JOURNAL OF BACTERIOLOGY, Aug. 1998, p. 4102–4110 Vol. 180, No. 16
0021-9193/98/$04.00+0

Rhs Elements Comprise Three Subfamilies Which Diverged Prior to Acquisition by *Escherichia coli*

YONG-DONG WANG, SHENG ZHAO,† AND CHARLES W. HILL*

Department of Biochemistry and Molecular Biology, Pennsylvania State College of Medicine, Hershey, Pennsylvania 17033

Received 28 January 1998/Accepted 2 June 1998

The *Rhs* elements are complex genetic composites widely spread among *Escherichia coli* isolates. One of their components, a 3.7-kb, GC-rich core, maintains a single open reading frame that extends the full length of the core and then 400 to 600 bp beyond into an AT-rich region. Whereas *Rhs* cores are homologous, core extensions from different elements are dissimilar. Two new *Rhs* elements from strains of the ECOR reference collection have been characterized. *RhsG* (from strain ECOR-11) maps to min 5.3, and *RhsH* (from strain ECOR-45) maps to min 32.8, where it lies in tandem with *RhsE*. Comparison of strain K-12 to ECOR-11 indicates that *RhsG* was once present in but has been largely deleted from an ancestor of K-12. Phylogenetic analysis shows that the cores from eight known elements fall into three subfamilies, *RhsA-B-C-F*, *RhsD-E*, and *RhsG-H*. Cores from different subfamilies diverge 22 to 29%. Analysis of substitutions that distinguish between subfamilies shows that the origin of the ancestral core as well as the process of subfamily separation occurred in a GC-rich background. Furthermore, each subfamily independently passed from the GC-rich background to a less GC-rich background such as *E. coli*. A new example of core-extension shuffling provides the first example of exchange between cores of different subfamilies. A novel component of *RhsE* and *RhsG*, *vgr*, encodes a large protein distinguished by 18 to 19 repetitions of a Val-Gly dipeptide occurring with a eight-residue periodicity.

The *Escherichia coli* genome is commonly viewed as a mosaic derived from an ancient antecedent genome to which various genetic elements have been added through horizontal transfer. The potential for rapid evolution through such horizontal transfer has been long recognized (4, 14). *Rhs* elements are unusual among accessory elements not only in details of their structural organization but also in their distribution through the *E. coli* population (8). *Rhs* element distribution correlates closely with the *E. coli* population structure as defined by multilocus enzyme electrophoresis analysis (MLEE) of the ECOR reference collection (7), and this contrasts with observations of other accessory elements whose presence tends to vary widely through populations. Five *Rhs* elements are present in *E. coli* K-12 and make up 0.8% of its chromosome (9); other *E. coli* strains have as many as seven (8). However, *Rhs* elements are not present in all *E. coli* strains, and they are absent from *Salmonella enterica* LT2 (11).

Rhs elements are designated according to their chromosomal locations (9). The locations of six, *RhsA* through *RhsF*, were determined in previous work (9, 20) (Fig. 1A). A prototypical structure of an *Rhs* element is shown in Fig. 1B, which emphasizes their composite nature. A feature common to all is a 3.7-kb, GC-rich core. This core maintains a single open reading frame (ORF) that extends the full length of the core and then beyond the core limit into an AT-rich region called the core extension. To date, 10 distinct core extensions have been described (6, 16, 20, 21). The core extensions add as many as 159 codons to the core ORF. Each core ORF is immediately followed or overlapped by a shorter ORF, called the downstream ORF (dsORF). The various dsORFs are also AT rich. The dsORFs are generally not similar to each other in sequence. The N termini of most dsORFs appear to encode a signal peptide. The core extensions and linked dsORFs appear to have coevolved (9), and the DNA segment encoding them is designated ext/ds in Fig. 1B. The ext/ds regions can be shuffled between *Rhs* elements, with the consequence that the structure of a given element can vary from strain to strain. Another feature common to *Rhs* elements is the presence of one or more insertion sequences (ISs) positioned to the right of the dsORF. The most frequently observed IS is one called the H-rpt (21), although IS*1* homologs also have been found (20). Often the ISs appear defective due to deletion or other mutations. The *Rhs* core is preceded by a leader sequence that contains the presumed promoter for core ORF expression. The leader regions of different *Rhs* elements are largely dissimilar in sequence. The core ORFs are not expressed to a detectable extent during routine cultivation, at least in strain K-12 (9). The function(s) of the putative products of the *Rhs* core ORFs is unknown. However, the predicted core protein sequence is similar in several ways to the sequence of a *Bacillus subtilis* wall-associated protein, leading to speculation that the *Rhs* products are associated with the cell surface and have a binding function (9).

Rhs cores fall into distinct subfamilies based on sequence divergence (16). Cores comprising the *RhsA-B-C-F* subfamily differ by about 22% at the nucleotide level from those of the *RhsD-E* subfamily, while cores within the two subfamilies diverge less than 10%. In the absence of sequence data, Southern analysis can be used to assign elements to a particular core subfamily, since probes from one subfamily give strong signals for other cores from the same subfamily but much weaker signals for cores from a different subfamily. A survey of the ECOR reference collection of *E. coli* gave preliminary evidence that some strains contain additional *Rhs* elements that could not be assigned to the six known elements (8).

This report describes two new *Rhs* elements, *RhsG* and *RhsH*, that comprise a new core subfamily which is distinct

* Corresponding author. Department of Biochemistry and Molecular Biology, Pennsylvania State College of Medicine, Mail Services H171, Hershey, PA 17033-0850. Phone: (717) 531-8592. Fax: (717) 531-7072. E-mail: chill@psu.edu.

† Present address: Department of Biomathematics, M. D. Anderson Cancer Center, Houston, TX 77030.

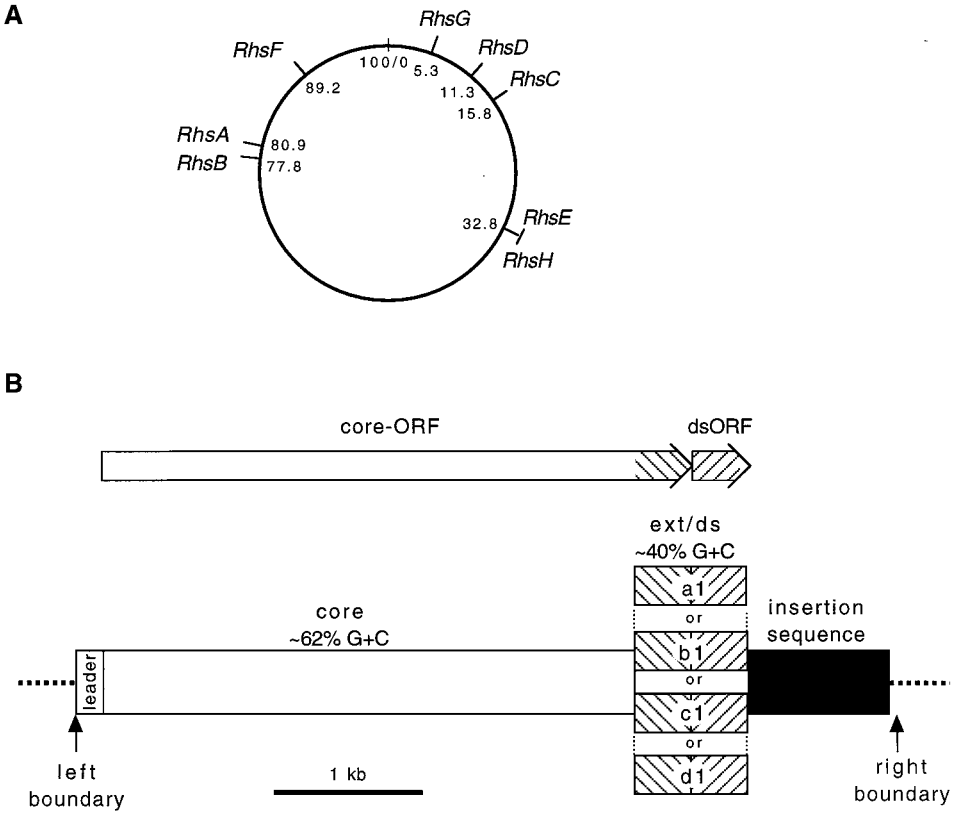

FIG. 1. *Rhs* element structure and location. (A) Locations are given with respect to the *E. coli* K-12 map. The locations of *RhsA* through *RhsF* are known from previous work (9, 20), while those of *RhsG* and *RhsH* are documented here. *RhsF* and *RhsH* are both absent from strain K-12, while portions of *RhsE* and *RhsG* appear to have been deleted in the original K-12 (see Discussion). (B) Structural components typical of an *Rhs* element. All actual elements have more complex structures. The core ORF is comprised of the core and an adjacent core extension. The respective core extensions and their linked dsORFs appear to have coevolved, and the DNA segments encoding them are designated ext/ds; individual ext/ds segments are designated a1, b1, etc.

from the *RhsA-B-C-F* and *RhsD-E* subfamilies. We present evidence based on sequence analysis that the three core subfamilies diverged in a GC-rich genome. Transfer into *E. coli* occurred after subfamily divergence, and consequently at least three different horizontal transfer events must have taken place. We also describe four new *Rhs* core extensions and document that a core extension can be shuffled between elements belonging to different core subfamilies; heretofore, examples of core-extension shuffling had been restricted to within subfamilies.

MATERIALS AND METHODS

Bacterial strains and molecular techniques. Sources of *E. coli* strains have been described previously (8). Cloning protocols were generally as described before (21). Primary clones obtained are identified in Table 1. Procedures for DNA extraction and Southern analysis have been described elsewhere (3).

DNA sequencing. Manual DNA sequencing was performed as described elsewhere (3). Oligonucleotide synthesis and automated DNA sequencing were performed by the Macromolecular Core Facility of the Penn State College of Medicine. For all new sequences, both strands were sequenced; for repetitive sequencing of homologous segments from different sources, sequencing from only one strand was sometimes used.

Nucleotide sequence accession numbers. The GenBank accession numbers for the sequences reported are AF044499 through AF044505.

RESULTS

A new *Rhs* core subfamily. A previous survey of the ECOR reference collection indicated that some *E. coli* strains possessed *Rhs* elements that could not be assigned to any of the six known elements (8). To characterize these putative new elements and extend our understanding of *Rhs* structure in general, we cloned portions of six different *Rhs* elements from the genomes of selected ECOR strains. The cloned segments were characterized by restriction analysis and sequencing; the structures deduced are summarized in Fig. 2. The cores of three elements (*RhsE* of ECOR-50, *RhsG* of ECOR-11, and *RhsH* of ECOR-45) were sequenced in their entirety, and partial se-

TABLE 1. Primary genomic clones

Plasmid[a]	Source	Insert identification[b]
pYW6561	ECOR-11 *RhsG*	*Hin*dIII (−26) to *Hin*dIII (−10)
pYW6517	ECOR-11 *RhsG*	*Mlu*I (−22) to *Mlu*I (3.2)
pYW6551	ECOR-11 *RhsG*	*Hin*dIII (−10) to *Hin*dIII (4.8)
pYW6571	ECOR-11 *RhsG*	*Mlu*I (3.2) to *Mlu*I (8.5)
pYW6582	ECOR-45 *RhsG*	*Mlu*I (−22) to *Mlu*I (3.2)
pYW6584	ECOR-45 *RhsG*	*Mlu*I (3.2) to *Mlu*I (5.1)
pYW6581	ECOR-45 *RhsG*	*Mlu*I (5.1) to *Mlu*I (9.6)
pSZ3950	ECOR-50 *RhsG*	*Bgl*II (−5.2) to *Bgl*II (1.5)
pSZ3885	ECOR-50 *RhsE*	*Eco*RI (−3.5) to *Bam*HI (−1.0)
pSZ3870	ECOR-50 *RhsE*	*Sal*I (−2.3) to *Hin*dIII (3.2)
pSZ3854	ECOR-50 *RhsE*	*Hin*dIII (3.2) to *Nsi*I (6.0)
pYW6521	ECOR-45 *RhsE/RhsH*	*Mlu*I (−5.6) to *Mlu*I (3.2) of *RhsE*[c]
pYW6651	ECOR-45 *RhsE/RhsH*	*Mlu*I (−3.9) to *Mlu*I (6.5) of *RhsH*[c]
pYW2103	ECOR-45 *RhsF*	*Sma*I (−0.2) to *Mlu*I (6.4) of *RhsF*

[a] pSZ and pYW plasmids use pUC19 and pUC21 (18), respectively, as vectors.
[b] Insert coordinates in kilobases, where base 0 is the 5′ end of the core (see Fig. 2).
[c] The *Mlu*I site at coordinate 3.2 of *RhsE* is identical to the *Mlu*I site at coordinate −3.9 of *RhsH*.

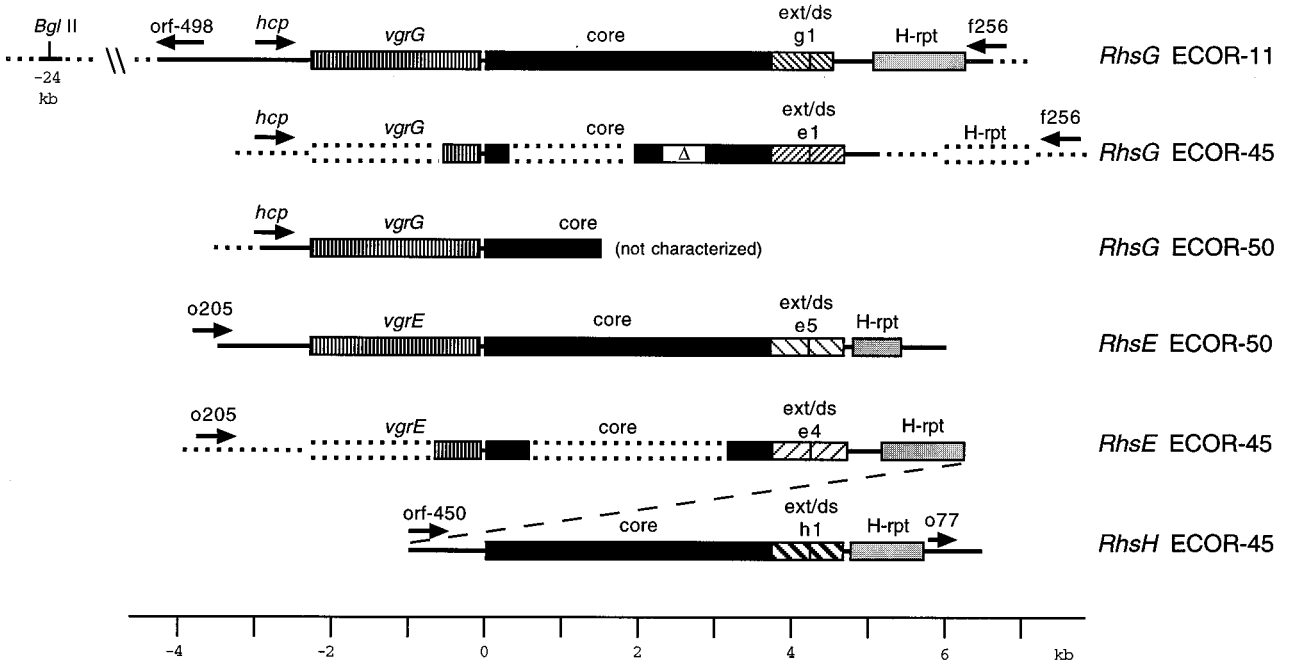

FIG. 2. Structures of new *Rhs* elements. Each element's designation and *E. coli* strain of origin are shown to the right. The regions specified by solid lines or filled boxes have been sequenced. Regions indicated by dotted lines have been characterized by restriction mapping and in some cases hybridization. Unique DNA sequences from the *E. coli* K-12 chromosomal framework are designated ORFs f256, o205, and o77 (2). The *Rhs* elements are aligned so that the start codon of the core ORF is base 1. *RhsE* and *RhsH* of ECOR-45 are linked in tandem (indicated by the diagonal dashed line). The core of ECOR-45 *RhsG* contains a 587-bp deletion, indicated by Δ.

quences were obtained for the other three (*RhsE* and *RhsG* of ECOR-45 and *RhsG* of ECOR-50). In many respects, the structures were similar to the prototypical structure shown in Fig. 1B. Each had a GC-rich core sequence similar to the cores of elements previously characterized in terms of size and predicted amino acid sequence. The sizes and G+C contents of the newly sequenced cores were compared to those for *RhsA*, *RhsC*, and *RhsD* of K-12 (Table 2). Each core maintained a single ORF, beginning at its left end but extending beyond the right limit of the core homology.

This effort brought the number of sequenced *Rhs* cores to eight. These eight were subjected to phylogenetic analysis using the PAUP program (17). An inspection of the sequences before this analysis revealed several short segments disrupting the continuity of their alignment. By subtracting these regions, we established for each a 3,693-bp sequence that allowed optimum alignment of all eight cores without interruption. The subtracted regions ranged from 3 to 57 nucleotides, and all were multiples of 3; thus, these adjustments had no effect on reading frames. The modified sequences were then used to prepare a tree (Fig. 3A), which indicated grouping into three core subfamilies. The two new elements, *RhsG* and *RhsH*, establish a new subfamily, *RhsG-H*. The grouping of *RhsE* with *RhsD* agreed with the previous designation of an *RhsD-E* subfamily based on the remnant of *RhsE* found in strain K-12 (16). The *RhsA-B-C-F* subfamily was defined previously (20), although a complete sequence for the *RhsF* core had not been reported previously. Simple pairwise comparisons were done to assess extents of similarity (Table 2). The divergence between cores from different subfamilies ranged from 22 to 29%. Within the subfamilies, *RhsG* and *RhsH* were the most divergent, at 11%.

Maintenance of *Rhs* structural features. In a typical *Rhs* element (Fig. 1B), the core ORF extends beyond the right-hand core limit into a core extension. The core extension, in turn, is immediately followed by a short dsORF. This ext/ds segment is typically AT rich, in contrast to the GC-rich core. To determine whether this arrangement was maintained for the new elements, the region to the right of each core was sequenced. (The relevant portion of ECOR-50 *RhsG* was not cloned.) In agreement with the prototype, each core ORF extended beyond the core boundary for an additional 432 to 480 bp. Each core extension, in turn, was closely followed or overlapped by a short dsORF, and the regions coding these features were all AT rich (Table 3). Each of the five ext/ds regions was quite different from the other four. With one exception, all five ext/ds regions were also different from any other ext/ds region previously sequenced. The single exception was the one found in *RhsG* of ECOR-45. That ext/ds region was nearly identical to ext/ds-e1 of *RhsE* from K-12 (16) (two mismatches in 943 nucleotides).

All *Rhs* elements so far characterized have had one or more

TABLE 2. Comparison of *Rhs* core sequences

Element	Source	Size[a] (bp)	G+C (%)	Sequence identity[b] (%)				
				RhsC	*RhsD*	*RhsE*	*RhsG*	*RhsH*
RhsA	K-12	3,741	61.5	97	78	77	71	72
RhsC	K-12	3,741	61.2		78	77	71	72
RhsD	K-12	3,774	63.4			94	76	77
RhsE	ECOR-50	3,783	64.2				75	78
RhsG	ECOR-11	3,771	63.0					89
RhsH	ECOR-45	3,780	63.2					

[a] A revised definition of the *Rhs* core adds 27 bp to its C terminus (see Discussion).

[b] From pairwise comparisons of a 3,693-bp sequence derived from each core, modified to eliminate all apparent insertions or deletions.

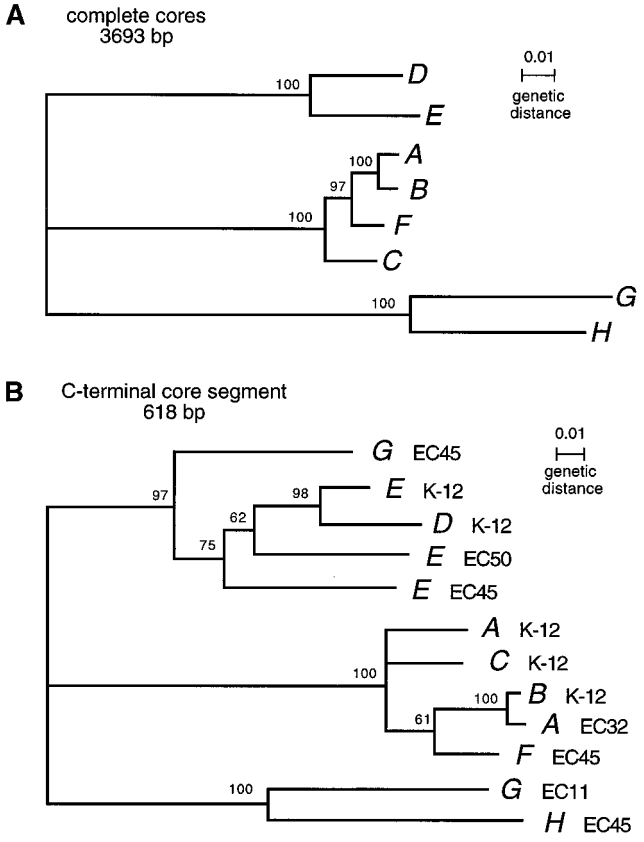

FIG. 3. Phylogeny of *Rhs* core sequences. (A) Tree defining three core subfamilies. The core sequences from each *Rhs* element were aligned, and regions of apparent insertion or deletion were subtracted as described in the text. The tree relating the remaining 3,693-bp sequences was prepared by using the PAUP program for phylogenetic analysis (17). Bootstrap values based on 1,000 trees are indicated at the nodes. (B) Tree relating the C-terminal portions of 12 *Rhs* cores. The last 618 bp of each core was used in the analysis; to adjust for the relative deletion present in *RhsG* of ECOR-11 and *RhsH* of ECOR-45 (encoding residues 1104 to 1108 in Fig. 4), 15 nucleotides were subtracted from the other 10 sequences. Bootstrap values based on 1,000 trees are indicated at the nodes. Grouping of ECOR-45 *RhsG* with the *RhsD-E* subfamily was supported by 100% of the trees.

insertion ISs present as components of the larger composite structure. The most frequently encountered IS has been the H-rpt, which in *RhsB* and *RhsE* of K-12 is 1,291 bp long and contains a 1,134-bp ORF (21). An element belonging to the IS*1* family, iso-IS*1*, is associated with *RhsF* (20). H-rpt homologies were found in association with each of these new elements (Fig. 2). In all cases but one, the H-rpt appeared defective through deletion. The single exception was the H-rpt of ECOR-11 *RhsG*, which retained its full length but nevertheless had its ORF disrupted by two frameshifts. *RhsE* of ECOR-45 contained two other small fragments that had interesting homologs found elsewhere. A small remnant of iso-IS*1* preceded its H-rpt by about 200 bp. It was only 134 bp in length but was 95% similar to a portion of the iso-IS*1* found in ECOR-50 *RhsF*. Between this iso-IS*1* fragment and the H-rpt was a 104-bp sequence related (86% similar) to the IntB intron found in some *E. coli* strains (5).

Another commonly observed feature of *Rhs* elements is the presence of small core fragments, generally located between the primary core and the IS. Two of these elements had such core fragments. *RhsE* of ECOR-50 had a 91-bp fragment apparently derived from the interior of a core, and *RhsH* of ECOR-45 had a 41-bp fragment apparently derived from the 5′ end of a core. They were both immediately adjacent to the left end of the H-rpt in the corresponding elements. These ISs and core fragments presumably are genetic debris left from ancient rearrangements leading to the current *Rhs* structures.

Map locations of *RhsG* and *RhsH*. The identification of an *Rhs* element is based on its chromosomal location. Therefore, *RhsG* and *RhsH* needed to be located with respect to the *E. coli* K-12 genetic map. Comparison of sequence flanking an *Rhs* element with sequence from a reference *E. coli* strain, lacking the element, has been used to locate the element as well as to mark its boundaries. We anticipated that *RhsG* was absent from K-12 and that K-12 would be a suitable reference strain. We expected that sequence to the right and left of ECOR-11 *RhsG*, if extended far enough, would match the K-12 sequence. Immediately to the right of the *RhsG* H-rpt (Fig. 2) was sequence identical to part of ORF f256 (no mismatches in 400 bp), which is located at min 5.3 of the *E. coli* K-12 chromosome (2). This was of particular interest because in K-12, a fragmentary H-rpt had been found precisely at this position (21). This was the only H-rpt homology in K-12 which had not been assigned to an *Rhs* element. In K-12, there is a *Bgl*II site preceding this fragmentary H-rpt by 33 bp. Inspection of the ECOR-11 restriction map revealed a *Bgl*II site 30 kb to the left of the H-rpt (Fig. 2). The sequence adjacent to this site was determined and found to be identical to that of K-12. We conclude that *RhsG* is located at the equivalent of min 5.3 on the K-12 genetic map. The existence of the 291-bp H-rpt remnant at this location in K-12 could be explained if the ancestral genome of both K-12 and ECOR-11 contained *RhsG*. In that case, a large deletion must have occurred in the K-12 lineage, beginning 33 bp to the right of the *Bgl*II site and terminating within the H-rpt. The putative deleted segment was flanked by a 5-bp repeated sequence, GTTCT, which could have had a role in the deletion event.

A similar strategy was used to locate *RhsH* of ECOR-45. ORF o77, which is adjacent to *RhsE* of K-12, was found immediately to the right of *RhsH* (Fig. 2). In other words, *RhsH* of ECOR-45 appeared to occupy the same chromosomal location as *RhsE* of K-12. This observation immediately brought to question the status of *RhsE* in ECOR-45. Examination of the sequence at the left end of the primary *RhsH* clone (pYW6651) revealed the presence of another *Rhs* core. This arrangement suggested that in ECOR-45, *RhsE* and *RhsH* lay in tandem. This was verified by cloning the remainder of ECOR-45 *RhsE* (pYW6521) and by selected sequencing. We found that the *RhsH* core was separated from the H-rpt of *RhsE* by 1,012 bp. This spacer contained a 450-bp ORF and was AT rich (37.9% G+C); it was not similar to any known sequence.

A new *Rhs* component, *vgr*. In ECOR-11, there was a long ORF immediately upstream of the *RhsG* core (Fig. 2). It was 2,139 bp long and slightly (55%) GC rich. Only 75 bp separated the *vgrG* ORF from the *RhsG* core. An essentially identical

TABLE 3. Core extension and dsORF sequences[a]

Source	ext/ds	Core extension		dsORF	
		Size (bp)	G+C (%)	Size (bp)	G+C (%)
ECOR-11 *RhsG*	g1	459	44.2	216	36.5
ECOR-45 *RhsG*	e1	447	40.9	480	27.5
ECOR-45 *RhsH*	h1	480	35.8	447	32.9
ECOR-45 *RhsE*	e4	432	41.7	393	28.0
ECOR-50 *RhsE*	e5	444	39.9	396	34.6

[a] *Rhs* components are identified in Fig. 1B.

relationship of *vgrG* to the core sequence was also found in *RhsG* of ECOR-50. However, the *vgrG* sequences of the two strains were significantly divergent (70 mismatches in 2,139 bp, or 3.3% divergence). The respective 75-bp spacers separating *vgrG* and the *Rhs* core were identical save for a single mismatch.

The close linkage of *vgrG* to the *RhsG* core prompted consideration of whether it should be included as part of the composite element. Decisions for inclusion in the composite are complicated by the fact that criteria for *Rhs* boundaries are themselves problematic: there are no conserved sequences or motifs marking the boundaries (9). Occurrence of a feature linked to the cores of two different *Rhs* elements has been used to justify inclusion. This criterion was satisfied for *vgr*. A *vgr* ORF was also associated with *RhsE* of ECOR-50, where it was separated from the *Rhs* core by only 66 bp (Fig. 2). Consequently, we consider *vgr* to be a component of the *Rhs* composite.

A sequence homologous to *Vibrio cholerae hcp*. There was a 516-bp ORF immediately to the left of the *vgrG* sequence of ECOR-11. Its predicted product was 72% identical to the *V. cholerae* hemolysin-coregulated protein of unknown function, Hcp (19). Hcp is the only *V. cholerae* protein known to be secreted without a signal sequence. The gene for the *E. coli* ECOR-11 Hcp homolog, *hcp*, was separated from *vgrG* by an apparent Rho-independent transcription terminator, AACAGCATCCGGCtatcagGCCGGATGCTGTTttt (capitalized nucleotides form a perfect dyad symmetry). The Hcp homolog was similarly located in *RhsG* of ECOR-50 (Fig. 2). A partial sequence from ECOR-50 revealed that the respective versions diverged at 9 of 464 positions (1.9% divergence). Yet further to the left in ECOR-11 was a 498-bp ORF (ORF-498), encoded by the complementary strand (Fig. 2). ORF-498 and *hcp* were separated by a 694-bp AT-rich (32.6% G+C) segment.

As with *vgr*, we considered whether *hcp* was part of the *RhsG* composite. A DNA probe specific for the Hcp homolog was used to check the 72 strains of the ECOR collection as well as K-12. K-12 and 36 of the ECOR strains gave no signal, but the other 36 ECOR strains were positive, each giving but a single signal. In most cases, this signal could be assigned to a position near *RhsG* analogous to the ECOR-11 and ECOR-50 arrangement (Fig. 2). However, six of the positive strains (ECOR-53, ECOR-59, ECOR-60, ECOR-61, ECOR-62, and ECOR-64) were from ECOR group B2, and these strains have no *Rhs* elements (8). Nevertheless, the location of the Hcp homolog in these strains appeared to be at min 5.3, the same as in ECOR-11. This conclusion was based on the fact that the Hcp signal coincided with the signal from a probe for ORF f256, the region downstream of *RhsG* as defined in ECOR-11 and K-12. Many ECOR strains were positive for the ORF f256 probe, and its source was presumed to be part of the common *E. coli* chromosomal framework. Our conclusion was that the Hcp homolog was probably part of an accessory element but was not part of the *RhsG* composite, since it could exist independently albeit at the same chromosomal location. Although its boundaries cannot be defined precisely as was done for *RhsA* or *RhsC* (6), *RhsG* extends from the transcription terminator that follows *hcp* to a point between the H-rpt and ORF f256 (Fig. 2).

DISCUSSION

We have identified eight *Rhs* elements, each having a distinct chromosomal location (Fig. 1A). The core sequences fall into three subfamilies, *RhsA-B-C-F*, *RhsD-E*, and *RhsG-H* (Fig. 3A). The amino acid sequences predicted for two members of each subfamily are aligned in Fig. 4. This alignment reveals positions where deletions or insertions distinguish the sequences. Some but not all of these interruptions correlate well with the subfamily affiliations. For example, *RhsG* and *RhsH* share a five-codon deletion at residues 1104 to 1108, and *RhsA* and *RhsC* share a seven-codon deletion at residues 845 to 851. Features previously noted for core proteins (16) are conserved despite the fact that roughly 40% of the positions in the peptide sequence are polymorphic in this set of six. For example, all core sequences are rich in hydroxylated amino acids, ranging from 19.8 to 20.8% in their combined content of Ser, Thr, and Tyr. Perhaps the most remarkable feature of the *Rhs* core protein is the presence of a repeated motif that can be written YDxxGRL(I/T). The locations of 36 reasonable matches to this motif are marked in Fig. 4 by large dots over the conserved positions. The motif repetition is particularly regular between coordinates 475 and 741, where Gly residues (marked by daggers) precede the YDxxGRL(I/T) motifs by 7 residues and where a periodicity of 20 to 21 residues is observed. Overall, the number and organization of motif repetitions are strongly conserved among all three subfamilies, indicating that the organization existed in a common progenitor. Previous examination of the *RhsA* and *RhsD* sequences had revealed a potential membrane-spanning region starting at residue 29 (16). *RhsG* and *RhsH* have a three-codon insertion that would introduce two and three lysines at position 42, while the *RhsE* sequence predicts an Arg at position 46. These substitutions would clearly compromise that proposed function. However, *RhsG* and *RhsH* also have Leu replacing Arg at 28. The result is the possible shift of the membrane-spanning domain to residues 20 to 39 (underlined in Fig. 4). Some regions of the core have been constrained through evolution more than others. For example, only 23% of the amino acid positions are polymorphic in the regions of residues 1 to 109 and 916 to 1082. On the other hand, 46% are polymorphic in the region bounded by residues 110 and 721.

The designation of the core C-terminal boundary in Fig. 4 includes nine more amino acids than used previously. In the set of sequences available originally (15), significant nucleotide divergence appeared to begin after a highly conserved proline codon at residue 1259. As more sequences became available, it became clear that the significant transition from conserved to divergent sequence is after residue 1268. Of the nine residues added, the first four show considerable variation among cores, but the last five are highly conserved. This region was formerly referred to as a nine-codon joint (9).

Several strong inferences about *Rhs* core evolution can be made from the sequence data. First, all cores have G+C contents of >61%, a value quite distinct from the *E. coli* average of 51%. Muto and Osawa (12) have observed that for a given bacterial species, the first, second, and third codon positions each have characteristic average G+C contents, each value in turn varying with the overall G+C content of the genome. The expected values for a genome of 51% G+C, such as that of *E. coli*, and for one of 63% G+C were calculated by using equations described by Lawrence and Ochman (10). These theoretical values are listed in Table 4, along with observed values for two *Rhs* cores from each subfamily. At the first and third positions, the observed values all agree well with the expectation for a 63% G+C genome. This finding suggests that the ancestral *Rhs* core evolved in a similarly GC-rich background and that these extant cores have not resided in *E. coli* long enough for the overall G+C contents to have ameliorated substantially. The second positions of the *Rhs* cores are somewhat more GC rich than predicted (48% versus 43.7%). This

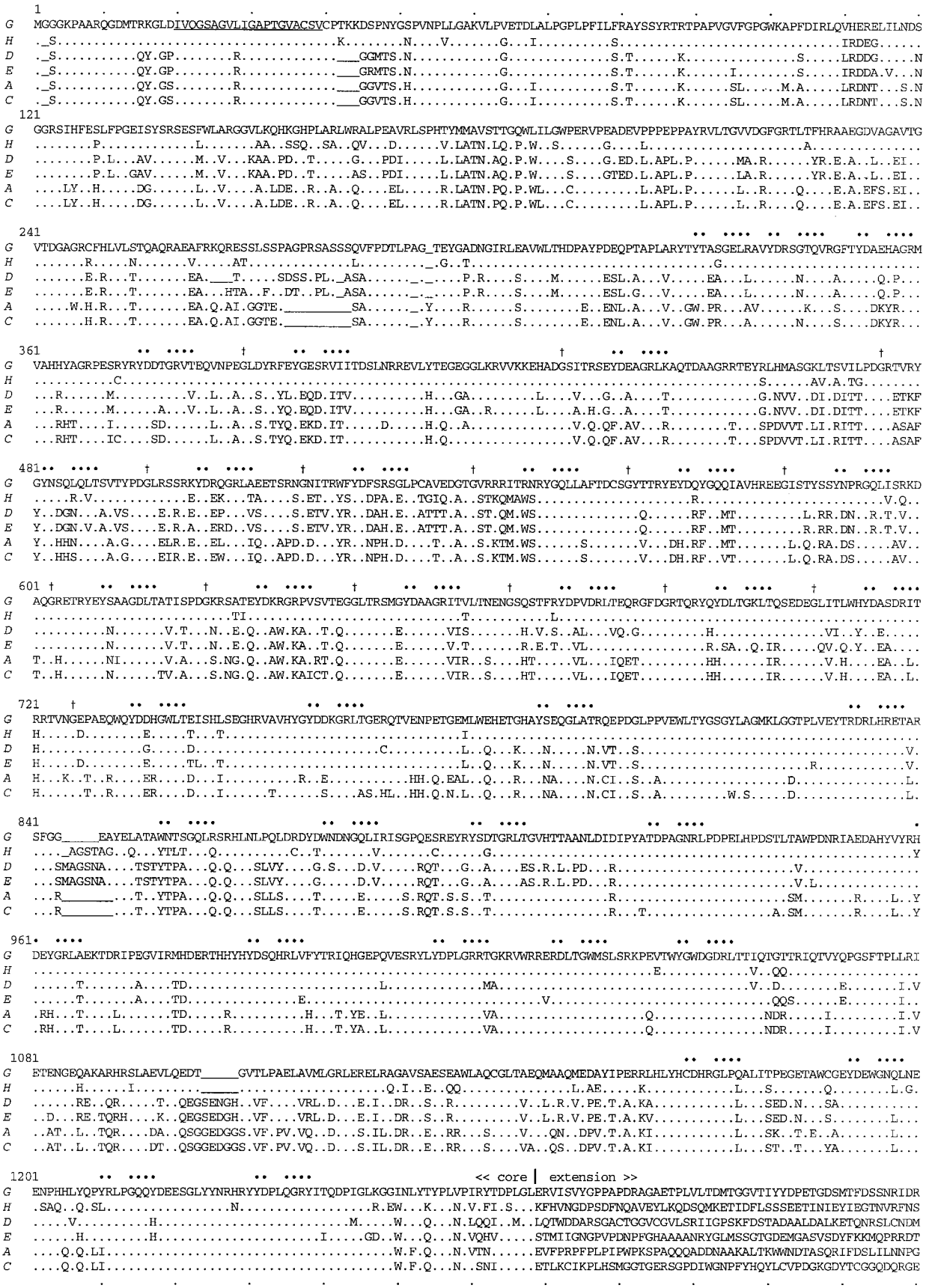

FIG. 4. Alignment of six predicted *Rhs* core proteins. The core sequences are *RhsG* (ECOR-11), *RhsH* (ECOR-45), *RhsD* (K-12), *RhsE* (ECOR-50), *RhsA* (K-12), and *RhsC* (K-12). Also included are the first 52 residues of the corresponding extensions of the core ORFs. Identity with the *RhsG* core sequence (.), a potential membrane-spanning domain for *RhsG* and *RhsH* (underlined), conserved positions of the repeated motif YDxxGRL(I/T) (●), and associated Gly residues (†) are marked.

TABLE 4. Base composition of *Rhs* cores[a]

Source	% G+C			
	Overall	1st position	2nd position	3rd position
Theoretical[b]	63.0	65.6	43.7	74.3
Theoretical	51.0	58.3	40.5	54.0
RhsA	61.5	64.5	47.9	72.1
RhsC	61.2	63.8	48.4	71.5
RhsD	63.4	66.1	48.3	75.9
RhsE	64.2	66.4	48.9	77.2
RhsG	63.0	65.6	48.4	75.0
RhsH	63.2	65.6	48.3	75.9

[a] Core sequences adjusted to 3,693 bp for optimal alignment.
[b] Calculated from the overall G+C content (10).

may be explained, at least in part, by the fact that the cores are quite hydrophilic, and codons with A at the second position all encode hydrophobic amino acids.

A point of considerable interest is whether a single ancestral *Rhs* core entered the *E. coli* species and then diversified into the array now seen or whether there were independent introductions into *E. coli*. The mere fact that there is up to 29% divergence between cores (Table 2) suggests that their divergence preceded the *E. coli-S. enterica* speciation (13). However, examination of informative substitutions gives additional insight, favoring an independent transition of a founder for each subfamily from the GC-rich background to a less GC-rich one. When a gene or element is transferred to a new host genome, the G+C content ameliorates in favor of the G+C content of the new host, with the third codon position changing more rapidly than the other positions (10). The second position ameliorates the slowest. Consequently, the degree of amelioration at each position can be used to assess evolutionary history. For our analysis, six cores, two from each subfamily, were aligned. We identified primary (1°) substitutions as positions at which both examples from one subfamily were identical but differed from the other four which in turn were identical. The minority base was tabulated as to its subfamily association, its position within the codon, and whether it was G or C or, alternatively, A or T (Table 5). Among the three subfamilies, there were a total of 695 sites with 1° substitutions. In addition, positions at which five cores were identical and one differed were identified as secondary (2°) substitutions. These marked divergence within a subfamily. There were a total of 314 2° substitutions.

The history of subfamily divergence is revealed by contrasting the first position to the third position among the 1° substitutions. At the first codon position, 60.0% of 1° substitutions were G/C for the *RhsA-C* subfamily, 71.0% were G/C for the *RhsD-E* subfamily, and 65.9% were G/C for the *RhsG-H* subfamily (Table 5). None of these values are significantly different from the 65.6% theoretical value expected for a genome with a 63% G+C content overall (Table 4). The persistence of high GC bias in this category implies that many of these substitutions occurred in a GC-rich background. In other words, subfamily divergence began in a GC-rich background. In contrast, at the third position, 43.7% of 1° substitutions were G/C for the *RhsA-C* subfamily, 50.7% were G/C for the *RhsD-E* subfamily, and 57.6% were G/C for the *RhsG-H* subfamily. These values are all significantly less than the 74.3% expected for the third codon position of a genome with a 63% overall G+C content (Table 4), and they distinctly indicate amelioration away from GC richness. This finding indicates that even though subfamily divergence may have begun in a GC-rich background, it continued in a less GC-rich background, possibly *E. coli*. Divergence within subfamilies is reflected in 2° substitutions. At both first and third positions, 2° substitutions show general bias toward A/T. This is consistent with their accumulation in a less GC-rich background, possibly *E. coli*. The parsimonious picture is that a common ancestor of the *Rhs* cores evolved in a GC-rich background; in the same (or similar) background, separation of the three subfamilies occurred, and the accumulation of 1° substitutions continued. At least one founder for each subfamily then independently entered a less GC-rich background, possibly *E. coli*. Additional 1° substitutions, now favoring A/T, accumulated to further separate the subfamilies. Finally, divergence within subfamilies began. The only values in Table 5 that do not favor this scheme are the first- and second-position 2° substitutions within the *RhsD-E* subfamily. These results indicate significant bias toward G/C, but they are in fact based on rather small numbers. There is a suggestion that the *RhsA-C* progenitor made the transition from the GC-rich background earlier than the *RhsG-H* progenitor since the third-position 1° substitutions distinguishing the *RhsA-C* subfamily appear significantly more highly biased toward A/T than those distinguishing the *RhsG-H* subfamily.

Multiple *Rhs* elements commonly occur in the same *E. coli* chromosome (8). This circumstance should provide many opportunities for recombination between cores (1). In fact, recombination between certain cores is known to be a common event: *RhsA* and *RhsB* recombine at a frequency of 10^{-5} in

TABLE 5. 1° and 2° core base substitutions[a]

Source	Base substitutions[b]					
	1st position		2nd position		3rd position	
	GC/AT	% G+C (mean ± SD)	GC/AT	% G+C (mean ± SD)	GC/AT	% G+C (mean ± SD)
1° substitutions						
RhsA-C subfamily	39/26	60.0 ± 6.1	34/21	61.8 ± 6.6	73/94	43.7 ± 3.8
RhsD-E subfamily	22/9	71.0 ± 8.2	18/15	56.3 ± 8.7	36/35	50.7 ± 5.9
RhsG-H subfamily	54/28	65.9 ± 5.2	38/36	51.4 ± 5.8	68/50	57.6 ± 4.5
2° substitutions						
RhsA-C subfamily	3/12	20.0 ± 10.3	4/6	40.0 ± 15.5	13/16	44.8 ± 9.2
RhsD-E subfamily	9/4	69.2 ± 12.8	15/10	60.0 ± 9.8	12/32	27.3 ± 6.7
RhsG-H subfamily	19/23	45.2 ± 7.7	25/28	47.2 ± 6.9	29/54	34.9 ± 5.2

[a] 1° substitutions unambiguously separate one core subfamily from the other two; 2° substitutions separate an individual core sequence uniformly from the other five.
[b] 1° and 2° substitutions at each codon position were scored as to whether they were to G or C or to A or T.

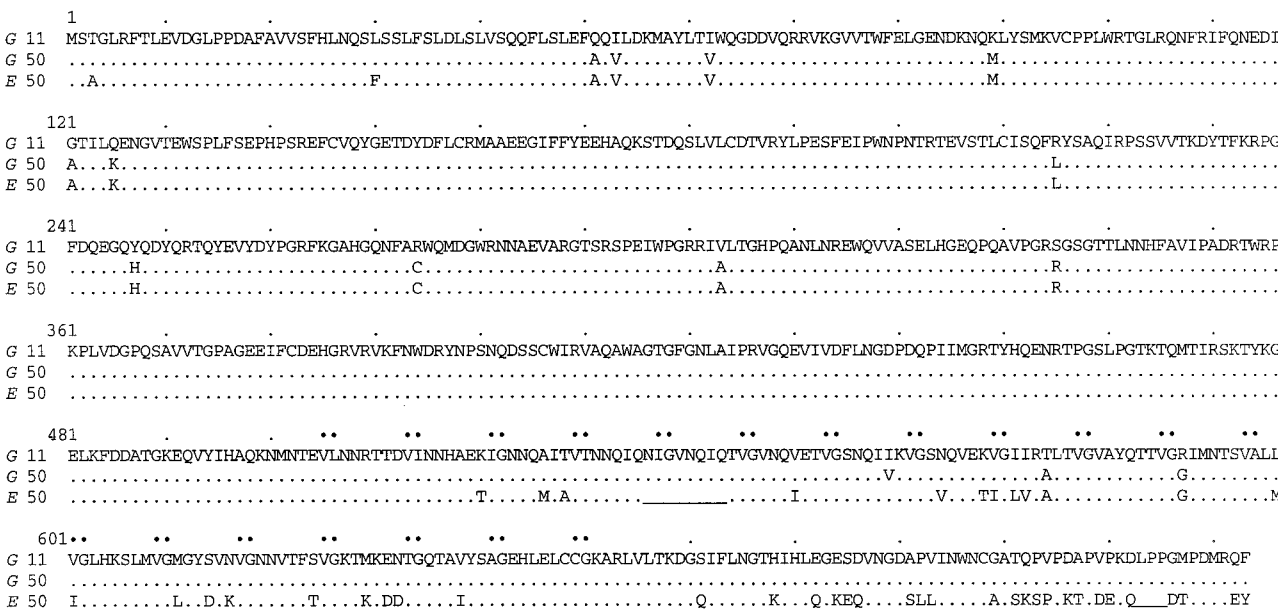

FIG. 5. Comparison of the proteins predicted for (from top to bottom) *vgrG* of ECOR-11 *RhsG*, *vgrG* of ECOR-50 *RhsG*, and *vgrE* of ECOR-50 *RhsE*. Positions of the Val-Gly dipeptide motif, repeated at intervals of 8 aa, are marked (●●).

K-12 (11). Another example of core recombination, involving *RhsD* and *RhsE*, has been noted (16). In both cases, recombination was between cores of the same subfamily. Recombination between cores of different subfamilies is an important issue, since it would increase sequence diversity within a subfamily but reduce sequence diversity between subfamilies. Inspection of aligned nucleotide sequences reveals little evidence of exchange between core subfamilies. As described above, the three subfamilies are distinguished by 695 informative sites distributed over 3,693 bp of homologous sequence. With respect to these sites, the longest interval of identity shared by two subfamilies was 101 bp, and the second longest was 79 bp. This places an upper limit on the sizes of segments exchanged between subfamilies. Thus, despite liberal opportunities for exchange between core subfamilies, little is evident. We suspect that in nature, exchange between subfamilies is negatively selected. (An exception to this generalization is discussed below as it relates to the shuffle of core extensions.) It appears that recombinational exchange is common within subfamilies but highly limited between subfamilies. It is important to note that recombination within subfamilies could reduce diversity; specifically, the 2° substitutions discussed above could be converted to 1° substitutions by gene conversion.

An intriguing aspect of *Rhs* diversification is the shuffling between elements of core extensions and their linked dsORFs. The four new extensions described here (Table 3) bring the total sequenced to 14. All 14 are AT rich (31 to 44%), and all have comparable lengths (130 to 168 amino acids [aa]). Of the 14 extensions, 11 show no detectable sequence similarity and 3 show modest similarity (discussed in reference 20). In previous work, the ECOR reference collection was scored for the location of five core extensions (8). Correlations were seen between the MLEE-based clonal groupings and two examples of core-extension shuffling. We now observe a third case of core-extension shuffling. Specifically, the same extension, ext-e1, was associated with *RhsE* of K-12 and *RhsG* of ECOR-45. (*RhsG* of ECOR-11 had a different extension, ext-g1.) This is the first example of extension movement from one core subfamily to another. We have suggested that shuffling involves simple homologous recombination, using conserved sequences to either side of the extension. Specifically, the core sequences might provide the homology to the left, while H-rpt sequences might provide the homology to the right. If the exchange were to involve cores of different subfamilies, the recipient element would have a recombinant core whose extreme 3′ end resembled the subfamily of the donor rather than that of the recipient. This appears to be true for ECOR-45 *RhsG*. Using the PAUP program, we compared the last 618 bp of ECOR-45 *RhsG* to the 3′ end of 11 other cores. The ECOR-45 *RhsG* sequence as well as the K-12 *RhsE* sequence clearly grouped with the *RhsD-E* subfamily rather than the *RhsG-H* subfamily (Fig. 3B) (bootstrap value of 100%). The sequence immediately to the left of the ECOR-45 *RhsG* core deletion (Fig. 2), on the other hand, clearly belonged to the *RhsG-H* subfamily (not shown). The most straightforward interpretation is that ext-e1 first resided at *RhsE* as it does in K-12 and subsequently was transferred, along with sequences from the 3′ end of the core, to *RhsG*.

Among the known *Rhs* elements, the *RhsE* locus is conspicuous for the extent of its diversity. Three different core extensions are associated with *RhsE* in different strains: ext-e1 in K-12 (16), ext-e4 in ECOR-45, and ext-e5 in ECOR-50 (Table 3). The status of *RhsE* in ECOR-45 is further distinguished from that in K-12 or ECOR-50 in that ECOR-45 *RhsE* is tandemly linked to *RhsH* (Fig. 2). *RhsH* is altogether absent from K-12, and we see no evidence that it was ever present in a K-12 antecedent. The 1,012-bp spacer separating the *RhsE* H-rpt from the *RhsH* core had no homology in the K-12 chromosome or elsewhere in the databases. Work in progress will determine the degree to which the structures of these *Rhs* elements, as well as the shuffling of their core extensions, correlate with the ECOR population structure.

Sequencing of *RhsE* and *RhsG* from ECOR-50 and *RhsG* from ECOR-11 revealed a new component of *Rhs* elements, the *vgr* ORF, which in *RhsG* contained 713 codons (Fig. 5). This ORF was separated from the respective cores of *RhsE* and

RhsG by 66 and 75 bp. These short spacers were quite different from each other. Neither contained obvious promoter or transcription terminator signals. Their close linkage raised the possibility that the *vgr* and core ORFs are cotranscribed, but no experimental information is yet available. A salient feature of the predicted *vgr*-encoded protein was a Val-Gly dipeptide repeated at intervals of 8 aa (marked by large dots in Fig. 5). In the C-terminal third of the ORF, there were 19 repetitions, all of which showed at least one of these two specified residues; 10 of the 19 had both. The *vgr* ORF of *RhsE* possessed one less 8-aa segment than *RhsG* (Fig. 5). Through the first two-thirds of their sequences, ECOR-50 *RhsG vgr* was more similar to the *RhsE* homolog from the same strain than it was to that of ECOR-11 *RhsG*. However, over the last third, the two *RhsG* sequences were much more alike than either was to *RhsE*. One could reasonably speculate that the N-terminal two-thirds of the two *vgr* loci in ECOR-50 became similar through intrachromosomal recombination.

Although they are not homologous, the *vgr* and *Rhs* core ORFs are similar in a number of respects. They are both large, their predicted products are hydrophilic, and they are both characterized by a regularly repeated peptide motif. These features are sometimes associated with ligand-binding proteins found either on the bacterial cell surface or secreted, and they were used to support speculation concerning such a role for the core-ORF (9). Possibly the *vgr* product has such a role.

Portions of both *RhsE* and *RhsG* in K-12 appear to have been deleted. In the case of *RhsE*, alignment of the K-12 and ECOR-50 sequences suggests that 4,437 bp, including the entire 2,106-bp *vgr* ORF, were deleted from K-12. The deletion began 67 bp upstream from the *vgr* ORF and included 2,195 bp of the core. In the case of *RhsG*, alignment of the ECOR-11 and K-12 sequences (see Results) suggested that a K-12 antecedent suffered a 30-kb deletion including both *vgr* and the core, leaving only 291 bp of the H-rpt. It appears likely, therefore, that a K-12 antecedent contained two copies of *vgr*, one at *RhsE* and one at *RhsG*, but both have been deleted. Strain ECOR-11, which is closely related to K-12 in the ECOR phylogeny (7), retains both copies. However, other close relatives of K-12, such as ECOR-1 and ECOR-8, also lack *vgr* homology at either location (our unpublished data). A study tracing these events through the ECOR collection is in progress.

Our understanding of the significance of *Rhs* elements for *E. coli* biology and evolution is severely limited by our ignorance of their function and of the conditions that influence the expression of their various ORFs. Nevertheless, many circumstances suggest that they have an important role. Eight distinct elements, defined by their map locations (Fig. 1A), are widely but not universally distributed among *E. coli* strains. As described above, each of the three *Rhs* core subfamilies moved independently from a GC-rich background into an AT-rich background, possibly *E. coli*. Although the wide *Rhs* distribution indicates genetic mobility, the close correlation of the distribution pattern with the ECOR clonal groupings (8) suggests that their degree of mobility is less than that associated with some ISs and prophages. Our view is that wide distribution in the absence of promiscuous mobility suggests that *Rhs* elements benefit the host cell significantly. Each of the eight *Rhs* cores is more than 3.7 kb long, yet examples of each have maintained an uninterrupted ORF of more than 1,200 codons. In order for such large ORFs to avoid nonsense and frameshift mutations while diverging up to 29%, each of the sequences must have been under strong selective pressure throughout the divergence process.

ACKNOWLEDGMENTS

We thank Du Chungen and Jill Hite for skilled assistance in DNA sequencing.

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

REFERENCES

1. **Bachellier, S., E. Gilson, M. Hofnung, and C. W. Hill.** 1995. Repeated sequences, p. 2012–2046. F. C. Neidhardt, R. Curtiss III, J. L. Ingraham, E. C. C. Lin, K. B. Low, B. Magasanik, W. S. Reznikoff, M. Riley, M. Schaechter, and H. E. Umbarger (ed.), *Escherichia coli* and *Salmonella typhimurium*: cellular and molecular biology, 2nd ed. ASM Press, Washington, D.C.
2. **Blattner, F. R., G. Plunkett III, C. A. Bloch, N. T. Perna, V. Burland, M. Riley, J. Collado-Vides, J. D. Glasner, C. K. Rode, G. F. Mayhew, J. Gregor, N. W. Davis, H. A. Kirkpatrick, M. A. Goeden, D. J. Rose, B. Mau, and Y. Shao.** 1997. The complete genome sequence of *Escherichia coli* K-12. Science **277:**1453–1462.
3. **Boyd, E. F., C. W. Hill, S. M. Rich, and D. L. Hartl.** 1996. Mosaic structure of plasmids from natural populations of *Escherichia coli*. Genetics **143:**1091–1100.
4. **Campbell, A.** 1981. Evolutionary significance of accessory DNA elements in bacteria. Annu. Rev. Microbiol. **35:**55–83.
5. **Ferat, J.-L., M. Le Gouar, and F. Michel.** 1994. Multiple group II self-splicing introns in mobile DNA from *Escherichia coli*. C. R. Acad. Sci. Paris **317:**141–148.
6. **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.
7. **Herzer, P. J., S. Inouye, M. Inouye, and T. S. Whittam.** 1990. Phylogenetic distribution of branched RNA-linked multicopy single-stranded DNA among natural isolates of *Escherichia coli*. J. Bacteriol. **172:**6175–6181.
8. **Hill, C. W., G. Feulner, M. S. Brody, S. Zhao, A. B. Sadosky, and C. H. Sandt.** 1995. Correlation of *Rhs* elements with *Escherichia coli* population structure. Genetics **141:**15–24.
9. **Hill, C. W., C. H. Sandt, and D. A. Vlazny.** 1994. *Rhs* elements of *Escherichia coli*: a family of genetic composites each encoding a large mosaic protein. Mol. Microbiol. **12:**865–871.
10. **Lawrence, J. G., and H. Ochman.** 1997. Amelioration of bacterial genomes: rates of change and exchange. J. Mol. Evol. **44:**383–397.
11. **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.
12. **Muto, A., and S. Osawa.** 1987. The guanine and cytosine content of genomic DNA and bacterial evolution. Proc. Natl. Acad. Sci. USA **84:**166–169.
13. **Ochman, H., and A. C. Wilson.** 1987. Evolutionary history of enteric bacteria, p. 1649–1654. *In* F. C. Neidhardt, J. L. Ingraham, K. B. Low, B. Magasanik, M. Schaechter, and H. E. Umbarger (ed.), *Escherichia coli* and *Salmonella typhimurium*: cellular and molecular biology, vol. 2. American Society for Microbiology, Washington, D.C.
14. **Riley, M., and A. Anilionis.** 1978. Evolution of the bacterial genome. Annu. Rev. Microbiol. **32:**519–560.
15. **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.
16. **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.
17. **Swofford, D. L.** 1993. PAUP 3.1.1. Smithsonian Institution, Washington, D.C.
18. **Vieira, J., and J. Messing.** 1991. New pUC-derived cloning vectors with different selectable markers and DNA replication origins. Gene **100:**189–194.
19. **Williams, S. G., L. T. Varcoe, S. R. Attridge, and P. A. Manning.** 1996. *Vibrio cholerae* Hcp, a secreted protein coregulated with HlyA. Infect. Immun. **64:**283–289.
20. **Zhao, S., and C. W. Hill.** 1995. Reshuffling of *Rhs* components to create a new element. J. Bacteriol. **177:**1393–1398.
21. **Zhao, S., C. H. Sandt, G. Feulner, D. A. Vlazny, J. A. Gray, and C. W. Hill.** 1993. *Rhs* elements of *Escherichia coli* K-12: complex composites of shared and unique components that have different evolutionary histories. J. Bacteriol. **175:**2799–2808.