JOURNAL OF BACTERIOLOGY, Apr. 1999, p. 2548–2554 Vol. 181, No. 8
0021-9193/99/$04.00+0
Copyright © 1999, American Society for Microbiology. All Rights Reserved.

Accessory DNA in the Genomes of Representatives of the *Escherichia coli* Reference Collection

ANA HURTADO AND FRANCISCO RODRÍGUEZ-VALERA*

División de Microbiología, Centro de Biología Molecular y Celular, Campus de San Juan, Universidad Miguel Hernández, 03550 San Juan de Alicante, Spain

Received 8 September 1998/Accepted 15 February 1999

Different strains of the *Escherichia coli* reference collection (ECOR) differ widely in chromosomal size. To analyze the nature of the differential gene pool carried by different strains, we have followed an approach in which random amplified polymorphic DNA (RAPD) was used to generate several PCR fragments. Those present in some but not all the strains were screened by hybridization to assess their distribution throughout the ECOR collection. Thirteen fragments with various degrees of occurrence were sequenced. Three of them corresponded to RAPD markers of widespread distribution. Of these, two were housekeeping genes shown by hybridization to be present in all the *E. coli* strains and in *Salmonella enterica* LT2; the third fragment contained a paralogous copy of *dnaK* with widespread, but not global, distribution. The other 10 RAPD markers were found in only a few strains. However, hybridization results demonstrated that four of them were actually present in a large selection of the ECOR collection (between 42 and 97% of the strains); three of these fragments contained open reading frames associated with phages or plasmids known in *E. coli* K-12. The remaining six fragments were present in only between one and four strains; of these, four fragments showed no similarity to any sequence in the databases, and the other two had low but significant similarity to a protein involved in the *Klebsiella* capsule synthesis and to RNA helicases of archaeal genomes, respectively. Their percent GC, dinucleotide content, and codon adaptation index suggested an exogenous origin by horizontal transfer. These results can be interpreted as reflecting the presence of a large pool of strain-specific genes, whose origin could be outside the species boundaries.

An essential aspect of bacterial population genetics that was seldom considered until recently is the large variation in DNA content of different strains in a single species. Plasmids and phages have been known for a long time to be highly variable in their presence in different isolates from the same species but often represent a small fraction of the total genome. The chromosome has generally been considered a well-conserved replicon. The fact that the genetic maps of *Escherichia coli* K-12 and *Salmonella enterica* LT2 are colinear (22) and of very similar sizes has weighed heavily in favor of assuming relative conservation of chromosome structure and content over wide phylogenetic distances. However, with the development of rapid chromosome mapping derived from pulsed-field gel electrophoresis techniques it has become more and more clear that chromosome size and organization are far from conserved in many species (27). Even in *E. coli*, differences of 1 Mb among the chromosomal sizes of different *E. coli* reference collection (ECOR) strains have been shown previously (2, 3). A similar result was found for two pathogenic strains, in which large DNA insertions were distributed throughout the genome (23). These data indicate that there could be significant differences in gene content among different strains of *E. coli*. We will refer to this DNA present in some but not all the strains in a bacterial species, regardless of its location (plasmid or chromosome), as accessory DNA. Naturally, an essential question is what is the nature of this accessory DNA. In principle, this extra DNA contained in some strains could be repeated genes or duplicated regions of the chromosome, and therefore, the significance of chromosome size variation in terms of the gene complement carried by different strains is unclear.

In one of the few studies addressing this issue (14), it was shown by subtractive hybridization that two strains of *S. enterica* belonging to subspecies I and V (considered to be the most widely divergent within the species) had about 20% of their genes not homologous. Some genes or genomic regions are known to have a variable presence in different strains belonging to the same species. The pathogenicity islands described for several pathogens during the last few years (8) are one example. *Rhs* elements are another well-known example of accessory DNA representing about 0.8% of the *E. coli* genome (9); they seem to have been acquired through separate transfer events from a GC-rich background (30). In the fully sequenced genome of *E. coli* K-12, ca. 2% of the genes are related to phages, plasmids, and transposons (4). They are also obvious candidates to represent accessory DNA. However, it is important to underline that, among the *E. coli* strains whose genome size is known, K-12 has one of the smallest (3). Comparative genomics of three or more strains representing divergent genealogical lineages of the same species would be an ideal way to analyze this problem. However, this kind of information is not yet available.

To contribute to the knowledge of the kind of genes that are present in only some strains of *E. coli*, we have developed an approach that has been already applied to detect subtractive genetic material in other situations (17). A random amplified polymorphic DNA (RAPD) study was carried out to identify amplified DNA bands present in some but not all the strains included in the ECOR collection. RAPD permits the amplification of genomic DNA with a single primer of arbitrary nucleotide sequence under conditions that favor relatively nonspecific binding of the primer to multiple sites of the template DNA (7). The loss of an RAPD marker can be due either to

* Corresponding author. Mailing address: División de Microbiología, Centro de Biología Molecular y Celular, Campus de San Juan, Universidad Miguel Hernández, 03550 San Juan de Alicante, Spain. Phone: 34 96 5919451. Fax: 34 96 5919457. E-mail: FRVALERA @UMH.ES.

Downloaded from http://jb.asm.org/ on September 27, 2020 by guest

the mutation of the primer annealing site or to the absence of the whole sequence or genomic region. To differentiate between these two situations, we carried out a hybridization study with RAPD bands as probes. We found two clearly different situations, one in which the band hybridized to all or most of the ECOR strains and another in which only a very limited number of strains showed a clear hybridization signal. The last RAPD fragments were assumed to represent accessory genes or DNA characteristic of some strains.

MATERIALS AND METHODS

Bacterial strains and culture conditions. The 72 *E. coli* strains of the ECOR reference collection, which includes isolates from a wide variety of hosts and geographic regions (18), were used. *E. coli* K-12 strain CECT 102 (Colección Española de Cultivos Tipo, Valencia, Spain) was used for reference purposes. Five strains of *E. coli* O157 were also included in the study: A8190 and E3406 (O157:H7), provided by T. S. Whittam (Pennsylvania State University, University Park), and E16159 (O157:H45, diffuse adherence), E39233 (O157:H43, heat-labile enterotoxin), and E76561 (O157:H$^-$, VT2 and *eae* genes) provided by H. R. Smith (Central Public Health Laboratory, London, United Kingdom). *S. enterica* serovar Typhimurium LT2 and *Halobacterium* sp. DNAs were used as controls. *E. coli* strains were grown on Luria-Bertani medium at 37°C and stored at −80°C in 15% glycerol.

Isolation of genomic DNA. Bacterial genomic DNA was extracted with InstaGene DNA purification matrix (Bio-Rad Laboratories, Inc.) according to the manufacturer's instructions. The concentration of the DNA samples was spectrophotometrically determined, and aliquots of 100 ng/μl were stored at −20°C.

RAPD PCR amplification. RAPD primers, previously demonstrated to generate polymorphic and reproducible patterns (7), were A1 (5′-TGCGGCTTAC-3′), A7 (5′-TCACGGTGCA-3′), and A10 (5′-GTAGACGAGC-3′). All three primers were synthesized in an Applied Biosystems 396 DNA-RNA synthesizer and diluted to 10 μM-concentration aliquots which were stored at −20°C.

The reaction mixtures of 50 μl contained 1 μg of genomic DNA, 10 mM Tris-HCl (pH 8.3), 50 mM KCl, 2 mM $MgCl_2$, 0.01% (wt/vol) gelatin, 0.2 mM (each) deoxynucleotide, 0.8 μM primer, and 2 U of *Thermus aquaticus* (*Taq*) polymerase (Gibco BRL, Life Technologies Ltd., Paisley, United Kingdom). PCR was performed with a PTC-100 automatic thermal cycler (MJ Research, Inc., Watertown, Mass.). The mixture was subjected to 35 cycles of the following incubations: denaturation at 94°C for 30 s, annealing at 36°C for 1 min, and extension at 72°C for 2 min. A final extension was performed at 72°C for 10 min. Negative controls without template DNA were included. The PCR products were analyzed by electrophoresis through 1.0% (wt/vol) agarose gels, and the bands were visualized by staining with ethidium bromide and excitation under UV light on a transilluminator.

DNA-DNA dot blot hybridization. A 5-μg aliquot of each genomic DNA was transferred to a Hybond N$^+$ nylon membrane (Amersham) with a vacuum-blotting apparatus (Bio-Dot apparatus; Bio-Rad Laboratories, Inc.). DNA was fixed to filters by baking it for 2 h at 80°C. Labeling of probes and blot development were performed by using the enhanced chemiluminescence (ECL) system (Amersham). Probes were prepared from selected RAPD PCR bands excised from the gel and purified with the Sephaglas BrandPrep kit (Pharmacia). Following a 1-h prehybridization at 42°C in ECL gold hybridization buffer supplemented with 0.5 M NaCl and 5% blocking agent, 200 ng of labeled probe was added and incubation was carried out overnight at 42°C. Hybridized filters were washed for 5 min at 42°C in 5× SSC (1× SSC is 0.15 M NaCl plus 0.015 M sodium citrate), followed by two 20-min stringent washes at 42°C in 6 M urea–0.4% sodium dodecyl sulfate–0.1× SSC. A final 5-min wash at room temperature with 2× SSC was carried out prior to exposing the membranes to the detection reagents. Chemiluminescence was detected on autoradiography films (Hyperfilm ECL; Amersham) after signal generation for 30 min.

Cloning and sequencing of RAPD fragments. Purified RAPD fragments were inserted into pCR 2.1 vector by ligation at 14°C overnight and transformed into INVαF′ One Shot competent cells (original TA cloning kit; Invitrogen Co., Carlsbad, Calif.) according to the manufacturer's instructions. The presence of insert was confirmed by PCR amplification with the M13 forward (−20) primer and the M13 reverse primer.

Nucleotide sequences of both strands of 15 inserts (over 12 kbp) were determined by primer walking with the Thermo Sequenase dye-deoxy terminator cycle sequencing premix kit (Amersham) and the ABI 377 automated DNA sequencer according to the manufacturer's instructions.

Computer analysis of nucleotide sequences. Putative open reading frames (ORFs) and G+C moles percent composition analyses were performed with the DNAStar package. Similarities between determined RAPD sequences and those from the GenBank database were calculated by BLAST searches (BLASTN and BLASTX) carried out at the network server of the National Center for Biotechnology Information. BLASTN similarity searches screen the nucleotide query sequence against nucleotide databases; BLASTX similarity searches compare a nucleotide query sequence translated in all reading frames against a protein sequence database. Alignments were performed with ClustalX. Dinucleotide GC relative abundance profiles (ρ^*_{GC}) were assessed through the odds ratio functional $\rho^*_{GC} = f_{GC}/f_G f_C$, where f_G denotes the frequency of nucleotide G and f_{GC} is the frequency of the dinucleotide GC in the sequence under study. Values of ρ^*_{GC} in the range of 1.20 to 1.35 are considered a genome signature for *E. coli* (10). The codon adaptation index (CAI) was calculated for each ORF according to the method of Sharp and Li (25) at the Virtual Genome Center website (27a). Values of CAI below 0.25 indicate unusual codon usage in *E. coli* and may signify recent immigration by horizontal transfer (4).

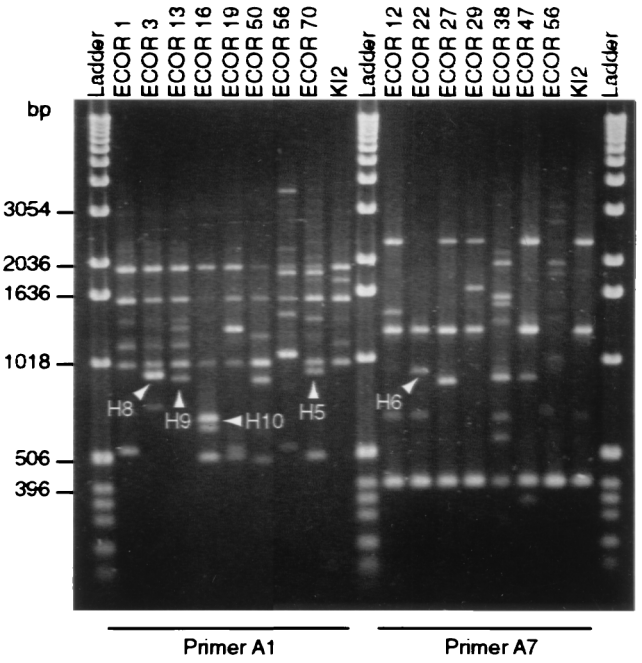

FIG. 1. Example of the electrophoresis patterns generated by RAPD amplification of some ECOR collection strains and *E. coli* K-12 with primers A1 and A7. The molecular size marker used is a 1-kb ladder (Gibco BRL, Life Technologies Ltd.). The strain number is indicated above each lane, and the primer used is shown below. Arrows point to some of the fragments used as hybridization probes in dot blot analysis.

Nucleotide sequence accession numbers. The sequences reported in this paper were deposited in GenBank with accession numbers from AF 127003 to AF 127017.

RESULTS

Analysis of RAPD products. Three 10-mer oligonucleotides previously demonstrated to generate polymorphic and reproducible patterns (7) were selected as RAPD primers. Genomic DNA from all 72 strains of the ECOR collection as well as five representatives of the *E. coli* serotype O157 (verotoxigenic strains) and the *E. coli* reference strain K-12 were amplified as described above (Materials and Methods). As expected (29), a high degree of polymorphism was found with the three primers. Figure 1 shows an example of the patterns obtained. Twenty-five different bands were included in the patterns generated with primer A1; among these, four bands were present in over 70% of the strains. In the case of primer A7, 20 different bands were amplified, and four of them were present in over 50% of the strains. A higher degree of polymorphism was found with primer A10, which generated 30 different bands, and among them, only three were present in more than 50% of strains. In total, 75 different bands were generated with the three primers, and only 12 of them were present in more than 50% of the strains. On the other hand, 10 RAPD fragments were present in less than 10% of the strains (see Table 1). These 10 plus three RAPD fragments with widespread distribution (H1, H2,

Downloaded from http://jb.asm.org/ on September 27, 2020 by guest

and H4) were selected for further study. Two of them (H4 and H13) were cloned and sequenced from two different strains.

Distribution of the RAPD fragments in the ECOR collection. The selected RAPD fragments were used as hybridization probes to determine the distribution of these sequences among the collection of strains tested. The selected bands (Table 1) were purified from the gel and used as probes against a dot blot containing total DNA from the whole set of strains, as well as DNA from other organisms included as controls (*S. enterica* LT2 and *Halobacterium* sp.). Hybridization results are summarized in Table 1 and Fig. 2, and they showed that (i) as expected, each probe hybridized strongly to the genomic DNA of the strain from which the band was isolated; (ii) among the RAPD fragments selected for their low frequency, six showed a very restricted distribution by hybridization, i.e., four fragments (H8, H9, H10, and H12) hybridized solely to the genomic DNA extracted from the strain of origin (indicating that those sequences were present in only a single strain within the set), one fragment (H11) hybridized with one additional strain, and another one (H13) hybridized with three strains; (iii) the remaining seven fragments, including the three selected for their high frequency (H1, H2, and H4), had a widespread distribution, hybridizing with between 40 and 100% of strains (Table 1). Those three plus H7 also hybridized with *S. enterica* LT2, showing a supraspecific distribution. None hybridized to *Halobacterium* sp. DNA. Several strains that did not generate the corresponding RAPD marker gave a positive hybridization result. This was probably due to the modification of the primer annealing sequence or the loss of a short DNA stretch including the primer target.

Sequencing of RAPD fragments. The 13 RAPD fragments used as probes were cloned and sequenced as indicated above (Materials and Methods). Additionally, two of them were sequenced in duplicate, i.e., the same band was cloned and sequenced from a second strain. The 15 RAPD fragments sequenced could be classified into three categories with different putative roles, frequencies of distribution, and putative origins (Table 1).

The first category consisted of the two most common RAPD fragments (H1 and H2), both of which included the central region of an ORF coding for basic cell functions (phosphoglycerol transferase I, involved in oligosaccharide biosynthesis, and the DNA adenine methylase DamX). They were present in all the strains tested and had values for G+C, ρ^*_{GC}, and CAI typical of *E. coli* K-12 genes. The second category included five fragments with a medium frequency of distribution that corresponded to either hypothetical proteins in intergenic regions or genes belonging to phages and/or plasmids. Fragment H3 showed high homology to the carboxyl-terminal end of PitA (a low-affinity phosphate transporter) and the complete hypothetical YhiO protein. The latter showed 99.1% amino acid identity to *E. coli* K-12 but at the nucleotide level had an extra 175-nucleotide block with respect to *E. coli* K-12 in the PitA-YhiO intergenic region. Also in this category was H4, which was found to be highly similar (99% amino acid similarity) to a paralogous copy of the essential chaperonin DnaK present in the LeuS-GltL intergenic region of *E. coli* K-12 (4). This fragment was sequenced from two strains (ECOR 2 and ECOR 44), and a 5.4% difference in nucleotides was found between them. This category also included fragments (H5, H6, and H7) that corresponded to genes belonging to phages and/or plasmids. They are promiscuous genetic elements and probably form a significant part (about 30% in this case) of accessory DNA. They show values for G+C and ρ^*_{GC} that fall outside the range considered normal for *E. coli* genes but not dramatically so; CAI values are within the normal limits described for *E. coli* K-12.

The rest of the fragments sequenced, H8 to H13 (category III), were present in a very restricted range of strains. They showed very low levels of similarity to sequences in the databases and in most cases (all except H9) had compositional parameters (i.e. percent G+C and ρ^*_{GC}) and CAI values markedly different from those typical of *E. coli* K-12 genes, for example, % G+C = 37.5, ρ^*_{GC} = 1.45, and CAI = 0.21 in the case of H10, which strongly suggest an extraspecific origin for this genetic material. H8, H9, H11, and H12 showed no significant similarity to any entry in the databases, either at DNA level or in the amino acids of the ORFs detected within them. The best BLASTX hit for H10 (60.8% identity) was to an 80.4-kDa protein coding for a gene located in the central region of the CPS gene cluster of *Klebsiella pneumoniae* serotype K2 which is essential for the capsular polysaccharide synthesis (1). The *Klebsiella cps* region consists of 13 ORFs that constitute a polycistronic operon which, interestingly, has an unusual G+C content (below 40%), in agreement with our results for H10. A lower similarity (53.4% amino acid identity) was found to a homologous protein (YccC) in *E. coli* K-12. However, these similarity levels were notably lower than those found for sequences in categories I and II. H13a and H13b (both representing the same RAPD band in two strains, ECOR 44 and ECOR 69) showed significant similarity to thermophilic RNA helicases described for archaeal genomes such as those of *Methanobacterium thermoautotrophicum* (26), *Archaeoglobus fulgidus* (12), and *Pyrococcus horikoshii* (11); however, no significant similarity was found to the *E. coli* K-12 RNA helicase. RNA helicases in *E. coli* have a role in postreplication repair and are also part of the degradosome, a multienzymatic complex involved in RNA processing and mRNA degradation (20). The sequences of ECOR 44 (multilocus enzyme electrophoresis [MLEE] cluster D) and ECOR 69 (MLEE cluster B1) were 99.3% identical, revealing a remarkable degree of conservation.

DISCUSSION

The approach that we have followed here permits the identification of sequences that are present in only some strains within a single species. The two fragments present in all the strains are part of ORFs with high similarity to *E. coli* K-12 housekeeping (HK) genes. Obviously, this kind of sequence was not the target of our study, and they were included only as a kind of blank or negative control. In both cases, the primer annealing target is included within the ORFs, which obviously contributes to the stability of the RAPD product. However, none of the bands were present in more than 75% of the RAPD patterns generated from the strains studied. In these strains, the absence of the RAPD marker must be due to point mutations in the annealing target. In fact, the conservation of the ORF coding for DamX in ECOR 43 was only 92.9% at the amino acid level, which is quite low for an HK gene (31). The rest of the RAPD fragments assayed (all the ones selected by their restricted distribution as RAPD markers) were detected by hybridization only in some strains. That supports this strategy for recovering accessory DNA from the whole genome of bacterial strains.

The five fragments included in category II are present in 42.3 to 97.4% of the assayed strains as shown by hybridization. H5, H6, and H7 are clearly related to *E. coli* phages and/or plasmids, and therefore, there is little doubt about their nature. The RAPD approach followed would retrieve both chromosomal and extrachromosomal fragments, and we have no in-

Downloaded from http://jb.asm.org/ on September 27, 2020 by guest

Downloaded from http://jb.asm.org/ on September 27, 2020 by guest

TABLE 1. Characteristics and distribution of the RAPD fragments classified into three categories[a]

Fragment designation	Strain	Primer	Fragment size (nucleotides)	ORF size (amino acids [aa])	% of RAPD positives	% of hybridization positives[b]	Similarity[c] (%, organism, gene and/or predicted function)	% G+C	ρ^*_{GC}[d]	CAI[e]
Category I										
H1	ECOR 10	A1	635	211 aa (central region)	72	100	99.0%, K-12, phosphoglycerol transferase I	51.97	1.14	0.40
H2	ECOR 43	A10	760	253 aa (central region)	70	100	92.9%, K-12, DamX, DNA adenine methylase	52.89	1.17	0.38
Category II										
H3	ECOR 44	A1	748	111 aa (complete)	2.6	78.2	99.1%, K-12, 13-kDa hypothetical protein in the PitA-UspA intergenic region	49.33	1.26	**0.23**
H4a	ECOR 2	A7	1,248	376 aa (COOH end)	66.7	88.5	99.0% (ECOR 2) and 96.0% (ECOR 44), K-12, paralogous to DnaK (heat shock protein), a 62-kDa protein in the LeuS-GltL intergenic region	52.00	1.37	0.29
H4b	ECOR 44	A7	1,248	376 aa (COOH end)				51.04	1.42	0.30
H5	ECOR 70	A1	948	288 aa (COOH end)	9.0	42.3	93.1%, K-12, phage T4 tail protein GP37	50.42	**1.15**	0.29
H6	ECOR 22	A7	980	314 aa (COOH end)	3.8	69.2	78.0%, *E. coli*, phi-R73 retronphage (putative DNA primase)	**59.49**	**1.12**	0.35
H7	ECOR 33	A7	923	260 aa (COOH end)	7.7	97.4	85.4%, *E. coli*, pColV-K30 plasmid, hypothetical protein	**56.88**	**1.10**	0.27
Category III										
H8	ECOR 3	A1	934	173 aa (complete)	0	0	No significant homology	**41.43**	**1.40**	**0.23**
H9	ECOR 13	A1	915	279 aa (NH_2 end)	0	0	No significant homology	47.98	1.35	0.26
H10	ECOR 16	A1	678	207 aa (NH_2 end)	0	0	60.8%, *Klebsiella*, 80.4-kDa hypothetical protein in CPS region	**37.46**	**1.45**	**0.21**
H11	ECOR 9	A10	684	132 aa (NH_2 end)	1.3	1.3	No significant homology	**41.37**	**1.07**	**0.19**
H12	ECOR 66	A10	422	127 aa (NH_2 end)	0	0	No significant homology	**47.16**	**0.81**	**0.21**
H13a	ECOR 44	A10	836	253 aa (NH_2 end)			46%, *Methanobacterium thermoautotrophicum*, ATP-dependent RNA helicase, postreplication repair	**48.92**	**1.15**	**0.20**
H13b	ECOR 69	A10	836	253 aa (NH_2 end)	3.8	3.8		**48.56**	**1.17**	**0.21**

[a] See the text. Boldface numbers correspond to unusual values of percent G+C, ρ^*_{GC}, and CAI for *E. coli* (4).
[b] All the probes hybridized with genomic DNA from the strain of origin; the percentage of other strains with which each probe hybridized is indicated in this column.
[c] Similarities were estimated by aligning with ClustalX the complete query ORF with the protein that gave the strongest BLASTX hit. Similarities below 40% along at least 60% of the query ORF were considered nonsignificant. BLASTN similarity searches were also carried out (data not shown).
[d] ρ^*_{GC}, dinucleotide GC relative abundance profiles; values in the range of 1.20 to 1.35 are considered a genome signature for *E. coli*.
[e] Values below 0.25 indicate unusual codon usage in *E. coli*.

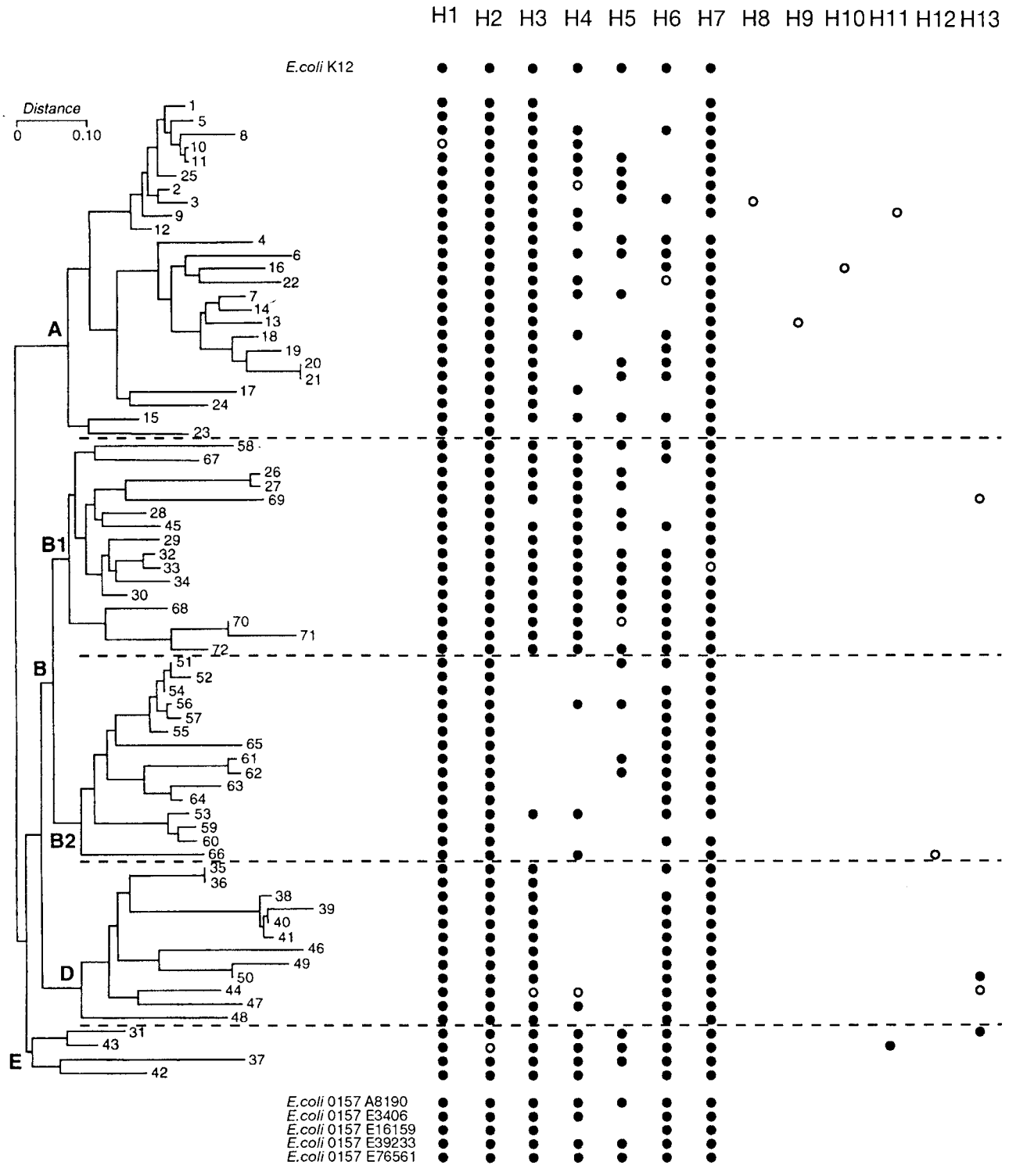

FIG. 2. Schematic representation of the distribution of the RAPD fragments among the *E. coli* ECOR collection strains, as well as the five *E. coli* O157 strains and K-12, as determined by dot blot hybridization results (H1 to H13 [see Table 1 for details]). Positive results in dot blot analysis are indicated by black dots; open dots correspond to the positive controls, i.e., the strain of origin of the RAPD fragment. The ECOR collection strains are grouped in a dendrogram based on MLEE results (24) with the different cluster designations on the outside branches.

formation as to their location. However, integration of plasmidic and/or phage material into the chromosome seems to be a common fact (6), and therefore, the location of gene clusters in chromosomal or extrachromosomal units can vary rapidly in evolutionary time. Plasmids and phages are the paradigm of accessory DNA, and, if anything, the relatively low number of fragments of this group found here is surprising. However, some of the fragments included in category III could also

Downloaded from http://jb.asm.org/ on September 27, 2020 by guest

represent this kind of genetic element, albeit unknown ones. It is remarkable that two of those phage-plasmid-related fragments are found in most of the ECOR strains, but the ORF related to phage T4 found in H5 has a distribution with a certain phylogenetic consistency, being present in most strains of groups A, B1, and E; scarce in B2; and absent in group D strains. Phage 21 has also been reported to have a distribution correlated with the MLEE groups, although its evolution is not correlated with that of the host (28). The consistency in the compositional parameters and codon usage found for these sequences is remarkable. They are slightly divergent from the average for the *E. coli* K-12 genome, and the CAI corresponds to that of *E. coli* K-12 genes expressed at a low level. H3 and H4 also had high similarities to sequences described for the *E. coli* K-12 genome. H3 contains the carboxyl end of PitA plus a part of the PitA-UspA intergenic region. This region, as in *E. coli* K-12, contains an ORF coding for a hypothetical protein, YhiO, which was very conserved in ECOR 44, with 99.1% similarity to *E. coli* K-12 at the amino acid level. Its function might be important because the degree of conservation is consistent with that of HK genes. Something similar applies to H4, which contains an ORF highly similar to a paralogous copy of *dnaK* found in the *E. coli* K-12 genome. Paralogous copies of essential genes have been shown to represent significant parts of the relatively large bacterial genomes fully sequenced (*E. coli* and *Bacillus subtilis*) (4, 13). In *E. coli* K-12, nearly a third of its genes have paralogous copies in the genome (4). The function of those supplementary copies of genes is unclear. The presence of this paralogous gene in only a limited group of strains could indicate that, in this case, it is either a nonessential copy with an adaptive function, i.e., required only to survive under certain conditions, or a pseudogene. In fact, the relatively high conservation found for the three strains in which this fragment has been sequenced (K-12, ECOR 2, and ECOR 44) supports the first notion. The absence of the RAPD fragments equivalent to H3 and H4 from many strains in which their presence was detected by hybridization can be justified by the location of one of the primer annealing targets for both fragments in intergenic regions that are possibly subjected to a higher rate of change.

The fragments classified in category III have indeed a very restricted distribution throughout the ECOR collection. Four of them were found in only one strain. This fact indicates a very recent acquisition and/or a high adaptive specificity of this genetic material. On the other hand, the fragments within this category that were found in more than one strain were present in relatively distant strains (H13 in groups B1, D, and E and H11 in groups B2 and E). The ORF within H10 shows a residual but significant similarity (53.4%) to a hypothetical protein described for *E. coli* K-12; for the rest, the lack of similarity with *E. coli* K-12 sequences confirms that they represent actual differential DNA without any homology to the DNA present in some *E. coli* strains. The compositional parameters (percent G+C, ρ^*_{GC}) and CAI are widely divergent from those of the average *E. coli* K-12 genome, and they are reminiscent of the 10% with atypical CAI values described for the K-12 genome (4). This is normally interpreted as the result of recent immigration by horizontal transfer (3, 15, 16). Low G+C values (around 40%) have also been found in pathogenicity islands (5, 19). The kinds of strain-specific sequences that we have retrieved are very unlikely to represent repeated copies of other parts of the genome. Furthermore, the fact that, among the 10 fragments with restricted distribution studied, four were restricted to the strain of origin stresses the importance of the differential gene complement carried by different strains.

Most population genetic studies of bacteria are concerned with sequence variation for a single gene (allelic variation). However, a major contribution to the intraspecific variation in bacteria could be ascribed to the presence of different pools of accessory genes in different strains rather than different alleles of the same genes (the latter is the case in the populations of higher organisms). If these different sets of adaptative genes carried by different lineages within the species are very large, they would be a major factor to consider in bacterial evolution. As stated elsewhere by Reanney (21), "In an ecosystem where the microbiota frequently exchange these genes, the size of the entity upon which selection acts may exceed by orders of magnitude the size of the gene pool of a clonal population."

ACKNOWLEDGMENTS

Financial support was from CICYT PM95-0111 and DGICYT PB96-0793-C04-02 to F.R.-V. A.H. was the recipient of a postdoctoral contract of the Spanish Ministry of Education and Science.

We thank H. R. Smith (Central Public Health Laboratory) and T. S. Whittam (Pennsylvania State University) for providing *E. coli* O157 strains. We are grateful to Kathy Hernandez for secretarial assistance and to Stuart Ingham for illustration.

REFERENCES

1. **Arakawa, Y., R. Wacharotayankun, T. Nagatsuka, H. Ito, N. Kato, and M. Ohta.** 1995. Genomic organization of the *Klebsiella pneumoniae cps* region responsible for serotype K2 capsular polysaccharide synthesis in the virulent strain Chedid. J. Bacteriol. **177:**1788–1796.
2. **Bergthorsson, U., and H. Ochman.** 1998. Distribution of chromosome length variation in natural isolates of *Escherichia coli*. Mol. Biol. Evol. **15:**6–16.
3. **Bergthorsson, U., and H. Ochman.** 1995. Heterogeneity of genome sizes among natural isolates of *Escherichia coli*. J. Bacteriol. **177:**5784–5789.
4. **Blattner, F. R., G. P. 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.
5. **Boyd, E. F., and D. L. Hartl.** 1998. Chromosomal regions specific to pathogenic isolates of *Escherichia coli* have a phylogenetically clustered distribution. J. Bacteriol. **180:**1159–1165.
6. **Cheetham, B. F., and M. E. Katz.** 1995. A role of bacteriophages in the evolution and transfer of bacterial virulence determinants. Mol. Microbiol. **18:**201–208.
7. **García-Martinez, J., A. J. Martínez-Murcia, F. Rodriguez-Valera, and A. Zorraquino.** 1996. Molecular evidence supporting the existence of two major groups in uropathogenic *Escherichia coli*. FEMS Immunol. Med. Microbiol. **14:**231–244.
8. **Hacker, J., G. Blum-Oehler, I. Muhldorfer, and H. Tschape.** 1997. Pathogenicity islands of virulent bacteria: structure, function and impact on microbial evolution. Mol. Microbiol. **23:**1089–1097.
9. **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.
10. **Karlin, S., J. Mrázek, and A. M. Campbell.** 1997. Compositional biases of bacterial genomes and evolutionary implications. J. Bacteriol. **179:**3899–3913.
11. **Kawarabayasi, Y., M. Sawada, H. Horikawa, Y. Haikawa, Y. Hino, S. Yamamoto, M. Sekine, S. Baba, H. Kosugi, A. Hosoyama, Y. Nagai, M. Sakai, K. Ogura, R. Otuka, H. Nakazawa, M. Takamiya, Y. Ohfuku, T. Funahashi, T. Tanaka, Y. Kudoh, J. Yamazaki, N. Kushida, A. Oguchi, K. Aoki, Y. Nakamura, T. F. Robb, K. Horikoshi, Y. Masuchi, H. Shizuya, and H. Kikuchi.** 1998. Complete sequence and gene organization of the genome of a hyper-thermophilic Archaebacterium, *Pyrococcus horikoshii* OT3. DNA Res. **5:**55–76.
12. **Klenk, H.-P., R. A. Clayton, J.-F. Tomb, O. White, K. E. Nelson, K. A. Ketchum, R. J. Dodson, M. Gwinn, E. K. Hickey, J. D. Peterson, D. L. Richardson, A. R. Kerlavage, D. E. Graham, N. C. Kyrpides, R. D. Fleischmann, J. Quackenbush, N. H. Lee, G. G. Sutton, S. Gill, E. F. Kirkness, B. A. Dougherty, K. McKenney, M. D. Adams, B. Loftus, S. Peterson, C. I. Reich, L. K. McNeil, J. H. Badger, A. Glodek, L. Zhou, R. Overbeek, J. D. Gocayne, J. F. Weidman, L. McDonald, T. Utterback, M. D. Cotton, T. Spriggs, P. Artiach, B. P. Kaine, S. M. Sykes, P. W. Sadow, K. P. D'Andrea, C. Bowman, C. Fijii, S. A. Garland, T. M. Mason, G. J. Olsen, C. M. Fraser, H. O. Smith, C. R. Woese, and J. C. Venter.** 1997. The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon *Archaeoglobus fulgidus*. Nature **390:**364–370.

Downloaded from http://jb.asm.org/ on September 27, 2020 by guest

13. **Kunst, F., N. Ogasawara, I. Moszer, A. Albertini, G. Alloni, V. Azevedo, M. Bertero, P. Bessieres, A. Bolotin, S. Borchert, R. Borriss, L. Boursier, A. Brans, M. Braun, S. Brignell, S. Bron, S. Brouillet, C. Bruschi, B. Caldwell, V. Capuano, N. Carter, S. Choi, J. Codani, I. Connerton, R. Daniel, F. Denizot, K. Devine, A. Dusterhoft, S. Ehrlich, P. Emmerson, K. Entian, J. Errington, C. Fabret, E. Ferrari, D. Foulger, C. Fritz, M. Fujita, Y. Fujita, S. Fuma, A. Galizzi, N. Galleron, S.-Y. Ghim, P. Glaser, A. Goffeau, E. Golightly, G. Grandi, G. Guiseppi, B. Guy, K. Haga, J. Haiech, C. Harwood, A. Henaut, H. Hilbert, S. Holsappel, S. Hosono, M.-F. Hullo, M. Itaya, L. Jones, B. Joris, D. Karamata, Y. Kasahara, M. Klaerr-Blanchard, C. Klein, Y. Kobayashi, P. Koetter, G. Koningstein, S. Krogh, M. Kumano, K. Kurita, A. Lapidus, S. Lardinois, J. Lauber, V. Lazarevic, S.-M. Lee, A. Levine, H. Lui, S. Masuda, C. Mauel, C. Medigue, N. Medina, R. Mellado, M. Mizuno, D. Moestl, S. Nakai, M. Noback, D. Noone, M. O'Reilly, K. Ogawa, A. Ogiwara, B. Oudeg, S.-H. Park, V. Parro, T. Pohl, D. Portetelle, S. Porwollik, A. Prescott, E. Presecan, P. Pujic, B. Purnelle, G. Rapoport, et al.** 1997. The complete genome sequence of the Gram-positive bacterium *Bacillus subtilis*. Nature **390:**249–256.
14. **Lan, R., and P. R. Reeves.** 1996. Gene transfer is a major factor in bacterial evolution. Mol. Biol. Evol. **13:**47–55.
15. **Lawrence, J. G., and H. Ochman.** 1997. Amelioration of bacterial genomes: rates of change and exchange. J. Mol. Evol. **44:**383–397.
16. **Lawrence, J. G., and H. Ochman.** 1998. Molecular archaeology of the *Escherichia coli* genome. Proc. Natl. Acad. Sci. USA **95:**9413–9417.
17. **Mathieu-Daudé, F., K. Evans, F. Kullmann, R. Honeycutt, T. Vogt, J. Welsh, and M. McClelland.** 1998. Applications of DNA and RNA fingerprinting by the arbitrarily primed polymerase chain reaction, p. 414–436. *In* F. J. de Bruijn, J. R. Lupski, and G. M. Weinstock (ed.), Bacterial genomes. Physical structure and analysis. Chapman & Hall, New York, N.Y.
18. **Ochman, H., and R. K. Selander.** 1984. Standard reference strains of *Escherichia coli* from natural populations. J. Bacteriol. **157:**690–693.
19. **Perna, N. T., G. F. Mayhew, G. Posfai, S. Elliott, M. S. Donnenberg, J. B. Kaper, and F. R. Blattner.** 1998. Molecular evolution of a pathogenicity island from enterohemorrhagic *Escherichia coli* O157:H7. Infect. Immun. **66:**3810–3817.
20. **Py, B., C. F. Higgins, H. M. Krisch, and A. J. Carpousis.** 1996. A DEAD-box RNA helicase in the *Escherichia coli* RNA degradosome. Nature **381:**169–172.
21. **Reanney, D.** 1976. Extrachromosomal elements as possible agents of adaptation and development. Bacteriol. Rev. **40:**552–590.
22. **Riley, M., and S. Krawiec.** 1987. Genome organization, p. 967–981. *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.
23. **Rode, C. K., L. J. Melkerson-Watson, A. T. Johnson, and C. A. Bloch.** 1999. Type-specific contributions to chromosome size differences in *Escherichia coli*. Infect. Immun. **19:**230–236.
24. **Selander, R. K., J. Li, E. F. Boyd, F.-S. Wang, and K. Nelson.** 1994. DNA sequence analysis of the genetic structure of populations of *Salmonella enterica* and *Escherichia coli*, p. 17–49. *In* F. G. Priest **et al.** (ed.), Bacterial diversity and systematics. Plenum Press, New York, N.Y.
25. **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.
26. **Smith, D. R., L. A. Doucette-Stamm, C. Deloughery, H. Lee, J. Dubois, T. Aldredge, R. Bashirzadeh, D. Blakely, R. Cook, K. Gilbert, D. Harrison, L. Hoang, P. Keagle, W. Lumm, B. Pothier, D. Qiu, R. Spadafora, R. Vicaire, Y. Wang, J. Wierzbowski, R. Gibson, N. Jiwani, A. Caruso, D. Bush, H. Safer, D. Patwell, S. Prabhakar, S. McDougall, G. Shimer, A. Goyal, S. Pietrokovski, G. M. Church, C. J. Daniels, J.-I. Mao, P. Rice, J. Nölling, and J. N. Reeve.** 1997. Complete genome sequence of *Methanobacterium thermoautotrophicum* ΔH: functional analysis and comparative genomics. J. Bacteriol. **179:**7135–7155.
27. **Trevors, J. T.** 1996. Genome size in bacteria. Antonie Leeuwenhoek **69:**293–303.

27a. **Virtual Genome Center Website.** 21 September 1995, revision date. [Online.] http://alces.med.umn.edu/VGC.html. [25 August 1998, last date accessed.]

28. **Wang, F.-S., T. S. Whittam, and R. K. Selander.** 1997. Evolutionary genetics of the isocitrate dehydrogenase gene (*icd*) in *Escherichia coli* and *Salmonella enterica*. J. Bacteriol. **179:**6551–6559.
29. **Wang, G., T. Whittam, C. M. Berg, and D. E. Berg.** 1993. RAPD (arbitrary primer) PCR is more sensitive than multilocus enzyme electrophoresis for distinguishing related bacterial strains. Nucleic Acids Res. **21:**5930–5933.
30. **Wang, Y.-D., S. Zhao, and C. W. Hill.** 1998. *Rhs* elements comprise three subfamilies which diverged prior to acquisition by *Escherichia coli*. J. Bacteriol. **180:**4102–4110.
31. **Whittam, T. S.** 1996. Genetic variation and evolutionary processes in natural populations of *Escherichia coli*, p. 2708–2722. *In* 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*: cellular and molecular biology, 2nd ed., vol. 2. ASM Press, Washington, D.C.

Downloaded from http://jb.asm.org/ on September 27, 2020 by guest