of the high level of variation.

Most bacterial species exist as a series of clones (Selander, Caugant and Whittam 1987), and in the case of *S. enterica*, these clones frequently differ in O antigen type. The O antigen variation was one of the criteria used to divide *S. enterica* into different species although all clones are now considered to belong to a single species, *S. enterica*, because of the high degree of nucleic acid relatedness among these isolates shown by DNA/DNA hybridization (Le Minor and Popoff 1987; Brenner 1984; Crosa et al. 1973). In this paper we use the old "species" names such as "typhimurium" or "anatum" as serovar names as recommended by Le Minor and Popoff (1987).

The O antigen is part of the lipopolysaccharide (LPS), present in the outer membrane of Gram-negative bacteria. LPS consists of three structural regions: the lipid A moiety, which is embedded in the outer bacterial membrane through hydrophobic interactions, an oligosaccharide core, and the O antigen. The O antigen comprises many repeats of an oligosaccharide unit, each unit consisting typically of three to six sugar residues. The variations among different O antigen structures are manifested in the types of sugars present, their arrangement within the repeat unit or the linkage between repeat units. The highly variable nature of the O antigen provides the basis for serotyping into group A, group B, etc., each group then being divided into serovars using other antigens.

The O antigen, as a polysaccharide, is not a primary gene product, but rather an assemblage of sugar moieties, each the product of a specific biosynthetic pathway. The genes which encode the enzymes of those pathways, or parts of pathways, specific to O antigen synthesis, plus the genes for the sugar transferases which assemble the O antigen oligosaccharide units, are in general located in the *rfb* gene cluster which maps at 42 min on the chromosome (Sanderson and Roth 1989). Most of the O antigen variation among *S. enterica* groups is thus the result of genetic variation in the *rfb* gene cluster. The polymerase which catalyzes the linkage between subunits of O antigen is encoded by *rfe* which is located far away from the *rfb* gene cluster in groups A, B and D (Mäkelä and Stocker 1984; Collins and Hacket 1991; Naide et al. 1965), but probably within the *rfb* gene cluster in group E1 (Nyman et al. 1979). We previously analyzed the genetic variation between *rfb* gene clusters of groups A, B and D (Brahmbhatt et al. 1988; Verma, Quigley and Reeves 1988; Wyk and Reeves 1989; Verma and Reeves 1989; Jiang et al. 1991;
events that replaced one set of genes by another unrelated set. We further suggest that the group E1 rfb region, like that of group B, arose in a low G + C content species and was acquired by S. enterica by lateral transfer.

MATERIALS AND METHODS

Bacterial strains and plasmids: S. enterica strain M32 which was used as the DNA source for cloning was kindly given by the Institute of Medical and Veterinary Science (IMVS), Adelaide, Australia. The plasmids derived from strain M32 are shown in Figure 2, together with the plasmids from strain LT2.

Enzymes and radiochemicals: Restriction enzymes, DNA polymerase I, bacterial alkaline phosphatase and T4 DNA ligase were purchased from Boehringer Mannheim Biochemicals, Indianapolis, Indiana; chemicals were obtained from Sigma Chemical Co. and Ajax Chemicals Pty. Ltd, Sydney; deoxy-nucleotides, deoxy-nucleotides and Taq DNA polymerase were purchased from Pharmacia LKB Biotechnology; an Applied Biosystems 370A DNA sequencer and fluorescent dye-labeled M13 universal primers were from Applied Biosystems; other radiochemicals were from Bresatec Limited, Adelaide, South Australia, and Amersham Corp., Arlington Heights, Illinois.

DNA techniques: Standard methods were used for DNA preparation, agarose gel electrophoresis, radioactive labeling of DNA, DNA hybridization, Southern blotting, autoradiography, ligation and bacterial transformation (Maniatis, Frisch and Sambrook 1982). The high and low stringency hybridization experiments were as described by Howley et al. (1979). All sequencing was by the chain termination technique of Sanger, Nicklen and Coulson (1977). Most of the sequencing was done using nested sets of deletions. Custom made oligonucleotides complementary to sequenced rfb DNA were used as primers for filling in a few gaps and for the sequencing of one small region.

Nested sets of deletions were made using exonuclease III (Henikoff 1984), and pT7T3-19U (from Pharmacia), pGEM-Tzf(+) (from Promega), pUC18 or pUC19 (Norlander, Kempe and Messing 1983) as the vector. pT7T3 and pGEM carry the IG region from M13 which enables derivatives to be packaged as single-stranded DNA (ssDNA) during superinfection with helper phage M13K07 (Vieira and Messing 1987). The inserts in pUC18 and pUC19 were transferred into M15 mp18 or M13 mp19 (Norlander, Kempe and Messing 1983) and ssDNA was made from these clones. In both cases the ssDNA was sequenced by the dideoxy method, in general using fluorescent dye-labeled M13 universal primers and the gels were read after autoradiography.

Computer analysis of sequence data: The sequence data were analyzed using the MBIS System (Buchholz and Reiner 1986) at the CSIRO Biotechnology Laboratory, Sydney, and the Australian National Genomic Information Service (ANGIS) at Sydney University; both incorporate several sets of programs. Sequence data was assembled using the SAP program of Staden (1982, 1986); molecular weights and hydrophobicity indices were calculated by the program ANALYSEQ (Staden 1982); the program PREDICT, which combines eight different methods was used to produce a joint protein secondary structure prediction (Elipoupolos et al. 1982); the program ALOM based on the
**RESULTS**

Homology between the rfb gene clusters of strains M32 and LT2, and cloning of the rfb gene cluster from strain M32: Cloned segments of strain LT2 DNA were used as probes in Southern blotting of strain M32 chromosomal DNA. The segments covering the two ends of the LT2 rfb gene cluster (pPR301 and pPR465, Figure 2) hybridized strongly to M32 chromosomal DNA whereas the segment covering the central region (pPR300) did not hybridize even at low stringency.

A SacI bank of M32 chromosomal DNA was constructed in plasmid pUC18 and screened using the insert of plasmid pPR301 as a radioactive probe. Plasmid pPR964 (Figure 2) was selected and used as a probe to screen an EcoRI bank of M32 chromosomal DNA in pUC18 and several clones were isolated. Preliminary restriction enzyme analysis showed that one of these clones, pPR966, extended away from the gnd gene (Figure 2). This chromosomal walking procedure was repeated using pPR966 to select pPR1030 from a M32 HindIII bank in pUC8, which was then

---

**Nucleotide sequence accession number:** The sequences from position 5,337 to 13,000 and from position 16,808 to 19,094 reported in this paper have been assigned the EMBL accession numbers X60665 and X60666, respectively.
used to isolate pPR1029 from a M32 EcoRI bank in pUC8.

The insert of pPR1029 (M32) hybridized strongly to the insert of pPR465 which covers the left end of the LT2 rfb region, and the insert of pPR964 (M32) hybridized strongly to the insert of pPR301 which covers the right end of the LT2 rfb region (Figure 2). This indicates that the inserts of pPR964, pPR966, pPR1030 and pPR1029 together cover the whole M32 rfb region.

Southern hybridization analysis using the inserts of pPR964, pPR966, pPR1030 and pPR1029 as probes against themselves and the M32 chromosomal DNA was carried out to compare patterns of EcoRI, HindIII and ScaI sites in plasmid and chromosomal DNA. No evidence was found for structural rearrangements during cloning.

**Detailed restriction enzyme mapping:** A detailed restriction enzyme map (Figure 2) of the rfb region of strain M32 was obtained. Each plasmid was first cleaved in single enzyme digests with each of seven enzymes. Double restriction enzyme digests were then carried out as required.

Comparison of the restriction enzyme map of the M32 rfb region with that of strain LT2 showed that most of the restriction sites from positions 0.0–5.6, from positions 17.8–21.4 and from positions 22.2–23.0 on the LT2 rfb map were conserved in strain M32 (Figure 2). Four sites (PstI at position 13.7, BamHI at position 13.9, ScaI at position 14.3 and ScaI at position 15.7) of the 14 restriction sites (all 6 base pair sites) within the region from position 13.3–17.2 on the M32 map were absent in strain LT2 (Figure 2): we estimate that there are approximately 5% (4 × 100/(14 × 6)) nucleotide differences between strains M32 and LT2 in this region. One site (ClaI at position 3.2) of the nine restriction sites within the region from position 0.0–5.6 on the LT2 rfb map was not present in strain M32 giving an approximate estimate of 1–2% nucleotide differences between strains M32 and LT2 in this region, or an average of 3% for the two regions.

**The regions unique to strain M32:** The first two of the regions of high similarity discussed above flank different central regions in each strain with neither similarity in restriction map nor detectable homology by Southern blotting. To define the junctions between these regions, we probed EcoRI and HindIII digested M32 clones (pPR1029, pPR1030, pPR966 and pPR964) with the inserts of LT2 clones, pPR465, pPR341, pPR301 and pPR472. The results of these hybridizations, summarized in Table 1, show that the first boundary is in the region from positions 5.3–7.2 of strain M32, and the second boundary is in the region from positions 11.8–13.3. The restriction maps also show differences in the region from positions 16.9–19.0 of strain M32. The 0.5-kb region from positions 21.7–22.2 of strain LT2 is replaced in strain M32 by 2.1-kb DNA.

**Sequencing:** Both strands of the M32 rfb regions from positions 5.3–13.0, and from positions 16.8–19.1 were sequenced (Figures 3 and 4).

Seven complete ORFs were found in these sequences, all on the same strand. The sequences of these seven ORFs were used to search the GenBank, NBRF and EMBL sequence databases, and no genes were found with significant similarity to them at either
Figure 3.-Part 1
**rfr Region of S. enterica Group El**

**Start of orf10.8**

**Start of homology with strain LT2**

**Start of orf11.9**

**Figure 3.—Part 5**
Figure 4.—The DNA sequence from position 16,808-19,094 of the rfb region of Salmonella strain M52, with the amino acid sequences of ORFs shown above the DNA sequence. One direction of this sequence was sequenced by cloning the 3.4-kb SacI-KpnI fragment (position 15.7-19.1) into pCEM and making nested deletions using KpnI and XbaI as 3' and 5' overhang enzymes. The other direction was sequenced by cloning the SacI-HindIII fragment (position 15.7-17.4) and the HindIII-KpnI fragment (position 17.4-19.1) into M13 mp18 and M13 mp19, respectively, and using custom made primers. The start position and Shine-Dalgarno sites, where present, are underlined. The sequences of the corresponding segments of strain LT2 are given underneath, with only the nucleotide and amino acid differences presented.
the nucleotide or amino acid levels. The opposite strands of these two regions were also analyzed but revealed no ORFs of more than 200 bases in length. The ORFs are named according to the start positions (in kb) on the map (e.g., orf17.4 (M32) starts at approximately base 17,400 on the M32 map). The putative translation start site and, where present, the Shine-Dalgarno sequence for each ORF are given in Figures 3 and 4 (Gold and Stormo 1987).

DISCUSSION

Comparison of the rfb gene clusters of groups E1 and B: We have cloned the rfb gene cluster of S. enterica strain M32 (group E1) as a series of overlapping clones, and compared it to that of strain LT2 (group B) by Southern hybridization and restriction mapping. The data revealed regions of homology, often nearly identical; and other regions with no significant homology at all (Figure 5). The regions unique to strain M32 were sequenced: the new sequences extended into the regions nearly identical in strain LT2 for which we have the sequence of the whole rfb region, and we thus have in effect the sequence of the whole of the rfb gene cluster of strain M32.

Analysis of the sequences revealed six complete ORFs in the region from positions 5,337–13,000 and one ORF in the region from positions 16,808–19,094 on the same strand. orf6.1, orf7.0 and orf11.9 are clearly homologous to genes in strain LT2 (Figure 3). However comparison of orf7.9, orf9.6, orf10.8 and orf17.4 with the ORFs of strain LT2 revealed no homology at nucleotide sequence level. There is thus a sharp distinction between genes which can be aligned and genes with no detected similarity.

Genes common to group B and group E1: The O antigen biosynthetic pathway for group B is shown in Figure 6. We have previously identified many of the rfb genes in strain LT2 (Wyk and Reeves 1989; Jiang et al. 1991). We expect the TDP-rhamnose and GDP-mannose biosynthetic pathway genes and the galactose transferase and rhamnose transferase genes all to be present in both strains; but the CDP-abequose biosynthetic pathway and transferase genes of strain LT2 to be absent in strain M32, and the mannose transferase genes of strains M32 and LT2 to be different. This expectation is borne out: the genes which have been identified in group B as encoding the TDP-rhamnose and GDP-mannose biosynthetic pathways, the galactose transferase and rhamnose transferase, are all conserved in strain M32 (Figure 5).

In strain LT2, three of the CDP-abequose pathway genes (rfbF, rfbG and rfbJ) have been identified (Figure 5). Comparison of the amino acid sequences of the protein products of these three genes with those of orf7.9, orf9.6 and orf10.8, which are located in the central region of the M32 rfb region, did not reveal any homology. The other two CDP-abequose pathway genes in strain LT2 have not yet been identified but comparison of rfb clusters of groups B and C2 (P. K. Brown, L. K. Romana and P. R. Reeves, unpub-
UDP-Gal

\[ \text{UMP} \]

\( \text{p-Und} \)

\( \text{Pi} \)

\( \text{pp-Und} \)

Abe

\( \text{Abe} \)

[Man-Rha-Gal]→pp-Und

[Man-Rha-Gal]→Man-Rha-Gal-pp-Und

Core LPS

Complete LPS

**Figure 6.**—Pathway of O antigen biosynthesis (McGrath and Osborn 1991). Steps 1, 2, 3, and 4 are catalyzed by galactose, rhamnose, mannose, and abequose transferases, respectively. Step 5 is catalyzed by O antigen polymerase. Abbreviations: Und, undecaprenol; pp, pyrophosphoryl; others as in Figure 1.

lished data) suggests that they are *orf7.6* and *orf10.4*, both of which are absent in strain M32. It appears that as expected the whole CDP-abequose biosynthetic pathway is absent in strain M32.

Most of the conserved genes are essentially identical in the two strains. However the M32 genes (*orf7.0* and *orf11.9*) which are located between the central unique region and flanking nearly identical regions show substantial differences from their corresponding genes (*orf7.0* and *orf16.5*) of LT2. The deduced protein products of *orf7.0* (represented as *orf7.0p*) of M32 and LT2 show 68.2% identity and 79.0% similarity, and the structural predictions for these two proteins show the same pattern (data not shown).

*orf11.9p* of M32 shows 75.2% identity and 85% similarity to *orf16.5p* of LT2. Both *orf7.0* and *orf11.9* of M32 are in the same positions in the gene cluster as the corresponding genes in strain LT2 and, as the level of similarity is high and there are no insertions and deletions observed in the comparisons, we believe that they are identical in function in the two strains.

**Genes unique to group E1:** *orf7.9p* (a putative membrane protein): *orf7.9* is one of the three ORFs in the central region with no homology to genes in strain LT2 at nucleotide sequence level. Structural analysis of *orf7.9p* predicts 12 transmembrane segments distributed throughout its length, suggesting that the whole protein is embedded in the membrane. No part seems likely to be hydrophilic, suggesting that this protein is not a transferase as we would expect a hydrophilic domain for the active site. This ORF is the only *rfb* gene encoding a 12-transmembrane segment protein in strain M32, and *orf12.8* (LT2) which is located in the same relative position is the only such gene in strain LT2 (Jiang et al. 1991). These data suggest that the two genes may have a similar function, although *orf7.9p* (512 amino acid residues) is 82 residues longer than *orf12.8p* (LT2), and there is only 50.5% similarity (20.8% identity) between the two genes at the amino acid level. They may be involved in O-antigen export (see Jiang et al. 1991).

Is *orf7.4* the O-antigen polymerase gene? *orf17.4* shows significant similarity to the sequence of the *S. enterica* strain C5 (serovar typhimurium, group B) *rfe* gene (Collins and Hackett 1991) which encodes the O-antigen polymerase; having 21.2% identity and 54.6% similarity at the amino acid level. The predicted secondary structure and hydrophobicity were also similar. Five thousand shuffle sequences of *orf17.4p* were generated, of which only eleven (0.22%) showed a level of similarity to *rfe* higher than 54% (the highest was 55.6% and the mean was 47.56%). We propose that *orf17.4p* catalyzes O-antigen polymerization in strain M32 (step 5 in Figure 6).

*orf9.6* and *orf10.8*: *orf9.6p* and *orf10.8p* are 392 and 361 amino acid residues in length, respectively, and do not show any significant similarity with LT2 *rfb* genes at either nucleotide or amino acid level. *orf9.6* encodes a highly hydrophobic protein with ten potential transmembrane segments, but none were found in *orf10.8p*. The linkages of mannose to rhamnose differ in the O antigens of strains M32 and LT2, and hence we expect the mannose transferase gene of strain M32 to differ from that in strain LT2. *orf9.6* and *orf10.8* are the only unidentified genes in M32, and either could be the mannose transferase gene.

**Structure of the *rfb* gene cluster of strain M32:** We have cloned the 19-kb *rfb* gene cluster of strain M32 as four overlapping clones, and hence have not directly determined the minimum region required for function. Several attempts to clone the whole 19-kb gene cluster in one piece were unsuccessful.

The strain LT2 *rfb* gene cluster extends from the end of *orf2.8* (position 3.7) to the end of *rfbP* (position 21.7) with no DNA outside this region needed for function (Jiang et al. 1991; L. Wang and P. R. Reeves, unpublished data). The region from positions 0.0-6.9 is conserved in strain M32 and the *rfb* gene cluster of strain M32 probably starts from position 3.7 as in strain LT2. The 141 bp sequence between *rfbP* (the last *rfb* gene in LT2) and *gnd* in strain LT2 (Jiang et al. 1991).
et al. 1991) is replaced by a 1,510-bp sequence in strain M32 (Figure 4). The nonidentity begins 3 bases after the end of rfbP and ends 22 bases before the start of gnd. orf17.4 (in M32) is located in this region and has considerable homology with the rfc gene of group B. We have not yet confirmed that orf17.4 (in M32) encodes the O antigen polymerase, but if so then the "rfb" region (including rfc) of strain M32 will extend to position 18,465 (the end of orf17.4).

Of the seven ORFs in the sequences presented in this paper, two (orf10.8 and orf11.9) do not have Shine-Dalgarno sequences (Figure 3). orf10.8 starts with an ATG codon two bases upstream of the stop codon of orf9.6, and the putative start codon of orf11.9 lies three bases after the stop codon of orf10.8: these two ORFs are presumably initiated by translational coupling. orf16.5 (LT2) which corresponds to orf11.9 (M32) starts immediately following the stop codon of the previous ORF (Jiang et al. 1991); and could also be initiated by translational coupling although it has a good Shine-Dalgarno sequence. The Shine-Dalgarno sequences of orf6.1 and orf7.0 are very similar to those of the corresponding ORFs in LT2 (Figure 3). orf7.9 and orf9.6 have three base Shine-Dalgarno sequences and no possibility of translational read-through, and are perhaps translated at low efficiency. orf17.4 has a good Shine-Dalgarno sequence.

Five programs from O'NEILL were used to search for promoters in the sequenced M32 rfb region. A large number of potential promoters were found, mostly with a Berg-von Hippel index too high for them to be likely to function unless under positive control. Among them, one with the mRNA startpoint at position 9502 is one of the top ranked and locates in the intergenic gap between orf7.9 and orf9.6. In strain LT2 there are two potential promoters in the flanking conserved region with the mRNA startpoints at positions 2803 and 3997 (Jiang et al. 1991). We suggest that the M32 rfb gene cluster is a single operon translated from the same promoter as strain LT2, but may have an additional internal promoter with mRNA startpoint at position 9502.

**Evolution of the rfb polymorphism:** The M32 rfb region is of low G+C content. The region described in this paper has a G+C content ranging from 0.29 to 0.37, and the ORFs have a codon usage characteristic of low G+C species, with P3 being lower than P1 or P2 (Table 2) (P1, P2 and P3 are the corrected average G+C contents for bases 1, 2 and 3 of codons; SUEOKA 1988). These values together with those for some of the genes common to strains M32 and LT2 are given in Table 2. The regions common to strains M32 and LT2 are also of low G+C content and have the codon usage of low G+C content strains (see Jiang et al. 1991). The G+C content of a gene reflects the long term effects of the balance between the mutation rates from G-C to A-T and A-T to G-C (SUEOKA 1988). We might expect that once a gene is in S. enterica, the mutational pressure which normally maintains a given G+C content would bring this value to about 0.5 which is usual for S. enterica genes. To account for the low G+C content and atypical (for S. enterica) codon usage of the group B rfb gene cluster, we proposed that it had been transferred to the S. enterica strain LT2 from a low G+C content species (Wyk and Reeves 1989; Jiang et al. 1991), and the same argument applies to the rfb gene cluster of group E.

There has in general been substantial synonymous substitution involving the third base since Escherichia and Salmonella diverged (OCHMAN and WILSON 1987), and had the rfb gene cluster been in Salmonella during this period, then the G+C content of the third base would have been adjusted to approximate that of these two genera. The fact that P3 is typical of a low G+C content organism suggests that the rfb gene cluster has been in Salmonella for only a small part of the time since the divergence.

The regions common to strains LT2 and M32 show very little difference and must be derived from a common ancestral sequence, but the group specific central regions, in contrast, have little in common and at most one of the two forms can have been associated with the ancestral flanking regions. It is not possible with only the evidence of the two current forms to say if both central regions have recently become associated with the flanking regions or if one has been associated for much longer and was replaced by the other in a relatively recent event.

The junctions between the central and flanking regions are interesting. orf7.0 and orf11.9 are adjacent

---

**TABLE 2**

<table>
<thead>
<tr>
<th>Gene</th>
<th>G + C</th>
<th>P1</th>
<th>P2</th>
<th>P3</th>
</tr>
</thead>
<tbody>
<tr>
<td>orf6.1(M32)</td>
<td>0.455</td>
<td>0.58</td>
<td>0.378</td>
<td>0.449</td>
</tr>
<tr>
<td>[orf6.1(LT2)</td>
<td>0.457</td>
<td>0.567</td>
<td>0.371</td>
<td>0.447</td>
</tr>
<tr>
<td>orf7.0(M32)</td>
<td>0.339</td>
<td>0.511</td>
<td>0.30</td>
<td>0.20</td>
</tr>
<tr>
<td>[orf7.0(LT2)</td>
<td>0.404</td>
<td>0.567</td>
<td>0.348</td>
<td>0.285</td>
</tr>
<tr>
<td>orf7.9(M32)</td>
<td>0.389</td>
<td>0.35</td>
<td>0.279</td>
<td>0.231</td>
</tr>
<tr>
<td>orf9.6(M32)</td>
<td>0.314</td>
<td>0.398</td>
<td>0.325</td>
<td>0.226</td>
</tr>
<tr>
<td>orf10.8(M32)</td>
<td>0.387</td>
<td>0.485</td>
<td>0.288</td>
<td>0.227</td>
</tr>
<tr>
<td>orf11.9(M32)</td>
<td>0.375</td>
<td>0.445</td>
<td>0.399</td>
<td>0.277</td>
</tr>
<tr>
<td>[orf16.3(LT2)</td>
<td>0.375</td>
<td>0.460</td>
<td>0.374</td>
<td>0.299</td>
</tr>
<tr>
<td>orf17.4(M32)</td>
<td>0.294</td>
<td>0.359</td>
<td>0.277</td>
<td>0.254</td>
</tr>
</tbody>
</table>

The G + C content of each gene and the values of P1, P2, and P3 as the corrected average G+C contents for codon position 1, 2 and 3 (SUEOKA 1988). The features of some LT2 rfb genes (Jiang et al. 1991) which are underlined are presented in brackets for comparison.
to the central region in strain M32, as are orf7.0 and orf16.5 in strain LT2 (Figure 2). The alignment of the orf7.0 genes is quite clear up to within 24 and 39 bases of the ends of orf7.0 (M32) and orf7.0 (LT2), respectively (orf7.0 (LT2) extends 15 bases beyond orf7.0 (M32), Figure 2), whereas there is no similarity between orf7.9 (M32) and orf7.6 (LT2) which follow orf7.0 in strains M32 and LT2, respectively. However although the alignment of the two genes is clear and the junction between similar and nonsimilar regions is sharp, the level of similarity decreases from within orf6.1 of both strains up to the junction (Figure 7), and a similar gradient is also found at the other junction with the level of similarity between orf11.9 (M32) and orf16.5 (LT2) increasing from the junction up to the ends of these ORFs (Figure 7, also see Figure 3). There have been several reports of intragenic recombination during the evolution of Escherichia coli (DUBOSE, DYKHUIZEN and HARTL 1988; BISERCIC, FEUTRIER and REEVES 1991; NEILSON, WHITTAM and SELANDER 1991). Such events allow those regions of rfb gene clusters which are common to two forms to undergo random genetic drift as one, while the nonhomologous regions must drift independently, as each must be independently maintained and there is no possibility of genetic recombination within these regions. We suggest that at the time of the recombination event which generated the junctions, orf7.0 (M32) and orf11.9 (M32) were very similar to orf7.0 (LT2) and orf16.5 (LT2) respectively, but proximity to the regions of non homology reduced homologous recombination, such that the neutral mutations accumulated in the two rfb gene clusters are different. The gradient reflects the increasing possibility of recombination with distance from the central region. The abrupt drop in the level of difference at about position 6,850 (Figure 7) could be due to a recent recombination between two forms of rfb clusters at this site (event 3 in Figure 8).

The rfb gene clusters of strains M32 and LT2 have two regions of non homology. One of them, orf7.9 to orf10.8 of strain M32 and orf7.6 to orf15.4 of strain LT2, is flanked at both ends by homologous DNA with a gradient of similarity; but the other, orf17.4 of strain M32, is flanked by DNA which is very similar to LT2 sequence right up to the junctions. This difference may reflect a difference in age of the recombination event involved; but if orf17.4 does indeed encode the group E1 O antigen polymerase, it must have evolved with the orf7.9 to orf10.8 region which includes some of the transferase genes. It is possible however that as in strain LT2, the rfc gene originally mapped elsewhere and has recently been brought into juxtaposition to the rfb gene cluster.

The level of difference in the common genes, away from those located in the junctions, is approximately 1–5% based on restriction enzyme data, and 1.2% based on sequence data. This is within the normal range for pairwise comparison between genes in the related E. coli (BISERCIC, FEUTRIER and REEVES 1991; DUBOSE, DYKHUIZEN and HARTL 1988; MILKMAN and CRAWFORD 1983), and presumably in S. enterica. This suggests that the genetic exchange, required for random genetic drift, occurs at normal frequency in the regions away from the junctions, although the sequences would only be present in strains carrying rfb regions of the closely related groups A, B, C2, D and E.

It is not easy to put specific times to the recombination events involved, but if we are correct in our interpretation of the gradient of similarity, then there has been time for substantially more substitution (up
to 50%) than has occurred in average genes (about 15–20%) since Salmonella and Escherichia diverged. We have then recombination events in the divergence of the two \(rfb\) gene clusters under discussion which predate divergence between \(E.\ coli\) and \(S.\ enterica\), although both forms appear to have been transferred to \(S.\ enterica\) since that date. This supports our earlier suggestion that the recombination events involved in the divergence of the \(rfb\) gene clusters of group B and D may have occurred in the donor low \(G + C\) species (\cite{Verma:1989, Jiang:1991}). The difference between groups B and D lies entirely within the segment absent in group E1 (Figure 5). Thus although we have compared group E1 with group B, the recombination events could equally well have involved the group D \(rfb\) gene cluster. We now have three \(rfb\) regions with substantial regions of homology, which are thought to have diverged in a low \(G + C\) species before lateral transfer to \(S.\ enterica\). In the case of groups D and B, the second transfer could have involved the homologous DNA at the two ends of the gene clusters (\cite{Verma:1989}). However the presence of \(orf17.4\), unique to group E1, at the end of the gene cluster requires a more complex set of events unless the location of \(orf17.4\) is secondary.

Since many synonymous changes involve base 3 of the codon, the \(G + C\) content of base 3 can reflect the mutational pressure more directly. The corrected \(G + C\) contents for base 3 (P3) of the M32 genes located in regions with gradient of similarity (\(orf7.0\) (M32) and \(orf11.9\) (M32)) are slightly lower than those of their corresponding genes in strain LT2 [\(orf7.0\) (LT2) and \(orf16.5\) (LT2)] (Table 2). Looking at the synonymous substitutions, we find in 491 codons, 81 sites with A or T in strain M32 but G or C in strain LT2; and 61 sites with G or C in strain M32 but A or T in strain LT2. We expect an increase over time in \(G + C\) content with P3 approaching 0.5, characteristic of \(S.\ enterica\) and \(E.\ coli\): the data thus suggest that the \(rfb\) cluster of strain LT2 has been in \(S.\ enterica\) for longer than that of strain M32.

**Summary:** Our general conclusions on the evolution of the \(rfb\) clusters of group B and group E1 \(S.\ enterica\) are shown in Figure 8. It appears that they arose from a common ancestral cluster in a low \(G + C\) content organism (Figure 8), and that either one or both of these clusters had the central region replaced by recombination to give forms A and B. The homologous regions of the two forms could undergo recombination among different \(rfb\) gene clusters such that this part evolved by random genetic drift, with a level of neutral variation that is normal for bacteria. However the two nonhomologous forms of the central regions have persisted, presumably reflecting selection for each form in different hosts, and hence each must drift independently. This independent drift must also apply to those parts of the homologous regions immediately adjacent to the nonhomologous central regions, to give the observed gradient of similarity. It should be noted that such a gradient was not
observed at the junctions in the nonhomologous region from positions 17,288-18,697 of strain M32 or the junctions found when group D and group B rfb gene clusters were compared (Verma, Quigley and Reeves 1988; Verma and Reeves 1989), perhaps because the recombination event in these regions were more recent. Both rfb gene clusters (forms A and C in Figure 8) must have been acquired by S. enterica by lateral transfer. We can not say which form represents the rfb gene cluster of group B or E1, but the one of group B was perhaps transferred into Salmonella first.

The O poly saccharide gene is adjacent to the rfb cluster in strain M32 but separated from it in strain LT2. The rfc genes must have coevolved with their rfb genes which determine the nature of the oligosaccharide which is polymerized, and the association of rfc and rfb gene cluster seems more natural, but is not considered in Figure 8.

Similar models of lateral transfer have been proposed for the generation of diversity of capsular polysaccharides in E. coli (Boulnois and Jann 1989) and in Haemophilus influenzae (Kroll and Moxon 1990). It is interesting that the two major polymorphisms in Gram-negative bacteria may have much in common in the origins of the variation.

This work was supported by a program grant from the Australian Research Council. L.W. is a recipient of a Faculty of Agriculture scholarship of the University of Sydney. S. enterica strain M32 (serovar anatum) was generously provided by the Institute of Medical and Veterinary Science, Adelaide (IMVS).

LITERATURE CITED


Devereux, J., P. Haeb erli and O. Smithies, 1984 A comprehen-


Neilson, K., T. S. Whittam and R. K. Selander,


Communicating editor: D. E. Berg