Rhs Elements of *Escherichia coli* K-12: Complex Composites of Shared and Unique Components That Have Different Evolutionary Histories

SHENG ZHAO,[†] CAROL H. SANDT, GREGORY FEULNER,[‡] DONALD A. VLAZNY,
JANE ANTAL GRAY, AND CHARLES W. HILL*

Department of Biological Chemistry, Milton S. Hershey Medical Center,
Pennsylvania State University, Hershey, Pennsylvania 17033

Received 9 November 1992/Accepted 7 March 1993

The complete sequences of the *RhsB* and *RhsC* elements of *Escherichia coli* K-12 have been determined. These sequence data reveal a new repeated sequence, called H-rpt (Hinc repeat), which is distinct from the *Rhs* core repetition that is found in all five *Rhs* elements. H-rpt is found in *RhsB*, *RhsC*, and *RhsE*. Characterization of H-rpt supports the view that the *Rhs* elements are composite structures assembled from components with very different evolutionary histories and that their incorporation into the *E. coli* genome is relatively recent. In each case, H-rpt is found downstream from the *Rhs* core and is separated from the core by a segment of DNA that is unique to the individual element. The H-rpt's of *RhsB* and *RhsE* are very similar, diverging by only 2.1%. They are 1,291 bp in length, and each contains an 1,134-bp open reading frame (ORF). *RhsC* has three tandem copies of H-rpt, all of which appear defective in that they have large deletions and/or have the reading frame interrupted. Features of H-rpt are analogous to features typical of insertion sequences; however, no associated transposition activity has been detected. A 291-bp fragment of H-rpt is found near min 5 of the *E. coli* K-12 map and is not associated with any *Rhs* core homology. The complete core sequences of *RhsB* and *RhsC* have been compared with that of *RhsA*. As anticipated, the three core sequences are closely related, all having identical lengths of 3,714 bp each. Like *RhsA*, the *RhsB* and *RhsC* cores constitute single ORFs that begin with the first core base. In each case, the core ORF extends beyond the core into the unique sequence. Of the three cores, *RhsB* and *RhsA* are the most similar, showing only 0.9% sequence divergence, while *RhsB* and *RhsC* are the least similar, diverging by 2.9%. All three cores conserve the 28 repetitions of a peptide motif noted originally for *RhsA*. A secondary structure is proposed for this motif, and the possibility of its having an extracellular binding function is discussed. *RhsB* contains one additional unique ORF, and *RhsC* contains two additional unique ORFs. One of these ORFs includes a signal peptide that is functional when fused to *TnphoA*.

The *Rhs* elements were discovered as the locations of a large repetition, the *Rhs* core, that provided homology for a RecA-dependent rearrangement of the *Escherichia coli* K-12 chromosome. The rearrangement resulted from unequal crossover between the *RhsA* and *RhsB* cores and occurred at a spontaneous frequency of about 10^{-5} (4, 17). Upon cloning of these two *Rhs* elements, a probe derived from one of the cores was used to identify the other three. The five *Rhs* elements collectively constitute nearly 1% of the *E. coli* K-12 chromosome. They are widely scattered through the chromosome as follows: *RhsD*, min 12; *RhsC*, min 16; *RhsE*, min 32; *RhsB*, min 77; and *RhsA*, min 81. All five elements are in the clockwise orientation. They have been classified as accessory genetic elements, because they are present in many but not all wild-type *E. coli* isolates (9, 17). In addition to the core, each *Rhs* contains DNA unique to the particular element, with the total size being much larger than that (3.7 kb) of the core. The largest element, *RhsC*, is 9.6 kb in length (9). The elements fall into two subfamilies on the basis of the degree of core sequence similarity. The *RhsA* core, representative of the *RhsABC* subfamily, diverges by 22% from the *RhsD* core, representative of the *RhsDE* subfamily (25).

The *RhsA* and *RhsD* cores are also distinguished by two small nonhomologous sequences. Despite their evolutionary divergence, the *RhsA* and *RhsD* cores share some striking common structural features. Each maintains a single open reading frame (ORF) throughout its entire length, with the start codon coinciding with the first base of the homologous sequence. The respective core ORFs then extend 139 and 177 codons beyond the ends of the homologous sequences into the unique sequences. Thus, they appear capable of producing large proteins (>156 kDa), with very similar 141-kDa amino-terminal portions and divergent carboxy termini. The amino acid sequences predicted for both cores contain a peptide motif that is repeated a total of 28 times.

A novel problem concerning the definition of the boundaries of the *Rhs* elements arises because the ends of the different elements are not homologous and there is no apparent structural relationship between the left and right ends of a given element. Because of the absence of structural landmarks at the *Rhs* boundaries, we have based our definition on a comparison of strains with and without core homology. By our definition, an *Rhs* element consists of all of the DNA which is (i) contiguous with the core homologous sequence and (ii) present in an *Rhs*⁺ strain but absent in a reference *Rhs*⁰ strain. The boundaries of *RhsA*, *RhsC*, and *RhsD* have been defined in this manner (9, 25).

The sequence of the entire *RhsA* element (9) and an extensive sequence analysis of *RhsD* and *RhsE* (25) have been reported, along with the cloning and partial character-

* Corresponding author.

[†] Present address: University of California at San Francisco, San Francisco, CA 94143-0984.

[‡] Present address: Pediatric Lipid Research, Johns Hopkins Hospital, Baltimore, MD 21205.

FIG. 1. Structures of the *RhsB* and *RhsC* elements of *E. coli* K-12. The limits of the *Rhs* elements are indicated by dashed vertical lines. The core homologous sequences and H-rpt homologous sequences (stippled and hatched boxes, respectively) are shown. Open box, a sequence shared by *RhsB* and *RhsA*. The remainder of each element is indicated by a thick line. ORFs are indicated by the arrows below the maps. The positions of the four blocks of repeated peptide motifs in the core ORF are shown as m-1, m-2, m-3, and m-4. Coordinates are in kilobases. The origins of probes used to detect the regions flanking *RhsB* are designated B-us and B-ds. Restriction sites: Ac, *AccI*; B, *Bam*HI; Bg, *Bgl*II; G, *Bgl*II; Hc, *Hinc*II; K, *Kpn*I; Nc, *Nco*I; P, *Pst*I; R, *Eco*RI; Rv, *Eco*RV; V, *Pvu*II. The *Taq*I site at the left of the *RhsB* sequence is shown also.

ization of *RhsB* and *RhsC* (9, 17, 24). Those results showed that the cores of *RhsB* and *RhsC* are much more closely related to *RhsA* than to *RhsD*. The major objective of this article is an analysis of the complete sequences of *RhsB* and *RhsC*. Particular questions addressed concern the degree of similarity between the respective cores, the details of the unique sequences that distinguish the elements, and the nature of a new, large homologous sequence called H-rpt (Hinc repeat), shared by *RhsB*, *RhsC*, and *RhsE*.

MATERIALS AND METHODS

Bacterial strains. The *E. coli* K-12 strains used for recombinant DNA work were described previously (9, 17, 24). NG556, a *recA* mutant that bears both pOX38 and pNG18 (Tn903 Kan^r Amp^r), was supplied by Nigel Grindley. CH1662 is an Amp^s segregant of NG556 that retains pOX38. The *TnphoA* test strains, CC118 and CC202, were obtained from Colin Manoil. The wild-type *E. coli* ECOR 32 and ECOR 55 were supplied by Robert Selander. Plasmid p5-2 was supplied by Christopher Elvin. λ 4A11 was supplied by William Hendrickson.

Plasmids and cloning techniques. The plasmids containing the *RhsA* (pRL276), *RhsB* (pRL389), and *RhsC* (pJG1626) core homologous sequences have been described elsewhere (17, 24). The sequences were subcloned primarily into pUC19 (21), by using standard techniques. To obtain reference DNA for the definition of *RhsB*, genomic DNA of the wild-type *E. coli* ECOR 32 was digested with *Asp* 718 and

*Bgl*III and ligated with pUC19, which had been previously digested with *Asp* 718 and *Bam*HI. The resultant library was screened for homology (9) by using probes designated B-us and B-ds (Fig. 1). A positive clone, pGF3540, was isolated.

Use of *TnphoA* to detect exported proteins. The procedures of Manoil and Beckwith (19) were used. A series of plasmids containing all or portions of *RhsA* and *RhsC* were introduced into the *TnphoA* test strain, CC202. Blue survivors were selected on high-kanamycin agar plates containing 5-bromo-4-chloro-3-indolylphosphate. This selection yields transposition mutants in which *TnphoA* has inserted in frame into a plasmid sequence encoding a protein with a signal peptide (19). pJG1748 was a *TnphoA* derivative of pJG1626, while pSZ3903 was a *TnphoA* derivative of pJZ3702. The transposition mutation present in pJG1748 was transferred to the chromosome through a strategy that relied on the inability of a pBR-derived plasmid to replicate autonomously in a *polA* mutant background (12, 15). CH1330 (*polA*) was transformed with pJG1748, and an Amp^r Kan^r transformant was selected. The transformant had a copy of pJG1748 integrated into the chromosome via a crossover between the plasmid insert and the homologous chromosomal region. The generalized-transduction phage P1 was grown on this mutant and used to transduce a *PolA*⁺ recipient. Kan^r transductants were screened for Amp^s; the presumption was that these would be mutants without vector DNA, which had the chromosomal *RhsC* element replaced by the mutant one. The location of the *TnphoA* insertion in chromosomal *RhsC* was verified both by P1 cotransduction and by Southern analysis.

Genomic DNA extraction and analysis. The techniques for genomic DNA isolation, restriction, and Southern analysis that were used were described previously (3). The *Rhs* core probes were the 0.86-kb *EcoRV-PvuII* fragment derived from the *RhsA* core (pJG1772) and the adjacent *AccI* fragments (combined length, 0.72 kb) from the *RhsD* core (pJG1861), while the H-rpt probe was the 1.4-kb *HincII-NcoI* fragment derived from the H-rpt cluster of *RhsC* (the pSZ204 insert).

Plasmid isolation, restriction enzyme analysis, and sequencing. Procedures for the isolation of plasmid DNA, the production of overlapping deletions, DNA sequencing, and sequence analysis have been described previously (9), with the following exception. In certain cases involving the sequences of large DNA repetitions such as the *Rhs* cores, the sequence of one representation was determined unambiguously by sequencing both strands independently. Other representations were then sequenced by doing side-by-side comparisons on the same gel with the original. This permitted unambiguous comparisons of mismatches between the repetitions. In these cases, the second strand was not always sequenced. Oligonucleotide primers necessary for sequencing were purchased from the Hershey Medical Center Macromolecular Core Facility. Sequence analysis and prediction of protein secondary structure was aided by use of the MacMol program (Soft Gene Berlin). The BLAST search program (1) provided by GenBank was used to search data bases for nucleic acid and amino acid sequence homology.

Mapping and cloning of the secondary H-rpt homologous sequence. A probe derived from the *RhsB* H-rpt was applied to Southern blots of *E. coli* K-12 genomic DNA representing all single and double digests (28 total) of the following seven enzymes: *Bam*HI, *Hind*III, *Eco*RI, *Bgl*II, *Asp* 718, *Pst*I, and *Pvu*II. The *Rhs*-associated homologies gave strong hybridization and were identified on the basis of their known sequences and positions on the Kohara physical map (16). The size was determined for each less intensely hybridizing, unidentified fragment, and an unambiguous restriction map for the immediate neighborhood of the secondary H-rpt homologous sequence was derived. A scaled transparency of this map was used to locate by visual inspection a uniquely close match between it and a graphic representation of the physical map. The approximately 6.2-kb *Bam*HI fragment containing both the secondary H-rpt homologous sequence and the carboxy-terminal portion of *mutD* was cloned from the Kohara λ phage 4A11 (miniset 126).

Nucleotide sequence accession numbers. DNA sequences have been submitted to GenBank with the following accession numbers: *E. coli* K-12 *RhsB* (ECORHSBAA), L02370; K-12 *RhsC* (ECORHSCA), L02373; K-12 *RhsE* H-rpt (ECORHSEA), L02372; and ECOR 32 *RhsB*⁰ (ECORHSBAB), L02371. The *RhsC* sequence begins at an *Eco*RI site which is the same as the site that initiates the ECOKDPABC file of GenBank. This identity was verified by sequencing across the site. The distal portion of the *RhsC* sequence from the *AccI* site through the distal *Pvu*II site overlaps the ECOPHRORF file. When ECORHSCA is merged with ECOKDPABC and ECOPHRORF, the three provide 17,071 bp of contiguous sequence at min 16.

RESULTS

Complete sequence of *RhsB* and *RhsC*. In previous articles, we described the cloning of the *E. coli* K-12 *RhsB* and *RhsC* cores (24), and in the case of *RhsC*, we presented evidence

that the entire *RhsC* element was included within the insert of plasmid pJG1626 (9). The original *RhsB* clone, pRL389, terminated at a *Hind*III site 2.7 kb beyond the end of the core homology (Fig. 1). As we extended our analysis, it became apparent that the right limit of *RhsB* was possibly beyond this *Hind*III site, and a source of this additional DNA was needed. The *pit* locus is linked to *RhsB* (17), and a clone of *pit* (8) appeared to overlap our clones. We obtained the *pit* clone, p5-2, and verified that its left terminus was the *Bam*HI site shown for *RhsB* in Fig. 1. From p5-2, we subcloned a 0.5-kb *Hind*III-*Bgl*II fragment, pGF3536, and this procedure provided material that was contiguous with the pRL389 insert.

The *RhsB* and *RhsC* regions have now been sequenced completely. This effort extends the limited sequencing reported previously (9, 24). The positions of the core homologous sequences, along with significant ORFs and other features, are shown in Fig. 1. The *RhsC* sequence began at an *Eco*RI site 0.7 kb upstream of the *RhsC* core and extended 10,484 bp through the distal *Pvu*II site. The boundaries of the *RhsC* element itself were determined previously (9), and they occur at the positions indicated in Fig. 1. The complete sequence revealed *RhsC* to be 9,570 bp long. Our sequence of the *RhsB* region began at a *Taq* I site 0.7 kb upstream of the *RhsB* core and extended 7,673 bp to a *Bgl*II site. The boundaries of *RhsB* had not been determined previously, and this issue is addressed in the next section.

Definition of the boundaries of *RhsB*. As discussed above, our definition of an *Rhs* element is the DNA covering the core homologous sequence which is absent in a reference *Rhs*⁰ wild-type strain. Application of this definition to *E. coli* K-12 *RhsB* required isolation of the corresponding chromosomal segment from a reference strain. The limits of K-12 *RhsB* were roughly determined by testing for homology between K-12 DNA segments and genomic DNAs from wild-type *RhsB*⁰ strains from the ECOR collection (22). When used as probes, both the 0.6-kb segment commencing at the *Taq*I site upstream of the *RhsB* core (designated B-us in Fig. 1) and the 0.5-kb *Hind*III-*Bgl*II segment (designated B-ds) gave positive signals to blots of the DNAs of such strains, indicating that, by our definition, these *E. coli* K-12 DNA segments were at least in part outside *RhsB*. The reference strain used for cloning the empty *RhsB* site was ECOR 32. Preliminary work showed that ECOR 32 does not have *RhsB*, although it does have *Rhs* elements at other chromosomal locations. A plasmid library of ECOR 32 DNA was prepared and screened for homology with the B-us and B-ds probes (see Materials and Methods). A positive clone was isolated, and from it a 3.8-kb *Hpa*I fragment was subcloned. This subclone, pGF3541, retained homology to both the B-us probe and the B-ds probe. pGF3541 was used for sequence determination and comparison with the K-12 *RhsB* sequence (Fig. 2). Starting from the left, the sequences of K-12 and ECOR 32 were nearly identical. However, 532 bp from the K-12 *Taq*I site, the sequences diverged completely. The sequence similarity resumed 34 bp beyond the *Hind*III site of K-12. Consequently, the K-12 *RhsB* element is 6,687 bp long. In its stead, ECOR 32 carries an unrelated 807-bp segment. Interestingly, the positions of divergence between K-12 and ECOR 32 are very close to stop codons that terminate ORFs converging on K-12 *RhsB*. The positions of these stop codons are shown in Fig. 2.

Homologies shared by the *Rhs* elements. As established by restriction analysis and partial sequencing (9), *RhsB* and *RhsC* have core homologies that are very similar to *RhsA*. Completion of the sequences showed that all three cores are

FIG. 2. Comparative sequences at the boundaries of *E. coli* K-12 *RhsB*. The homologous sequences of K-12 and ECOR 32 are aligned, and identities (:) and mismatches (*) are indicated. The positions of divergence and convergence are separated by 6,687 bp in K-12 and 807 bp in ECOR 32. The positions of stop codons terminating external ORFs that converge on *RhsB* are underlined. Short inverted repetitions are shown by arrows above the sequences.

exactly 3,714 bp in length. On the basis of pairwise comparisons among the three, *RhsB* and *RhsA* are the most similar, differing at only 33 of 3,714 nucleotides (0.9%), while *RhsC* diverges from *RhsA* at 99 positions (2.7%) and from *RhsB* at 108 positions (2.9%). Most significantly, the reading frame initiated at the first base of the core is open throughout the entire core in *RhsB* and *RhsC*, as was found for *RhsA*. These core ORFs then extend into unique sequences. The 173-codon extension of *RhsB* and the 159-codon extension of *RhsC* were described previously (9). A repetition of the last 1,020 bp of the core homology was found in *RhsC*, and its position is shown in Fig. 1. Similar repetitions of the terminal portion of the core were also found in *RhsA* and *RhsD*, but those repetitions were much smaller than the one in *RhsC*. The core repetition of *RhsC* is followed by a 161-codon extension of its ORF. Its sequence shows no homology to any of the other core extensions.

Inspection of the comparative sequences of *RhsB* and *RhsC* revealed another distinct homology that has not been described in detail. A 1,291-bp sequence in *RhsB* was partially or wholly repeated three times in tandem in *RhsC*. This homologous sequence has been named H-rpt (Hinc repeat) for the presence of a characteristic *HincII* site. H-rpt homology is also found in *RhsE* (25). The positions of the H-rpt units are shown in Fig. 1, and their sequences, along with a consensus sequence derived from the five occurrences, are aligned in Fig. 3. A large (1,134-bp) ORF-H is present in the H-rpt's of *RhsB* and *RhsE*.

Upon observing that H-rpt homology is associated with three of strain K-12 *Rhs* elements, we wished to determine whether it occurs at other chromosomal locations. This was tested by Southern analysis by successively using probes specific for the *Rhs* core subfamilies and H-rpt. Both *SalI* and a combination of *SalI* and *HindIII* were used to digest *E. coli* K-12 DNA (Fig. 4). The H-rpt probe strongly hybridized to only three fragments in each case, and those fragments corresponded in size to the fragments expected for *RhsB*, *RhsC*, and *RhsE*. An additional weakly hybridizing band was

observed. The source of this signal was located between *mutD* and *pepD* at min 5 on the *E. coli* map (2). Specifically, the last 291 nucleotides of an H-rpt (Fig. 3) lie in the clockwise orientation beginning 1.9 kb to the right of the *EcoRI* site which serves as the right endpoint of the published sequence of *mutD* (6) (see Materials and Methods).

Absence of H-rpt transposition activity. Certain features of H-rpt might suggest that it is capable of transposition or some other type of site-specific recombination. For example, four of the five H-rpt homologous sequences begin at exactly the same base (base 1 in Fig. 3) and four of the five end at the same base (base 1291). The occurrence of a repeated sequence abutted against several different sequences is most easily explained by a site-specific recombination mechanism. In addition, the first and last 11 bp of the H-rpt consensus sequence are related as inverted repeats ("i" in Fig. 3). Similar inverted repetitions are often found at the ends of insertion sequences (10). We attempted to detect H-rpt transposition activity through a replicon fusion test (13). The source of H-rpt was a pUC19 derivative, pJG1645, that contained the 2.8-kb *BamHI-HindIII* segment of *RhsB* (Fig. 1) inserted into the vector *SmaI* site. This plasmid was transformed into a *recA* recipient, CH1662, that carried the F⁺ derivative, pOX38. pOX38 has had the resident insertion sequences of F⁺ deleted. Grindley and Joyce (13) showed that pOX38 could transfer the drug resistance of a nonconjugative plasmid to a Rec⁻ recipient if the plasmid contained a sequence capable of causing cointegration. They found that Tn903 could do this at a frequency of 10⁻⁶ to 10⁻⁷. We performed this test with pJG1645 in a manner that would have detected cointegrants if they had formed at a frequency of 10⁻⁹. We detected two fusions between pJG1645 and pOX38; however, neither seemed to involve relevant sequences of the *Rhs* element. Instead, they involved one of the residual hot spots for illegitimate recombination on pOX38 that was identified by O'Connor and Malamy (23). Thus, no transposition activity could be detected by this test.

FIG. 3. Sequence of H-rpt homology. H-rpt units from *RhsB*, *RhsC*, and *RhsE*, along with the weakly hybridizing homologous sequence at min 5, are aligned with a consensus (cons.) sequence derived by inspection of the set. Also included is the 20 bp preceding and following each representation. Six hundred base pairs are omitted for brevity. A match with the consensus is indicated by a dot. ○, absence of any base. Nucleotide substitutions that do not change the derived amino acid sequence are shown by lowercase letters. Potential start codons, in-frame stop codons, and ribosome binding sites (Shine-Dalgarno sequences [S-D]) in the forward sense are singly underlined, while those in the reverse sense are doubly underlined. Inverted repetitions i and ii are indicated (>>>).

FIG. 4. *Rhs* core and H-rpt homologies in *E. coli* K-12. *E. coli* K-12 DNA was digested with *SalI* (lanes a to c) or a combination of *SalI* and *HindIII* (lanes d to f). After agarose gel electrophoresis and Southern transfer, the paper was treated successively with probes specific for *RhsA* core (lanes a and d), *RhsD* core (lanes b and e), and H-rpt (lanes c and f). Bands in lanes a to c are identified on the left; bands in lanes d to f are identified on the right. The letters A to E and min 5 to the sides of the gel identify the *Rhs* element corresponding to each band. Note that the *RhsD* core probe gives significant cross-hybridization with the *RhsABC* subfamily fragments and that *HindIII* separates the core and H-rpt sequences of *RhsE*.

A coding sequence in *RhsC* produces a signal peptide capable of mediating export. It was of interest to determine whether the *Rhs* elements produce proteins that are either membrane proteins or proteins that are transported into the periplasmic space. The *TnphoA* transposon can be used to screen for such coding sequences on multicopy plasmids (19), and we used it to test *RhsA* and *RhsC*. The pBR325 derivatives, pRL276 (*RhsA*) and pJG1626 (*RhsC*) were transformed into the *TnphoA* test strain, CC202, and Kan^r PhoA⁺

derivatives were selected (see Materials and Methods). In both cases, many transpositions into the vector β -lactamase gene were found, as expected. In addition, transpositions of *TnphoA* into sequences derived from *RhsC*, but not from *RhsA*, were found. All of the transpositions into *RhsC* went somewhere into the *Bam*HI-*Hinc*II interval (6.3 to 7.7 kb in Fig. 1). The transposition test was repeated on an *RhsC* subclone, pJZ3702, which contains the *Pvu*II-*Hinc*II (6.1- to 7.7-kb) fragment. The two transpositions observed occurred at 6.8 and 7.1 kb. The transposition found in pSZ3903 was at 6.8 kb, and the exact position of the insertion was determined by sequencing. The *PhoA* coding sequence was found to be inserted at base 6837 in a 189-amino-acid ORF, ORF-C3. It was inserted in frame, 17 codons from its amino terminus (Fig. 5). The sequence of the first 19 amino acids of this coding sequence conforms nicely to expectations for a signal peptide (31).

The color development in colonies of these *TnphoA* transposition mutants clearly indicated that there was significant expression of ORF-C3 when the ORF was present on the multicopy plasmids. This observation did not prove that the chromosomal copy of ORF-C3 is expressed constitutively. To check this, we constructed a mutant in which the ORF-C3 of the chromosomal *RhsC* element was replaced by the version containing *TnphoA* (see Materials and Methods). Colonies of this mutant, CH3763, produced no color even upon prolonged incubation of the plates. We conclude that there is little or no expression of the ORF-C3 coding sequence when the ORF is present as a single copy at its normal location in the chromosome.

DISCUSSION

The *Rhs* elements were defined initially on the basis of the long core homology present in each (17). We now report a distinct homology, H-rpt, associated with some but not all of the *Rhs* elements. H-rpt was found once in both *RhsB* and *RhsE*, and at least a portion of it was found three times in *RhsC* (Fig. 1). H-rpt boundaries are clearly defined in that several representations of the homology begin and end at

FIG. 5. Sequences of the regions encoding the ends of the core extensions and the beginnings of the unique internal ORFs. In each example, the amino acid sequence of the core extension is shown below the nucleotide sequence, while the unique ORF is shown above. The presumed ribosome binding site (Shine-Dalgarno sequence [S-D]) and start codons for each of the unique ORFs are underlined. Stop codons terminating the core extensions are indicated (***). The position of the *TnphoA* insertion into the *RhsC* ORF-C3 is indicated, and its probable signal peptide is doubly underlined. Coordinates are as indicated in the legend to Fig. 1.

coincident bases (Fig. 3). The copies in *RhsB* and *RhsE* are 1,291 bp in length, and they contain an 1,134-bp ORF, called ORF-H. ORF-H terminates 11 bp before the end of the homology. The nucleotide sequences of ORF-H in *RhsB* and *RhsE* differ at only 21 positions, 10 of which cause a change in coding. The three copies present in *RhsC* all seem defective. The first has only the first 381 bp of the homology, the second has had its ORF interrupted by a termination signal after the 17th codon, and the third has a 313-bp deletion (Fig. 3). In addition, a base substitution in the third copy terminates the ORF just before the deletion. Whether any of the H-rpt ORFs are expressed under conditions of routine laboratory cultivation is being investigated, but we do note in Fig. 3 sequences which conform closely to the *E. coli* -10 and -35 consensus promoter. A search of data bases for proteins homologous to the predicted product of ORF-H yielded two significant matches. One match was a hypothetical protein encoded by a *Xanthobacter autotrophicus* plasmid (14). This ORF is linked to the *dhlA* (halokane dehalogenase) gene (accession no. M26950), and its predicted amino acid sequence is 37% identical to ORF-H over the first 286 codons of ORF-H. The other match was three closely linked ORFs which precede the *Vibrio cholerae* *rfbT* locus (accession no. X59554), which is implicated in serotype determination (28). These ORFs, *rfbQ*, *rfbR*, and *rfbS*, occur in frame with spacings of only 6 bases between *rfbQ* and *rfbR* and 12 bases between *rfbR* and *rfbS*. In other words, what appears to be three peptides in *V. cholerae* is a single larger peptide in the *E. coli* H-rpt. The overall similarity is high, with the three *V. cholerae* peptides collectively showing 51% amino acid identity with ORF-H. The *V. cholerae* and *X. autotrophicus* sequences show 33% identity. Since no function for any of these hypothetical proteins is known, they provide little insight into H-rpt function.

Several circumstances suggest that H-rpt is an insertion sequence. These include its multiple occurrence as a discrete homologous sequence, its size, which is comparable to other insertion sequences, and the presence of inverted repetitions at its termini (10). In addition, there is a 7-bp inverted repetition covering the putative ribosome binding site of ORF-H ("ii" in Fig. 3). Similar arrangements are present in IS10 and IS50, in which they have been shown to prevent the expression of the transposase from read-through transcripts (7, 26). Unlike many insertion sequences, however, no repetition of flanking host sequence is apparent for any of the copies. In an effort to test whether H-rpt behaves as an insertion sequence, we sought evidence of transposition activity associated with the copy in *RhsB*. As described in Results, we were unable to observe the participation of H-rpt in replicon fusion down to a frequency of 10^{-9} . Despite our negative results, it remains difficult to explain the presence of multiple copies of a precise sequence like H-rpt without invoking the activity of a site-specific recombination system.

The physical association of H-rpt with the *Rhs* core is provocative, although the significance is obscure. An H-rpt-specific probe produced four signals in K-12 genomic DNA (Fig. 4). Three of the signals were strong and correspond to *RhsB*, *RhsC*, and *RhsE*. The origin of the weak signal maps to min 5. It is a highly conserved fragment of H-rpt, consisting of its last 291 bp (Fig. 3). It is interesting that there is a 79-codon ORF (ORF-Hrev in Fig. 1) that overlaps ORF-H in the reverse direction. Essentially all of ORF-Hrev is retained in this fragment at min 5, suggesting the possibility that provision of the ORF-Hrev product was once its function. Whatever the original situation, ORF-Hrev of the

min 5 fragment now terminates after only 41 codons because of a base substitution.

The *Rhs* elements seem to be composites assembled from components with very different evolutionary histories. This idea is based on two considerations. The first is that homologous components of different elements show widely disparate degrees of divergence. The second is that different components differ radically in their GC contents. For example, the H-rpt units of *RhsB* and *RhsE* diverge at only 27 of 1,291 positions (2.1%). This divergence is much smaller than that observed for the core sequences of *RhsB* and *RhsE*, which differ at 323 of 1,529 positions (21%). This complexity is increased by consideration of the AT-rich but seemingly featureless distal portion of *RhsA*. It contains a 234-bp segment (bp 8522 to 8755 of ECORHSA) that is 76% similar to a portion of the H-rpt consensus sequence (bp 428 to 662 in Fig. 3). Possibly, *RhsA* once had an H-rpt, but it is now both highly divergent and degenerate. Preliminary sequence data show that an analogously divergent and degenerate H-rpt also occurs in *RhsE* upstream from the conserved H-rpt. This 1,060-bp segment of *RhsE* is 62% similar to the ORF-H consensus sequence (13a). H-rpt is AT rich, having a GC content of only 40%. This composition contrasts with the composition of the *Rhs* core, which is GC rich (62%). Both values differ markedly from the *E. coli* norm of 50%, in keeping with the idea that both components have exotic origins and that they entered *E. coli* populations long after the divergence of *E. coli* as a species. The difference in their compositions further suggests that the original sources of the core and H-rpt homologies were themselves very different.

RhsB and *RhsC* both have unique internal ORFs. These are designated ORF-B2, ORF-C2, and ORF-C3 in Fig. 1. Each lies immediately downstream from an extension of a core ORF. The sequences are highly organized in that the stop codon of the core extension overlaps the initiation codon of the unique ORF (Fig. 5). For ORF-C2 and ORF-C3, the initiation codons share bases with the stop codon of the preceding ORF. For ORF-B2, the coding sequences overlap by 26 bp and occur in different reading frames. A similar situation was observed for *RhsA* (9), in which ORF-A2 follows the core extension by 21 bp. All of these internal ORFs are very AT rich, with ORF-C2 being extreme at 31.5% GC (Table 1). All have very low *E. coli* codon adaptation indices (CAI) (27) (Table 1). (A CAI of 0.26 to 0.3 is typical of *E. coli* proteins expressed at low levels, and none of the genes analyzed by Sharp and Li [27] had indices lower than 0.21.) Low CAI values and low GC content are also characteristic of the core extensions (Table 1). This further supports the hypothesis of an exotic origin of the elements (20, 27). The CAI values of the GC-rich cores are about 0.33, much higher than those of the other *Rhs* components. One of the unique internal ORFs, ORF-C3, served as a target for *TnphoA* and, therefore, provided a signal sufficient to get alkaline phosphatase into the periplasm. This suggests that the ORF-C3 product is exported into or through the cytoplasmic membrane (19). Its sequence predicts a 189-amino-acid protein with an excellent signal peptide (Fig. 5). Replacement of the chromosomal ORF-C3 by the ORF-C3::TnphoA construct resulted in a reduction of alkaline phosphatase expression to undetectable levels, implying that the chromosomal ORF-C3 itself is not expressed under the conditions of laboratory culture. The *Rhs* core ORFs are also expressed at very low, if not negligible, levels (28a).

Comparison of *RhsB* with *RhsA* revealed one additional homology. *RhsB* contains a 399-bp segment (open box in

TABLE 1. Base compositions and CAI of ORFs from the *Rhs* elements and from flanking chromosomal regions

Element	ORF ^a	Size (bp)	GC content (%)	CAI ^b
<i>RhsA</i>	Core	3,714	61.7	0.332
	Core extension	417	40.5	0.157
	A2	840	33.7	0.196
	A3	459	32.5	0.176
	Aus	>531	52.7	0.289
<i>RhsB</i>	Core	3,714	61.9	0.331
	Core extension	519	37.4	0.189
	B2	381	33.6	0.216
	Bus	399	57.4	0.380
	Bds	>450	57.3	0.287
<i>RhsC</i>	Core	3,714	61.6	0.332
	Core extension	477	40.7	0.185
	C2	324	31.5	0.110
	Core repeat extension	483	42.9	0.155
	C3	567	36.9	0.185
	Cus	204	50.5	0.371
	ORF169	507	50.5	0.287
H-rpt	H	1,134	41.6	0.221
	Hrev	237	43.6	0.216

^a The ORFs associated with *RhsB* and *RhsC* are identified in Fig. 1. The *RhsA*-associated ORFs were described previously (9) and are included for comparison. The Aus and Bds ORFs have been only partially sequenced.

^b The Sharp and Li CAI (27) values were calculated by using the DNA Strider program (Christian Marck, Gif-Sur-Yvette, France).

Fig. 1) that is similar to the proximal portion of ORF-A2 of *RhsA*. The homologous sequences differ by 11 of the 399 bp. Whereas ORF-A2 encodes 280 amino acids, the residual ORF in *RhsB* is terminated after 71 codons. This is the only example in which even a portion of one of the internal ORFs appears in a different element.

Comparison of the *E. coli* K-12 chromosome to that of ECOR 32 allowed a definition of the *RhsB* boundaries. The left boundary is 200 bp upstream of the core. This distance between the left boundary and core start is similar to the 190 bp found for *RhsA* and the 189 bp found for *RhsC*, despite the dissimilarity between the upstream *RhsB* sequence and that of *RhsA* or *RhsC*. Of the *Rhs* elements of K-12, only *RhsA* and *RhsC* are highly homologous in this segment preceding the core; the corresponding *RhsB* and *RhsD* sequences have limited homology with each other and, to an even lesser extent, with *RhsA* and *RhsC*. This diversity is intriguing, because the promoter for core expression is presumably in this segment.

The *Rhs* elements have not shown a simple insertion-deletion relationship when *Rhs*⁺ strains are compared with *Rhs*⁰ strains. They generally replace a much smaller sequence in the reference strain. For *RhsC*, the sequence replaced is only 10 bp (9). For *RhsA*, the replaced sequence is 32 bp (9), and for *RhsD* it is 224 bp (25). The situation for *RhsB* was found to be more complex. First, the alternative sequence in ECOR 32 is larger, containing 807 bp. Second, other strains have different and much larger alternative sequences at this location. The *Rhs*⁰ strain ECOR 55 has an alternative sequence which is larger than 3.4 kb. This alternative sequence has not been cloned in its entirety, so we were not able to delimit its ends relative to K-12. It shows no homology to either K-12 or ECOR 32 genomic DNA by

Southern analysis (9a). The 807-bp substitution present in ECOR 32 was tested for homology to GenBank sequences as well as to the *Rhs* sequences. No significant homologies were found. It is moderately AT rich, having a GC content of 41%. The longest internal ORF is 330 bp, but this ORF is not preceded by a good ribosome binding site.

RhsB lies between two ORFs of unknown function that are oppositely oriented, converging on the element and terminating within a few nucleotides of its boundaries (Fig. 2). Both are somewhat GC rich at 57%. ORF-Bus has a CAI of 0.380 (Table 1), suggesting significant expression. Transcripts covering these ORFs would extend into *RhsB*. However, short inverted repeats are present near both ends of *RhsB* (Fig. 2), and we suggest that they cause rho-dependent termination, thus protecting the element from transcription from external promoters. *RhsC* is also closely linked to flanking ORFs (Fig. 1). ORF-Cus terminates 55 bp upstream of *RhsC*, and there is a good rho-independent transcription terminator between it and the *RhsC* boundary. Transcription of ORF169 has been studied by Lorence et al. (18). Its start codon is 40 bp to the right of *RhsC*. The GC content of each of these ORFs is about 50%, and ORF-Cus has a relatively high CAI of 0.371 (Table 1).

Comparisons of the *RhsB* and *RhsC* cores with each other and with *RhsA* revealed relationships with implications for the evolution of the elements. In Fig. 6, we have tabulated the number of mismatches in each 100-bp interval of the core. Synonymous substitutions are shown as open bars, and substitutions affecting the amino acid sequence are shown as solid bars. The mismatches between *RhsA* and *RhsB* are clustered at the ends of the core, with most of them occurring toward the 3' end. Of the 33 mismatches, 26 occur in the last 16% of the sequence. In fact, there are only three mismatches in the 3,026-bp interval beginning at nucleotide 82. Two explanations seem plausible. *RhsA* and *RhsB* may have diverged a very long time ago, as reflected by the 4.3% divergence in the last 606 bp of the core, but a relatively recent gene conversion has made the bulk of the respective cores nearly identical. This explanation is supported by the fact that recombination between the *RhsA* and *RhsB* cores has been shown to occur (4, 17). On the other hand, the two elements may have diverged very recently, and the great divergence at the 3' end could reflect an exchange between one of them and a more divergent variant. Such an exchange might have not only altered the 3' end of the core but also brought in the unique core extension currently associated with that element.

The *RhsC* core shows a different relationship with *RhsA*. Overall, there are 99 mismatches, 41 of which predict amino acid substitution. Many of the substitutions occur through the interior of the core. Evidently, these cores diverged a rather long time ago and gene conversion has not caused homogenization of their sequences.

The *RhsD* core belongs to a core subfamily that is roughly 22% divergent from *RhsA* (25). In addition to the simple sequence divergence, there are two small nonhomologous sequences that distinguish *RhsA* and *RhsD*; a 15-codon sequence in *RhsD* replaces an unrelated 10-codon sequence in *RhsA*, and *RhsD* has what can be considered a 7-codon insertion relative to *RhsA*. This raises the possibility that these regions might be more variable than other portions of the cores; however, *RhsB* and *RhsC* are identical to *RhsA* at these positions.

The partial core repetitions such as the one present in *RhsC* (Fig. 1) are intriguing. A rather obvious idea is that they, along with their unique extensions, were once part of a

FIG. 6. Distribution of base substitutions in pairwise comparisons of the *Rhs* cores. The 3,714-bp core sequences were divided into 36 100-bp intervals, and the last 114 bp was included in the 37th interval. The *Rhs* cores were compared in pairs, and the number of mismatches for each interval was counted. Synonymous changes are indicated by open bars and nonsynonymous changes are indicated by solid bars. (a) Comparison of *RhsA* and *RhsB*; (b) comparison of *RhsA* and *RhsC*; (c) comparison of *RhsB* and *RhsC*.

primary core but they have been displaced by different versions. We can look for clues to the origin of a core repetition in its sequence. The one in *RhsC* is 1,020 bp long. Its reading frame has not been interrupted by a termination codon. In terms of sequence divergence, it bears roughly the same relationship to the primary cores of *RhsA* (2.0%) and *RhsC* (2.4%) but is somewhat more distant from *RhsB* (3.1%). The absence of stop codons in the remaining core ORF despite this degree of mutational drift suggests that it was functional through most of its evolutionary history and became nonfunctional relatively recently.

A distinctive feature of the *Rhs* core is the presence of a peptide motif which is repeated 28 times (9). The motif has the sequence XXGX₍₁₋₃₎XXRYXYDXXGRL(I or T)XXXX. The peptide motif repeats occur in four blocks (Fig. 1). Block m-1 contains 3 representations (amino acids 328 to 393), block m-2 contains 16 representations (amino acids 416

FIG. 7. Secondary structure of the repeated core peptide motif. Secondary structure was predicted (11) for each amino acid position in 14 consecutive occurrences of the peptide motif in the *RhsA* core, starting at amino acid 416 and ending at 710 (9). For each position within the motif, all 14 decision values were averaged and plotted. ●, β -sheet; ○, β -turn.

to 755), block m-3 contains 8 representations (amino acids 806 to 1005), and block m-4 contains only one representation (amino acids 1160 to 1181). This arrangement is conserved not only in the *RhsABC* subfamily but also in *RhsD*. As the function of a core protein is unknown, we can only speculate as to the significance of the motif. One possibility is that the core protein is a binding protein, either outside the cell or on the cell surface. Peptide repetitions in such proteins have been observed. In a potentially relevant example, a set of bacterial proteins that bind carbohydrate moieties share a tandemly repeated motif (30). This motif includes a sequence, YYFD, bracketed by conserved glycines, a circumstance that somewhat resembles the *Rhs* motif. Application of an algorithm (11) for predicting protein secondary structure suggests that each repetition of the *Rhs* motif contains two β -sheet- β -turn regions. We have calculated the decision value for each position within 14 motif repetitions (selected from block m-2). The values for each position were averaged and plotted, with the result shown in Fig. 7. If this secondary structure prediction is correct, the sections of β -sheet could fold back on each other within a single motif and successive motifs could stack in a very regular array. Successive β -sheet- β -turn structures are found in bacterial porins (5, 29), although no repetition of primary sequence has been noted. Interestingly, each putative segment of β -sheet in the *Rhs* motif contains 7 amino acids, thought to be the minimum number required to cross a membrane. We are attempting to determine the locations of the *Rhs* core proteins.

The results reported in this article, taken together with previous results, paint a picture of extreme complexity with regard to *Rhs* structure and evolution. Portions of the same element can have very different GC contents (e.g., cores versus internal ORFs). Pairs of elements can have distinct homologies which, in turn, show dramatically different degrees of divergence (e.g., cores versus H-rpt's). Some as yet undiscovered mechanism must generate many different core extensions and attach them to the 3' ends of the cores. Thus far, for *E. coli* K-12, we have identified nine distinct extensions, and only two of them (the extensions of the primary cores of *RhsD* and *RhsB*) show any homology beyond the first nine codons (9). We are searching for evidence of movement and/or alteration of these extensions, using wild-type *E. coli* from the ECOR collection.

ACKNOWLEDGMENTS

We thank Shani Williams for her contributions to the characterization of H-rpt.

This work was supported by Public Health Service grant GM16329 from the National Institutes of Health.

REFERENCES

- Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. Basic local alignment search tool. *J. Mol. Biol.* **215**:403–410.
- Bachmann, B. J. 1990. Linkage map of *Escherichia coli* K-12, edition 8. *Microbiol. Rev.* **54**:130–197.
- Brody, H., A. Greener, and C. W. Hill. 1985. Excision and reintegration of the *Escherichia coli* K-12 chromosomal element *e14*. *J. Bacteriol.* **161**:1112–1117.
- Capage, M., and C. W. Hill. 1979. Preferential unequal recombination in the *glyS* region of the *Escherichia coli* chromosome. *J. Mol. Biol.* **127**:73–87.
- Cowan, S. W., T. Schirmer, G. Rummel, M. Steiert, R. Ghosh, R. A. Paupit, J. N. Jansonius, and J. P. Rosenbusch. 1992. Crystal structures explain functional properties of two *E. coli* porins. *Nature (London)* **358**:727–733.
- Cox, E. C., and D. L. Horner. 1986. DNA sequence and coding properties of *mutD* (*dnaQ*) a dominant *Escherichia coli* mutator gene. *J. Mol. Biol.* **190**:113–117.
- Davis, M. A., R. W. Simons, and N. Kleckner. 1985. *Tn10* protects itself at two levels from fortuitous activation by external promoters. *Cell* **43**:379–387.
- Elvin, C. M., N. E. Dixon, and H. Rosenberg. 1986. Molecular cloning of the phosphate (inorganic) transport (*pit*) gene of *Escherichia coli* K12. *Mol. Gen. Genet.* **204**:477–484.
- Feulner, G., J. A. Gray, J. A. Kirschman, A. F. Lehner, A. B. Sadosky, D. A. Vlazny, J. Zhang, S. Zhao, and C. W. Hill. 1990. Structure of the *rhsA* locus from *Escherichia coli* K-12 and comparison of *rhsA* with other members of the *rhs* multigene family. *J. Bacteriol.* **172**:446–456.
- Feulner, G., C. H. Sandt, and C. W. Hill. Unpublished data.
- Galas, D. J., and M. Chandler. 1989. Bacterial insertion sequences, p. 109–162. In D. E. Berg and M. M. Howe (ed.), *Mobile DNA*. American Society for Microbiology, Washington, D.C.
- Garnier, J., D. J. Osguthorpe, and B. Robson. 1978. Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *J. Mol. Biol.* **120**:97–120.
- Greener, A., and C. W. Hill. 1980. Identification of a novel genetic element in *Escherichia coli* K-12. *J. Bacteriol.* **144**:312–321.
- Grindley, N. D. F., and C. M. Joyce. 1980. Genetic and DNA sequence analysis of the kanamycin resistance transposon *Tn903*. *Proc. Natl. Acad. Sci. USA* **77**:7176–7180.
- Hill, C. W., and J. A. Gray. Unpublished data.
- Janssen, D. B., F. Pries, J. van der Ploeg, B. Kazemier, P. Terpstra, and B. Witholt. 1989. Cloning of 1,2-dichloroethane degradation genes of *Xanthobacter autotrophicus* GJ10 and expression and sequencing of the *dhlA* gene. *J. Bacteriol.* **171**:6791–6799.
- Kingsbury, D. T., and D. R. Helinski. 1970. DNA polymerase as a requirement for the maintenance of the bacterial plasmid colicinogenic factor E1. *Biochem. Biophys. Res. Commun.* **41**:1538–1544.
- Kohara, Y., K. Akiyama, and K. Isono. 1987. The physical map of the whole *E. coli* chromosome: application of a new strategy for rapid analysis and sorting of a large genomic library. *Cell* **50**:495–508.
- Lin, R.-J., M. Capage, and C. W. Hill. 1984. A repetitive DNA sequence, *rhs*, responsible for duplications within the *Escherichia coli* K-12 chromosome. *J. Mol. Biol.* **177**:1–18.
- Lorence, M. C., S. D. Maika, and C. S. Rupert. 1990. Physical analysis of *phr* gene transcription in *Escherichia coli* K-12. *J. Bacteriol.* **172**:6551–6556.
- Manoil, C., and J. Beckwith. 1985. *TnpH*: a transposon probe for protein export signals. *Proc. Natl. Acad. Sci. USA* **82**:8129–8133.
- Médigue, C., T. Rouxel, P. Vigier, A. Hénaut, and A. Danchin. 1991. Evidence for horizontal gene transfer in *Escherichia coli* speciation. *J. Mol. Biol.* **222**:851–856.
- Norlander, J., T. Kempe, and J. Messing. 1983. Construction of improved M13 vectors using oligodeoxynucleotide-directed mutagenesis. *Gene* **26**:101–106.
- Ochman, H., and R. K. Selander. 1984. Standard reference strains of *Escherichia coli* from natural populations. *J. Bacteriol.* **157**:690–693.
- O'Connor, M. B., and M. H. Malamy. 1984. Role of the *F* factor *oriV1* region in *recA*-independent illegitimate recombination. Stable replicon fusions of the *F* derivative pOX38 and pBR322-related plasmids. *J. Mol. Biol.* **175**:263–284.
- Sadosky, A. B., A. Davidson, R.-J. Lin, and C. W. Hill. 1989. *rhs* gene family of *Escherichia coli* K-12. *J. Bacteriol.* **171**:636–642.
- Sadosky, A. B., J. A. Gray, and C. W. Hill. 1991. The *RhsD-E* subfamily of *Escherichia coli* K-12. *Nucleic Acids Res.* **19**:7177–7183.
- Schulz, V. P., and W. S. Reznikoff. 1991. Translation initiation of IS50R read-through transcripts. *J. Mol. Biol.* **221**:65–80.
- Sharp, P. M., and W.-H. Li. 1987. The codon adaptation index—a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.* **15**:1281–1295.
- Strocher, U. H., L. E. Karageorgos, R. Morona, and P. A. Manning. 1992. Serotype conversion in *Vibrio cholerae* O1. *Proc. Natl. Acad. Sci. USA* **89**:2566–2570.
- Vlazny, D. A., and C. W. Hill. Unpublished data.
- Weiss, M. S., T. Wacker, J. Weckesser, W. Welte, and G. E. Schulz. 1990. The three-dimensional structure of porin from *Rhodobacter capsulatus* at 3 Å resolution. *FEBS Lett.* **267**:268–272.
- von Eichel-Streiber, C., M. Sauerborn, and H. K. Kuramitsu. 1992. Evidence for a modular structure of the homologous repetitive C-terminal carbohydrate-binding sites of *Clostridium difficile* toxins and *Streptococcus mutans* glucosyltransferases. *J. Bacteriol.* **174**:6707–6710.
- von Heijne, G. 1985. Signal sequences: the limits of variation. *J. Mol. Biol.* **184**:99–105.