CRISPR Distribution within the *Escherichia coli* Species Is Not Suggestive of Immunity-Associated Diversifying Selection

Marie Touchon, Sophie Charpentier, Olivier Clermont, Eduardo P. C. Rocha, Erick Denamur, and Catherine Branger

Institut Pasteur, Microbial Evolutionary Genomics, Département Génomes et Génétique, F-75015 Paris, France; CNRS, UR2171, F-75015 Paris, France; UPMC Université Pierre et Marie Curie, Atelier de Bioinformatique, F-75005 Paris, France; AP-HP, Hôpital Louis Mourier, Service de Microbiologie-Hygiène, Colombes, France; INSERM, UMR-S 722, F-75018 Paris, France; and Université Paris Diderot, Faculté de Médecine, Site Xavier Bichat, F-75018 Paris, France

Received 29 October 2010/Accepted 7 March 2011

In order to get further insights into the role of the clustered, regularly interspaced, short palindromic repeats (CRISPRs) in *Escherichia coli*, we analyzed the CRISPR diversity in a collection of 290 strains, in the phylogenetic framework of the strains represented by multilocus sequence typing (MLST). The set included 263 natural *E. coli* isolates exposed to various environments and isolated over a 20-year period from humans and animals, as well as 27 fully sequenced strains. Our analyses confirm that there are two largely independent pairs of CRISPR loci (CRISPR1 and -2 and CRISPR3 and -4), each associated with a different type of *cas* genes (*Ecoli* and *Ypest*, respectively), but that each pair of CRISPRs has similar dynamics. Strikingly, the major phylogenetic group B2 is almost devoid of CRISPRs. The majority of genomes analyzed lack *Ypest* cas genes and contain CRISPR3 with spacers matching *Ypest* cas genes. The analysis of relatedness between strains in terms of spacer repertoire and the MLST tree shows a pattern where closely related strains (MLST phylogenetic distance of \(<0.005\) corresponding to at least hundreds of thousands of years) often exhibit identical CRISPRs while more distantly related strains (MLST distance of \(>0.01\)) exhibit completely different CRISPRs. This suggests rare but radical turnover of spacers in CRISPRs rather than CRISPR gradual change. We found no link between the presence, size, or content of CRISPRs and the lifestyle of the strains. Our data suggest that, within the *E. coli* species, CRISPRs do not have the expected characteristics of a classical immune system.

Clustered, regularly interspaced, short palindromic repeats (CRISPRs) were initially discovered in *Escherichia coli* (19) and have been recently identified in most archaea and many bacteria (35). CRISPRs provide acquired immunity against viruses and plasmids by targeting nucleic acid in a sequence-specific manner (18). They typically consist of short (23 to 47 bp) and highly conserved direct repeats regularly separated by stretches of variable sequences called spacers. These spacer sequences sometimes match sequences from phages and plasmids (protopacers) and are therefore thought to derive from them. CRISPRs have been shown experimentally to protect *Streptococcus thermophilus* and *Staphylococcus epidermidis* from phages (3) and plasmids (25), respectively. However tempting it might be to establish a link between this interference and the evolution of CRISPRs as an adaptive immune system, it is yet unclear if CRISPRs are sufficiently relevant in highly diverse natural environments to be conserved solely based on that characteristic (24).

*E. coli*’s CRISPRs are in two pairs of loci, CRISPR1 and -2 and CRISPR3 and -4, located at 62 and 20 min on the chromosome, respectively (10, 38). Their diversity has been described in the 72 strains of the ECOR collection, which are representative of the species genetic diversity, and in a heterogeneous set of 34 strains for which the complete genomes were sequenced (10, 38). From the available data, it is not clear whether the CRISPRs can have a role *in natura* in the adaptation of the *E. coli* strains to mobile genetic elements, nor whether this is a function of their diverse lifestyles. The number of processed CRISPR transcripts in *E. coli* seems to be very low (5), and protection against phage infection and lysogenization has been observed only in laboratory-engineered strains, i.e., an *hns* global transcriptional repressor mutant (11, 29) or a strain overexpressing *LeuO*, a LysR-type transcription factor (40). Neither insertion of new spacers nor interference with targeted invaders by nonmanipulated CRISPR loci has been documented so far (27). Finally, genomic analysis showed that *E. coli* CRISPRs are small and remain unchanged for long periods of time, which is at odds with the dynamics of an immune system (38).

In order to get further insights into the role of CRISPRs in natural isolates of *E. coli*, we (i) analyzed the CRISPR diversity in a large, well-characterized collection of 263 *E. coli* strains exposed to various environments and isolated in the 1980s and the 2000s, as well as in 27 fully sequenced strains representative of the phylogenetic diversity of the species, and (ii) mapped this diversity onto the phylogenetic history of the strains assessed by multilocus sequence typing (MLST). This allows an unparalleled analysis of the dynamics of CRISPR loci in light of the evolutionary history of the strains within a species.
**MATERIALS AND METHODS**

Bacterial strains. Strains were isolated in the 1980s (32 strains) and the 2000s (21 strains) from humans (217 strains) or animals (46 strains) living in various regions of France and belong to our previously published collection (4, 14, 28, 32, 33). One hundred fourteen strains originated from feces under commensal conditions (48 and 46 human and animal strains, respectively). Among the animal strains, 38 were from mammals (22 from wild animals and 16 from farm animals) and 8 were from birds (2 from wild birds and 6 from farm birds). One hundred forty-four strains were isolated in extraintestinal infections, all from humans (77 in urinary tract infections, 46 in septicemias, and 21 in miscellaneous infections). Five strains were isolated in extraintestinal colonization situations in humans. A detailed list of the strains and their main characteristics is given in Table S1 in the supplemental material. A set of 27 strains, for which the complete genomes were sequenced, representative of the 6 main phylogenetic groups was also included in the study: K-12 MG1655, K-12 W3110, K-12 DH10B, Sakai, EDL933, CFT073, 1A39, UTB9, 536, APECO1, UMN026, 55099, ED1a, SS88, 1A1, ECE4115, HS, E24377A, SMS35, E2348/69, ATCC 8739, SE11, BW2592, B842, BL21(DE3), BL23, and TW14359.

MLST analysis. MLST was performed using the Pasteur Institute schema based on the partial sequences of 8 genes (HDNA, iclA, putB, putF, putP, trpA, trpB, and uidA) as described previously (20). The molecular phylogeny of the strains has been explored by the construction of multiple sequence alignments of the concatenated genes with MUSCLE v3.6 (21). After alignment, the phylogenetic tree was reconstructed using the maximum likelihood method implemented in the PhyML program with a GTR substitution model (16) (see Methods in the supplemental material). Reliability for internal branching was assessed using the aLRT test (1). The evolutionary distance between two strains was defined as the sum of the lengths of all branches connecting them in the phylogenetic tree. The topology of this tree is congruent with previous whole-genome phylogenetic analyses of E. coli (e.g., see reference 37). The B2 and D subgroups were defined as in references 23 and 9, respectively. Of note, D subgroups VII, VIII, and IX correspond to the F phylogenetic group (20).

**CRISPR amplification and sequencing.** CRISPR loci were amplified by PCR followed by sequencing. The primers used were designed following the alignment of conserved sequences of available E. coli complete genomes flanking the four distinct CRISPR loci. CRISPRs were amplified using the Taq polymerase Expand High Fidelity™ (Roche Diagnostics), with primers C1Fw and C1Rev (CRISPR1), C2Fw or C2GFw and C2Rev (CRISPR2), C3Fw and C3Rev (CRISPR3), and C4Fw and C4Rev (CRISPR4) (see Fig. S2 in the supplemental material; also see Table S2 for the sequences of the primers). The PCR program was an initial denaturation step for 2 min at 94°C and then 94°C for 30 s, 55°C for 30 s, and 72°C for 1 min 30 s for 25 cycles.

To compare the spacer contents of all strains in each CRISPR locus, the spacer repertoire relatedness was measured for all pairs of strains as the fraction of shared similar spacers in the smallest CRISPR (34). Thus, when considering two strains, A and B, this corresponds to dividing the number of spacers common to strains A and B by the minimal number of spacers in A or B. Two spacers with at least 95% identity and less than 10% difference in sequence length were considered similar. As the first spacer of CRISPR1 and the first spacer as well as the following spacers S (small) and L (large) of CRISPR2 were conserved among the strains, they were not considered in the measurement of the spacer relatedness. Spacers S is smaller than the mean of the spacers whereas spacer L is a 454-bp sequence. Spacer repertoire relatedness was used to compute a distance matrix of all pairs of genomes with a CRISPR containing more than 5 spacers. The smaller CRISPRs were excluded because they carry too little signal to be of use for this analysis and are likely to be mere vestiges of previously functional CRISPRs. This matrix was then used to calculate a phylogenetic tree using the BIONJ algorithm (15). The Markov clustering algorithm (13) was used to cluster and organize the CRISPRs into distinct groups. Thus, we defined 15 and 8 large (i.e., with at least 4 strains) spacer relatedness groups in CRISPR1 and CRISPR2, respectively.

RESULTS AND DISCUSSION

The global organization of the CRISPRs within the phylogenetic structure of the species. In the large panel of 263 natural isolates belonging to all known E. coli phylogenetic groups and subgroups and having various lifestyles, we amplified and sequenced the four CRISPR loci. To this data set, we added the CRISPR sequences of 27 fully sequenced E. coli strains. Totals of 2,331, 1,829, 485, and 183 spacers were found in CRISPR1, -2, -3, and -4, respectively, of which 610, 497, 92, and 72, respectively, were distinct (see Table S3 in the supplemental material). Contrasting occupancy of the CRISPR loci was observed in that CRISPR1 and CRISPR2 exhibit numerous repeats, whereas CRISPR3 has mostly 1 to 5 repeats and CRISPR4 does not have repeats except in rare cases (Fig. 1). Closer examination of the data in the phylogenetic framework of the species indicates a strong dependence of the phylogenetic group on the CRISPR content. Within CRISPR1, the median number of repeats is between 12 and 14 according to the phylogroup (minimum, 2 to 5; maximum, 22 to 24), except for the B2 strains, which are completely devoid of repeats. CRISPR2 exhibits a similar pattern with a median number of repeats of 10 to 14 (minimum, 2 to 5; maximum, 25 to 32) for all phylogenetic groups except the B2 group, which has a median number of repeats of 4 (minimum, 2; maximum, 4). Within CRISPR3, almost all the strains, whatever their phylogroup, have a median number of two repeats. The few exceptions are 16 (17%) B2 phylogroup strains that exhibit more repeats, with a maximum of 18 repeats. Lastly, all strains lack repeats within CRISPR4, except for 14 (15%) B2 phylogroup strains that have more than 5 repeats, with a maximum of 21. These latter strains all have a large CRISPR3, thereby showing agreement with the idea of a joint dynamic between CRISPR3 and CRISPR4, which was suggested but not statistically tested before because of a lack of sufficient data (38). Graphical representation of the spacers across CRISPR1 to 4 is given in Fig. S3 to S5.

Two subtypes of cas genes have been reported in the E. coli species: the Ypest subtype, present only in the B2 phylogroup strains, and the Ecoli subtype, present in the remaining strains of the species (10, 38). We confirmed this association and observed a link between the number of repeats and the presence of cas gene subtypes. In the B2 strains, the absence of CRISPR1 is associated with the absence or with a relic of CRISPR2 and with the deletion or truncation of at least one associated cas gene of the Ecoli type, cas2. As a result, no CRISPR1 amplification was obtained using a consensus reverse primer designed based on the cas2 gene (see Fig. S2 in the supplemental material). Out of the 193 strains of the other phylogenetic groups, only 15 lack CRISPR1 and they all also lack the Ecoli-type cas genes. For these 15 strains, with the

Blastn was used for similarity searches between CRISPR spacer sequences and the 834 complete prokaryote genomes, 1,725 complete plasmid genomes, and 522 virus genomes available in GenBank. Only matches showing an E value of <1.10^-10 and less than 10% difference in sequence length were retained; matches to sequences found within CRISPR loci were ignored.

**Nucleotide sequence accession numbers.** Sequences of the CRISPRs from the 263 natural isolates have been deposited in GenBank with accession numbers JF495780 to JF496196.
exception of one strain having no CRISPR2, the CRISPR2 number of repeats was low, as for the B2 strains.

In strains lacking CRISPR4, we amplified the region of CRISPR3 with primers designed based on core genes flanking the loci clpA and infA (see Fig. S2 in the supplemental material), thereby showing the absence of the associated Ypest cas genes. Conversely, by obtaining amplifications with primers designed upward and downward from the Ypest-type cas genes, in cas1 and cys4 (Fig. S2), we showed that strains with CRISPR4, all but one of them from the B2 group, carried the Ypest-type cas genes. The association between cas genes and the size of the CRISPR loci suggests that the presence of functional cas genes is essential for active incorporation of CRISPR spacers. In their absence, we speculate that CRISPRs decrease in size by successive sequence deletions that are not compensated for by the integration of new spacers.

A striking feature of E. coli CRISPR3 is that some spacers match the cas genes of the Ypest subtype. Since Ypest cas genes are required for the function of CRISPR3 and CRISPR4, when these spacers are present, the cas genes are absent and the CRISPR shows even lower diversity. These elements have been called anti-CRISPR systems to highlight their putative role in inhibiting functional CRISPRs, and they might be involved in the manipulation of mobile elements with CRISPRs by the bacterial host (38). We performed an exhaustive search of protospacers in the CRISPRs of the studied strain collection. In all the strains having fewer than 6 repeats at CRISPR3, the majority of spacers match Ypest-subtype cas genes. Conversely, B2 group strains with CRISPR4 and a Ypest-subtype cas gene do not have a CRISPR3 spacer matching cas genes. No cas gene match was observed in the other CRISPRs.

These results, in agreement with previous analysis on a smaller number of strains (10, 38), show a strong footprint of the phylogeny on the CRISPR loci, with an almost complete lack of CRISPRs in the B2 phylogenetic group strains. This fact is striking, as the B2 phylogenetic group is a major group within the E. coli species. B2 group strains are frequently isolated under commensal conditions (36) and represent the major source of extraintestinal infections (28). It has been proposed that the B2 phylogenetic group is basal in the species (22). As CRISPR1 and -2 are present in Escherichia fergusonii, the most parsimonious scenario is a loss of the Ecoli cas gene operon in the B2 branch, followed by loss of CRISPR1 and -2 activity. Furthermore, as E. fergusonii also has an anti-CRISPR (CRISPR3), our data argue for an ancestral acquisition of the anti-CRISPR, followed in some cases within the B2 group by a loss of these elements, allowing the acquisition of CRISPR4 with Ypest cas genes by horizontal gene transfers. Such a scenario is supported by the presence of phylogenetic incongruence at this locus (37), as previously suggested (38).

**Spacer repertoire relatedness in the phylogenetic context.**

To have a better appreciation of the CRISPR content in relation to the phylogenetic history of the strains, we plotted for each pair of strains the spacer repertoire relatedness against the evolutionary distance in the MLST tree (see Materials and Methods). CRISPRs with fewer than 6 repeats were considered residual, and since they might be nonfunctional, they were removed from the subsequent analyses. Figure 2 illustrates the results obtained from the analysis of the spacers of CRISPR1. A high relatedness is observed when the phylogenetic distance is small (≤0.005). At longer distances, we observe a rapid decrease of spacer relatedness (Fig. 2A and B). Using various molecular clocks (22, 30, 41), it can be estimated that an MLST distance of 0.005 corresponds to hundreds of thousands to 5 million years. We found similar trends in the analysis of CRISPR2 spacer relatedness (data not shown). We then studied in more detail the distribution of spacer repertoire relatedness among the 106 strains having a phylogenetic distance of

---

**FIG. 1.** General features of the 4 distinct CRISPR loci in the E. coli species. In the five major phylogenetic groups of the species, namely, A, B1, B2, D, and F, we have represented the proportion of strains lacking a CRISPR in light gray, the proportion of strains containing a residual CRISPR in medium gray, and the proportion of strains having a CRISPR which we consider putatively functional in black (i.e., with a number of repeats above 5). This analysis was done for each of the 4 CRISPRs previously defined in reference 38. Schematic phylogenetic relationships between strains are represented on the left of the figure; E. fergusonii was used as an outgroup; the sizes of the triangles are proportional to the number of strains in each phylogenetic group. Strains of the C group are not represented.

A high relatedness is observed when the phylogenetic distance is small (≤0.005). At longer distances, we observe a rapid decrease of spacer relatedness (Fig. 2A and B). Using various molecular clocks (22, 30, 41), it can be estimated that an MLST distance of 0.005 corresponds to hundreds of thousands to 5 million years. We found similar trends in the analysis of CRISPR2 spacer relatedness (data not shown). We then studied in more detail the distribution of spacer repertoire relatedness among the 106 strains having a phylogenetic distance of
0 to 0.005. This shows a particular distribution with two extreme behaviors: either the spacer content is highly similar (79% of the strains have similar spacer contents with 80 to 100% relatedness), or it is completely different (the other 21% of the strains) (Fig. 2B). There are very few cases where only a fraction of spacers coincide. A closer examination of the 22 strains with radically different spacers shows that in 12 strains the spacer content is identical to spacers found in other spacer repertoire groups, corresponding probably to intraspecies horizontal transfer of the CRISPR. In the remaining strains, we found a complete replenishing of the locus with unique spacers.

These results suggest that CRISPR spacers are rarely replaced but that, when they are, the replacements are radical, so that very few ancient spacers are left in place. Altogether, these data argue for a non-clock-like behavior of the turnover of spacers in CRISPRs, rather than a gradual change.

Furthermore, the evolutionary histories of the two pairs of CRISPRs are not independent but linked. When the spacer relatedness groups of more than 4 strains (i.e., those represented in Fig. 3) are considered, 94% of the strains that clustered together in the CRISPR1 group are also found together in the CRISPR2 group. The same result is observed for CRISPR3 and CRISPR4 (see Fig. S6 in the supplemental material).

**An example of fine-scale analysis: the E. coli C phylogenetic group.** We analyzed in more detail the data of the spacer repertoire relatedness of CRISPR1 and -2 from strains of a small clonal group of the species, group C, which is closely related to the B1 phylogenetic group (26) and diverged from the A/B1 group ancestor between 100,000 and 5 million years ago (22, 30, 41) (MLST distance of 0.005). This group is representative of the various E. coli lifestyles, encompassing commensal and pathogenic strains from both humans and animals. All the spacer repertoires of the clonal group C strains were clustered together in spacer groups that were named group 2 (Fig. 3). Five non-group C strains belong to group 2 of CRISPR1 and -2 (Fig. 3 and 4). It has been previously shown that new spacers are acquired in a polarized fashion, with new units being added close to the leader, at the right end of CRISPR1 and -2 (10, 38). Surprisingly, the most recently acquired spacer is highly persistent and thus shared by nearly all CRISPR1s (spacer 10) but also CRISPR2s (spacer 6) of the group C strains (Fig. 4). The order of the spacers is always conserved between the strains; the few differences concern the absence of some spacers, likely due to deletions mediated by DNA polymerase slippage on the repeats. In fact, the strains CRISPRs among distant strains have no similar spacers, leading to instability in the estimation of the branches near the root of the tree. All together, these results indicate strong phylogenetic inertia of the CRISPR loci at short distances and very weak inertia at long distances.

We then built a tree from a distance matrix of all pairs of genomes based on spacer repertoire relatedness for putatively active CRISPR1 and -2 (Fig. 3) and CRISPR3 and -4 (see Fig. S6 in the supplemental material). We used a Markov clustering algorithm (13) to cluster and organize the CRISPRs into distinct spacer relatedness groups. We found that the CRISPR clusters matched the phylogenetic groups: more than 90% of strains clustered in a given CRISPR group belong to the same phylogenetic group/subgroup. The very few exceptions are likely to correspond to horizontal gene transfers, as for example strain 690, which belongs to the A phylogroup but is clustered with the D phylogroup strains (Fig. 3). Also, strains 549 and R384, which belong to the A and B1 phylogroups, respectively, are clustered with phylogroup C strains (see below). On the other hand, the relationships between the groups of spacer repertoire relatedness do not follow the strain phylogeny (Fig. S1). This might be caused by ancient horizontal transfers. Alternatively, it might simply reflect the observation that CRISPRs between strains have no similar spacers, leading to instability in the estimation of the branches near the root of the tree. All together, these results indicate strong phylogenetic inertia of the CRISPR loci at short distances and very weak inertia at long distances.

**FIG. 2.** Spacer repertoire relatedness of CRISPR1. (A) Association between spacer repertoire relatedness of CRISPR1 and phylogenetic distance between each pair of genomes. For clarity, phylogenetic distances were distributed into equidistant bins (intervals of 0.01); the average and standard error of the spacer repertoire relatedness of each bin are indicated. (B) Magnification of panel A in the distance range of 0 to 0.01. Each axis was divided into 10 equidistant bins; the point size is proportional to the number of strains present in each bin.
that we studied from clonal group C have been exposed to various environments and each have a specific life history: they have been isolated over a 20-year period in several locations in France (Paris, Brittany, and the Pyrenees), from humans and animals (domestic and wild), and under commensal (feces or tracheal secretion) or pathogenic (urinary tract infection) conditions (see Table S1 in the supplemental material). The high spacer relatedness that we observed in these epidemiologically unrelated group C strains shows that among phylogenetically closely related strains (“closely” meaning hundreds of thousands of years) exhibiting various lifestyles, there is rarely acquisition of novel spacers.

This conservation of spacers among the strains exhibiting identical MLST patterns is also observed when considering strains from other spacer relatedness groups (see Fig. S3 and S4 in the supplemental material). A striking illustration is the uropathogenic strain UMN026, isolated in the United States in the 1990s and having exactly the same CRISPR1 and -2 spacers as those of 5 strains isolated more than 10 years later in various parts of France, either as commensals or as pathogens. All strains are indicated by their designation, the phylogenetic group and subgroup, the number of repeats, and the CRISPR group. esfe, E. fergusonii. Incongruent strains discussed in the text are indicated by red blots. Strains for which we have only the phylogroup determined by the Clermont et al. method (6) (see Table S1 in the supplemental material) are not represented in this figure.
these strains belong to the same MLST phylogenetic group, DIV (phylogenetic distance = 0) (Fig. S1).

To compare these results with those of a locus known to be under strong diversifying selection, we analyzed the rfb locus encoding the O antigen in the 10 strains of phylogroup C with spacer relatedness group 2 CRISPR1 and -2 (Fig. 4). We found that these strains all exhibit a distinct O antigen based on MboII digestion of the rfb long-range PCR product (7) (see Fig. S7 in the supplemental material), in accordance with the heterogeneity of O antigens in group C strains (8). This shows that at these evolutionary scales, genes under diversifying selection are indeed evolving fast, whereas CRISPRs are not.

The presence of CRISPRs and bacterial lifestyles. If CRISPRs evolve mostly in functions of local parasites, it could be relevant to compare the CRISPR contents according to the origins of the strains in a systematic way. We first compared the total numbers of repeats in commensal versus extraintestinal pathogenic strains, as these two categories of strains face distinct environments and adaptive challenges. A first global analysis shows that the number of repeats is significantly higher in commensal (mean, 21.2) than in pathogenic (mean, 16.2) strains (Fig. 5A). However, we have observed a very specific pattern of CRISPR in B2 phylogenetic group strains, and it is well known that extraintestinal strains belong mainly to phylogenetic group B2 (28). Therefore, we repeated the comparison, excluding the B2 group strains, and found no difference between commensal and pathogenic strains (Fig. 5B). The same

hols true when splitting the pathogenic strains into uropathogenic and septicemic strains (see Fig. S8A in the supplemental material). We also did not find any difference in CRISPR size according to association with human or animal hosts, whether or not the B2 strains were considered (Fig. S8B). It appears that, using our large panel of strains, there is no obvious link between the presence of CRISPRs and a specific lifestyle or origin.

Analysis of spacer contents. When the total number of distinct spacers in the 4 CRISPRs (1,278) is considered, only 88 (7%) are homologous to sequences in the data bank, excluding other CRISPR spacers (see Table S3 in the supplemental material). This extends previous results (10, 38) showing 12.5% and 8% of distinct spacers having protospacers, respectively. We also find that 12, 31, and 57% of the distinct spacers match nonprophagic chromosomal regions, plasmids, and phages, respectively (Table S3). A specialization is observed for CRISPR3 and -4, which match plasmids in strains having Ypest cas genes, and CRISPR3, which matches cas genes in strains lacking Ypest cas genes, as discussed before (10, 38).

We then looked to see if we could identify differences in the presence/absence of protospacers according to the lifestyle of the strains. No difference was observed between commensal and pathogenic strains in CRISPR1, -2, and -3 without Ypest cas genes (data not shown).

Concluding remarks. We observed in a large collection of natural isolates of E. coli a disparity in the repartition of CRISPR content, with strains of the main phylogenetic group B2 lacking CRISPRs, extending the previous analyses based on a smaller set of strains (10, 38). Deeper analysis showed a strong phylogenetic inertia at short distances and little signal at longer distances. These results urge caution on the issue of CRISPRs as epidemiological markers (16). First, CRISPRs are lacking in a major group of the species. Second, some very distant strains resemble one another in terms of their CRISPRs because of horizontal gene transfer. Third, most comparisons between strains show no similarity from which to draw evolutionary distances. However, CRISPR typing could be useful in association with MLST to differentiate strains from a single clonal group, as illustrated for the C group (Fig. 4).

It is difficult to conceive the E. coli CRISPRs as an active immune system against phages, as such a system should provide a selective advantage and thus diffuse in all the species. It has been shown previously that the sensitivity of an E. coli collection representative of the species diversity to a set of 59 coliphages (21) did not correlate with CRISPR content (10). Furthermore, no link was evidenced between a specific lifestyle and the CRISPR content. Lastly, almost all the strains possess an anti-CRISPR that prevents invasion by genetic elements containing functional CRISPRs (38). The overall congruence between the CRISPR trees and the evolutionary history of the strains is clearly not expected for an immune system involved in defense against infectious agents such as phages that are both numerous and extremely diverse. As a comparison, the two major “bastions of polymorphism” observed in the E. coli genome, where a high level of incongruence is observed with the strain phylogeny, correspond to the rfb and fim loci, coding for the O antigen and the Fim adhesin, respectively (37). These molecules are both immunogenic and subjected to important diversifying selection.
The present results reinforce the idea that, within the E. coli species, CRISPRs do not have characteristics of an active immune system in that their evolutionary patterns do not fit the quick and constant evolutionary pace of such a system. At this stage of understanding bacterium-CRISPR-mobile genetic element interactions, one cannot provide precise quantitative models of what such an immune system would be. Yet, given the huge diversity of phages and plasmids and the rapidity with which they can generate further diversity, one would expect an adaptive immune system to adapt very fast to new challenges. This does not fit our observations in E. coli. Our observations are in sharp contrast with those for Leptospirillum in acido-philic microbial biofilms, where CRISPR diversification is fast enough to promote individuality in otherwise nearly clonal isolates, a hallmark of population-level response to the rapidly changing selective pressure of phage predation (39). Rapid dynamics were also found in S. thermophilus in dairy cultures (3). The ecology of E. coli does not seem to have selected CRISPRs as a rapid way to acquire adaptive immunity against foreign DNA. This could be in agreement with the fact that the predatory viral-microbial dynamic is notably absent in the intestine (31), the primary habitat of E. coli. But, since CRISPRs are maintained in most E. coli lineages, this certainly suggests that CRISPRs have additional functions, such as recently shown for DNA repair (2).

ACKNOWLEDGMENTS

This work was partly funded by a grant from the Assistance Publique-Hôpitaux de Paris (Contrat d’Initiation à la Recherche Clinique 05 103) and by the Fondation pour la Recherche Médicale. M.T. and E.P.C.R. are supported by the CNRS and the Institut Pasteur. We are grateful to Olivier Tenaillon for helpful discussions during this work.

REFERENCES