Microbiology (2010), **156**, 1351–1361 **DOI** 10.1099/mic.0.036046-0

Diversity of CRISPR loci in *Escherichia coli*

C. Díez-Villaseñor, C. Almendros, J. García-Martínez and F. J. M. Mojica

Departamento de Fisiología, Genética y Microbiología, Facultad de Ciencias, Universidad de Alicante, E-03080, Spain

Correspondence
F. J. M. Mojica
fmojica@ua.es

Received 10 November 2009
Revised 17 December 2009
Accepted 29 January 2010

CRISPR (clustered regularly interspaced short palindromic repeats) and CAS (CRISPR-associated sequence) proteins are constituents of a novel genetic barrier that limits horizontal gene transfer in prokaryotes by means of an uncharacterized mechanism. The fundamental discovery of small RNAs as the guides of the defence apparatus arose as a result of *Escherichia coli* studies. However, a survey of the system diversity in this species in order to further contribute to the understanding of the CRISPR mode of action has not yet been performed. Here we describe two CRISPR/CAS systems found in *E. coli*, following the analysis of 100 strains representative of the species' diversity. Our results substantiate different levels of activity between loci of both CRISPR types, as well as different target preferences and CRISPR relevances for particular groups of strains. Interestingly, the data suggest that the degeneration of one CRISPR/CAS system in *E. coli* ancestors could have been brought about by self-interference.

INTRODUCTION

A novel prokaryotic immunity-like system (CRISPR/CAS) has been recently discovered (Barrangou *et al.*, 2007; Brouns *et al.*, 2008; Marraffini & Sontheimer, 2008), involving two main constituents: (i) clusters of regularly interspaced short palindromic repeats (CRISPR) and (ii) CAS (CRISPR-associated sequence) proteins. CRISPR, originally reported in Bacteria (Ishino *et al.*, 1987), and later on in Archaea (Mojica *et al.*, 1993), have been described as a novel family of short regularly spaced repeats (SRSRs) after an analysis of about 20 prokaryotic genomes from both domains (Mojica *et al.*, 2000). Repeats alternate with similarly sized spacers that derive from sequences (proto-spacers) of diverse origin, notably from mobile genetic elements (Bolotin *et al.*, 2005; Mojica *et al.*, 2005; Pourcel *et al.*, 2005). Various CRISPR/CAS systems that combine specific CRISPR types (Kunin *et al.*, 2007) and CAS repertoires (Haft *et al.*, 2005; Makarova *et al.*, 2006) have been established. Arrays of the same CRISPR are commonly followed by the leader (Jansen *et al.*, 2002; Mojica *et al.*, 2000), an AT-rich sequence typically located at the opposite edge with respect to a degenerated terminal repeat. The leader appears to promote transcription towards the repeats (Brouns *et al.*, 2008; Hale *et al.*, 2008; Lillestøl *et al.*, 2006, 2009; Marraffini & Sontheimer, 2008), generating the RNAs that constitute the molecular base of the interference action (Brouns *et al.*, 2008). For recent descriptions of the CRISPR/CAS systems, see Sorek *et al.* (2008) and van der Oost *et al.* (2009).

The first experimental contribution to unravelling the molecular mechanism of CRISPR processing came from *Escherichia coli* studies (Brouns *et al.*, 2008). Further analysis of this model organism is expected to contribute to CRISPR/CAS characterization. A comprehensive description of the systems of this species is of interest. At present, two arrays of the same 29 bp CRISPR motif (iap repeat) have been described in *E. coli* (Ishino *et al.*, 1987; Nakata *et al.*, 1989), one starting 24 bp from the *iap* 3′ end, and a second array at a distance of about 24 kb, downstream of the *ygcF* gene. Additionally, the presence of one or two copies of a different motif (Ypest repeat) has been reported in strains of the species (Haft *et al.*, 2005). The CRISPR classification proposed by Kunin *et al.* (2007) assigns iap repeats to type 2, and Ypest repeats to type 4 (hereinafter referred to in the text as CRISPR2 and CRISPR4 repeats, respectively). Leader sequences adjacent to both CRISPR2 arrays have been identified (Mojica *et al.*, 2009). A set of eight *E. coli* subtype CAS genes (namely, *cas2-cas1-cse3-cas5e-cse4-cse2-cse1-cas3*) has been detected next to the iap array (Haft *et al.*, 2005), defining the CRISPR2.1/CAS-E locus. The aforementioned study by Brouns and coworkers was performed with components of this locus. Here we describe the CRISPR/CAS systems of 100 *E. coli* strains, including 28 available genomes and the *E. coli* reference (ECOR) collection of 72 strains (Ochman & Selander, 1984). This collection is thought to represent the genetic

Abbreviations: CAS, CRISPR-associated sequence; CRISPR, clustered regularly interspaced short palindromic repeats; ECOR, *Escherichia coli* reference; MLEE, multilocus enzyme electrophoresis.

The GenBank accession numbers for the original sequences reported in this paper are GU260715–GU260889.

Five supplementary figures, with a supplementary reference, are available with the online version of this paper. The figures show variation of CRISPR sequences, location of spacers with homologues along CRISPR arrays of representative strains, phage susceptibility within the ECOR collection, generation of CRISPR2.2-3 repeats by recombination, and CRISPR2.1/CAS-E loci of *Shigella* spp.

diversity of the species, henceforth providing a reliable framework for assessing CRISPR variability. ECOR strains fall into four main groups (A, B1, B2 and D), plus a minor group (E), as defined by multilocus enzyme electrophoresis (MLEE) analysis (Selander *et al.*, 1986). We will refer to MLEE groups as a reference. Insights into the activity and dynamics of CRISPR loci are discussed, suggested by multiple observations from our study.

METHODS

PCR amplification. PCRs for CRISPR loci amplification of ECOR strains were performed with recombinant *Taq* polymerase from Invitrogen in a Mastercycler Gradient (Eppendorf) thermal cycler. Oligonucleotide primers were designed following the alignment of conserved sequences that flank repeat arrays of available genomes. Oligonucleotide C2.1F (5′-TGGTGAAGGAGTTGGCGAAGG-3′), hybridizing to the *iap* 3′-end, was used for CRISPR2.1 amplifications in combination with alternative reverse primers matching sequences in the *cas3–cysH* intergenic region (C2.1R1, 5′-TCTCTTCTTTGCAGGGAGGC-3′), the leader (C2.1R2, 5′-GTTGGTAGATTGTTGATGTGGA-3′; C2.1R3, 5′-GGTTGGTGGGTTGTTTTTATGG-3′) or the adjacent *cas2* gene (C2.1R4, 5′-GAAAATGTCCCTCCGCGCTTACG-3′). The *ygcE–ygcF* region, containing the other CRISPR2 arrays, was amplified with primers 5′-CGATCCAGAGCTGGTCGAATG3-3′ (*ygcE* 3′ end) and 5′-AGTGCTCTTTAACATAATGGATG-3′ (leader region). Oligonucleotide 5′-AGCACAAGGCGGAAGCAGC-3′, hybridizing to the *clpA* 3′ end, was used in combination with 5′-AATGCGCCTCGGACGATTGC-3′ (named C4.1-2R, upstream of *infA*) for amplification of CRISPR4.1-2, or with 5′-CGCGTTTGGAGTGGAGAATGG-3′ (5′ end of the associated *cas1*) for CRISPR4.1. CRISPR4.2 was amplified with 5′-GCGCAACCGCCACTATTCC-3′ (*csy4* gene) and C4.1-2R. Standard conditions with an annealing temperature of 55 °C were employed for amplifications of CRISPR2.1 involving C2.1R1, and similarly for CRISPR4. The touchdown method (Don *et al.*, 1991) was applied for the remainder. The PCR program for C2.1R2, C2.1R3 and C2.1R4 consisted of the following steps: (i) 94 °C for 2 min, (ii) 11 cycles of 94 °C for 15 s, annealing for 20 s, decreasing the temperature by 1 °C per cycle from 62 °C to 52 °C, 68 °C for 2 s, and 72 °C for 2 min, (iii) 35 cycles of 94 °C for 15 s, 60 °C for 20 s, 68 °C for 2 s and 72 °C for 2 min, and (iv) final extension at 72 °C for 10 min. PCRs for *ygcE–ygcF* were conducted with the following program: (i) 94 °C for 2 min, (ii) 11 cycles of 94 °C for 15 s, annealing for 20 s, decreasing the temperature by 1 °C per cycle from 58 °C, 68 °C for 10 s, and 72 °C for 3 min, (iii) 35 cycles of 94 °C for 15 s, 60 °C for 20 s, 68 °C for 10 s and 72 °C for 3 min, and (iv) 72 °C for 10 min.

Sequencing and sequence analysis. PCR products were purified with the QIAquick PCR purification kit (Qiagen) and sequenced with the Big Dye Terminator Cycle Sequencing kit in an ABI PRISM 310 DNA Sequencer following the manufacturer's instructions (Applied Biosystems). Additional CRISPR arrays were detected by searches with the BLASTN program (Altschul *et al.*, 1997) performed at the websites of the NCBI (http://blast.ncbi.nlm.nih.gov/Blast.cgi) and the Wellcome Trust Sanger Institute (http://www.sanger.ac.uk/Projects/Escherichia_Shigella/), as well as through the analysis of publicly available *E. coli* genomes (http://www.ncbi.nlm.nih.gov/sites/genome/) with a computer program designed by our group (Mojica *et al.*, 2000).

Spacer homologues (proto-spacers) were identified as sequences located outside CRISPR loci, showing at least 28 identities with spacers. Searches were performed with BLASTN run against the nr database at the NCBI website, with the parameters that the application automatically sets for short queries. The significance of the alignments was determined as previously described (Mojica *et al.*, 2005).

RESULTS

We analysed CRISPR/CAS systems of the ECOR collection and 28 *E. coli* genomes. Table 1 summarizes the most relevant features of the CRISPR loci found. The different layouts of CRISPR/CAS loci detected are illustrated in Fig. 1.

CRISPR2/CAS-E system

In addition to the two arrays of CRISPR2 repeats reported, hereinafter CRISPR2.1 (adjacent to the *iap* gene, in the *iap–cysH* region) and CRISPR2.3 (downstream of the *ygcF* coding sequence), we found a third array (CRISPR2.2) located downstream of *ygcE*, at 0.5 kb from CRISPR2.3. Examples in the *ygcE–ygcF* region of a single array (CRISPR2.2-3) and a complete absence of repeats were detected, replaced invariably in the latter case by a sucrose operon. CRISPR2.1 repeats are also absent in some ECOR strains and genomes analysed.

Amplicons from CRISPR2.1 loci were obtained for 70 ECOR strains, 50 of which have repeats (CRISPR2.1$^+$), varying from three to 30 units. Of these CRISPR2.1$^+$ strains, 42 also have *E. coli* subtype CAS genes (CAS-E) following the leader. Similar proportions were found in the 28 available genomes: 17 are CRISPR2.1$^+$/CAS-E$^+$, two are CRISPR2.1$^+$/CAS-E$^-$, and the remaining nine are CRISPR2.1$^-$/CAS-E$^-$.

PCR amplification of the *ygcE–ygcF* region gave products for all ECOR strains: 50 have CRISPR2.2 and CRISPR2.3, 17 have a single array, and five lack CRISPR in this region. Of the 28 sequenced genomes tested, 17 harbour both arrays, six carry CRISPR2.2-3 and five have no CRISPR. The repeat content of each cassette is quite different: whilst CRISPR2.2 invariably consists of three units, CRISPR2.2-3 usually has two (18 out of 23 strains), and CRISPR2.3 varies from two to 29 units.

A total of 536 CRISPR2.1 spacers, arranged in 40 combinations, referred to as alleles, were found in the 50 ECOR strains with repeats in that locus (Fig. 2). Additionally, 196 spacers were analysed in the sequenced genomes with CRISPR2.1 repeats (19 strains). Each of these genomes has a different CRISPR2.1 allele, and only that of strain 101.1 is represented in the ECOR collection. Taken together, of all 732 CRISPR2.1 spacers, 303 different sequences (unique spacers) were found. These spacers are arranged in 58 CRISPR2.1 alleles, resulting in 84 % diversity (proportion of alleles found in the 69 CRISPR2.1$^+$ strains).

In CRISPR2.3, 561 spacers arranged in 37 alleles were detected for the 50 ECOR strains carrying the array (Fig. 2). The subset of 17 sequenced genomes with that locus encompasses 206 spacers. Each genome has a

Table 1. Number of repeats and main features of CRISPR arrays from the *E. coli* strains analysed

Strain (MLEE)	CRISPR array						
	2.1	2.2	2.3	2.2-3	4.1	4.2	4.1-2
SMS-3-5	23†	¶	¶	¶	¶	¶	1
101.1	7‡	3	16	¶	¶	¶	2
H10407	4†	3	8	¶	¶	¶	2
K12	14†	3	7	¶	¶	¶	2
ATCC 8739	22†	3	29‖	¶	¶	¶	4
53638	10†	3	5	¶	¶	¶	4
IAI1	19†	3	22	¶	¶	¶	3
HS	10†	¶	¶	20	¶	¶	4
55989	3†	3	16	¶	¶	¶	2
SE11	24†	3	21	¶	¶	¶	2
E24377A	14†	3	26	¶	¶	¶	3
E22	12†	3	9	¶	¶	¶	2
E110019	11†	3	12	¶	¶	¶	2
B7A	5†	3	18	¶	15	20*	¶
B171	9†	3	5	¶	¶	¶	2
BL21	6‡	3	14	¶	¶	¶	2
ED1a	0	¶	¶	1	18	14*	¶
UTI89	0	¶	¶	2	9	7*	¶
APEC O1	0	¶	¶	2	6	7*	¶
S88	0	¶	¶	2	6	7*	¶
CFT073	0	¶	¶	2	¶	¶	2
F11	0	¶	¶	¶	¶	¶	3
536	0	¶	¶	¶	¶	¶	3
E2348/69	0	¶	¶	¶	¶	¶	1
IAI39	0	¶	¶	¶	¶	¶	3
UMN026	19†	3	10	¶	¶	¶	4
042	8†	3	3	¶	¶	¶	3
Sakai	5†	3	2	¶	¶	¶	2
EC1 (A)	18†	3	5	¶	¶	¶	2
EC2 (A)	6‡	3	7	¶	¶	¶	2
EC3 (A)	§	3	7	¶	¶	¶	2
EC4 (A)	3†	3	21‖	¶	¶	¶	4
EC5 (A)	7‡	3	7	¶	¶	¶	2
EC6 (A)	9†	3	10	¶	¶	¶	4
EC7 (A)	22†	3	7	¶	¶	¶	2
EC8 (A)	7‡	3	7	¶	§	§	§
EC9 (A)	§	3	6	¶	¶	¶	2
EC10 (A)	7‡	3	7	¶	¶	¶	2
EC11 (A)	13†	3	6	¶	¶	¶	2
EC12 (A)	7‡	3	7	¶	¶	¶	2
EC13 (A)	3†	3	7	¶	¶	¶	2
EC14 (A)	6‡	3	7	¶	¶	¶	2
EC15 (A)	15†	3	27	¶	¶	¶	5
EC16 (A)	10†	3	23	¶	¶	¶	5
EC17 (A)	14†	¶	¶	¶	¶	¶	2
EC18 (A)	20†	3	24	¶	¶	¶	2
EC19 (A)	21†‖	3	16	¶	¶	¶	2
EC20 (A)	22†‖	3	18	¶	¶	¶	2
EC21 (A)	22†‖	3	18	¶	¶	¶	2
EC22 (A)	16†	3	23‖	¶	¶	¶	3
EC23 (A)	0	¶	¶	2	¶	¶	2
EC24 (A)	3†	3	2	¶	¶	¶	1
EC25 (A)	7‡	3	7	¶	¶	¶	2
EC26 (B1)	30†	3	23	¶	¶	¶	2
EC27 (B1)	30†	3	19	¶	¶	¶	2
EC28 (B1)	14†	3	19	¶	¶	¶	2
EC29 (B1)	14†	3	26	¶	¶	¶	2
EC30 (B1)	14†	3	21	¶	¶	¶	2
EC31 (E)	10†	3	14	¶	¶	¶	3
EC32 (B1)	7†	3	18	¶	¶	¶	2
EC33 (B1)	7†	3	18	¶	¶	¶	2
EC34 (B1)	13†	3	22	¶	¶	¶	2
EC35 (D)	5†	3	11	¶	¶	¶	2
EC36 (D)	5†	3	14	¶	¶	¶	2
EC37 (E)	6†	3	5	¶	¶	¶	2
EC38 (D)	0	¶	¶	¶	¶	¶	3
EC39 (D)	0	¶	¶	¶	¶	¶	3
EC40 (D)	0	¶	¶	¶	¶	¶	3
EC41 (D)	0	¶	¶	¶	¶	¶	3
EC42 (E)	14†	3	9	¶	¶	¶	3
EC43 (E)	15†	3	14	¶	¶	¶	3
EC44 (D)	10†	3	4	¶	¶	¶	4
EC45 (B1)	28†	3	17	¶	¶	¶	2
EC46 (D)	7†	3	13	¶	¶	¶	2
EC47 (D)	17†	3	10	¶	¶	¶	5
EC48 (D)	0	3	4	¶	¶	¶	2
EC49 (D)	11†	¶	¶	20	¶	¶	2
EC50 (D)	6†	3	10	¶	¶	¶	2
EC51 (B2)	0	¶	¶	2	¶	¶	2
EC52 (B2)	0	¶	¶	2	¶	¶	2
EC53 (B2)	0	¶	¶	1	¶	¶	2
EC54 (B2)	0	¶	¶	2	¶	¶	2
EC55 (B2)	0	¶	¶	2	¶	¶	2
EC56 (B2)	0	¶	¶	2	¶	¶	2
EC57 (B2)	0	¶	¶	2	¶	¶	2
EC58 (B1)	5†	3	7	¶	¶	¶	2
EC59 (B2)	0	¶	¶	2	¶	¶	2
EC60 (B2)	0	¶	¶	1	¶	¶	2
EC61 (B2)	0	¶	¶	2	8	7*	¶
EC62 (B2)	0	¶	¶	2	8	7*	¶
EC63 (B2)	0	¶	¶	2	4	3*	¶
EC64 (B2)	0	¶	¶	2	¶	¶	3
EC65 (B2)	2	¶	¶	2	8	17*	¶
EC66 (B2)	0	¶	¶	2	¶	¶	2
EC67 (B1)	6†	3	10	¶	¶	¶	2
EC68 (B1)	11†	3	19	¶	¶	¶	2
EC69 (B1)	16†	3	5	¶	¶	¶	2
EC70 (B1)	10†	3	9	¶	¶	¶	2
EC71 (B1)	4†	3	9	¶	¶	¶	2
EC72 (B1)	11†	3	9	¶	¶	¶	2

†Array with adjacent CAS-E genes.
¶Not present.
‡Transposable element adjacent to the CRISPR array.
‖Transposable element within the CRISPR array.
*Array with adjacent CAS-Y genes.
§Not sequenced.

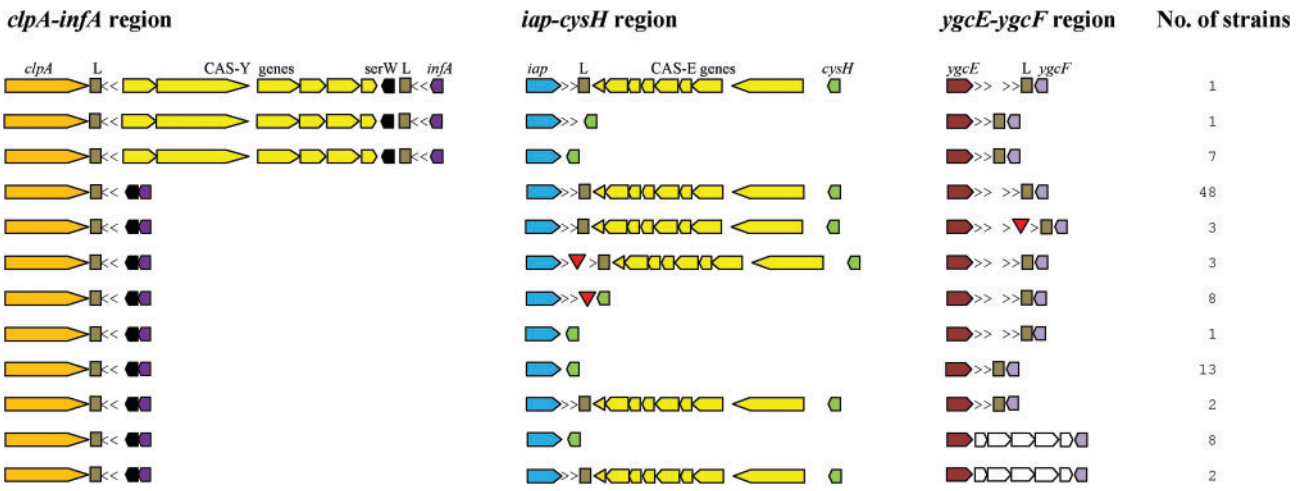


Fig. 1. Representation of the structural diversity of CRISPR regions. Genetic elements are arranged according to their relative position in the chromosome. Leader sequences (L), CAS and flanking genes (boxes pointing towards their direction of transcription) are identified and distinctly coloured. The sucrose operon is shown as empty boxes. Each CRISPR array is represented, irrespective of the number of repeats, by a pair of '>' symbols pointing towards the leader. Transposable elements are depicted as red triangles, either following two '>' symbols (when adjacent to a CRISPR array) or between them (when within a CRISPR array). The number of strains corresponding to each combination of CRISPR loci organizations is indicated.

different CRISPR2.3 allele, and only that of UMN026 is represented in the ECOR collection. Overall, of the 767 CRISPR2.3 spacers analysed, 298 are unique, arranged in 52 CRISPR2.3 alleles that result in 77.6 % allele diversity. The total number of unique spacers in the *ygcE–ygcF* region increases by two and 31 when CRISPR2.2 and CRISPR2.2-3 are considered, respectively.

CRISPR2 repeats vary to different degrees depending on the array and their position within it (Supplementary Fig. S1). We will refer to particular repeats as 'CRISPR array No.' – 'position within the array, numbers increasing towards the leader'. The two most frequent repeat variants give the CRISPR2 consensus CGGTTTATCCCCGCT-GGCGCGGGGAACWC. Main divergences are in the CRISPR2.2 array, the leader distal edge of CRISPR2.1 and, to a lesser extent, in the equivalent location of CRISPR2.3. Specifically, repeats in CRISPR2.2 differ from the consensus in up to 4 nt (2.2-2 and 2.2-3 repeats) or up to 10 nt (2.2-1). Repeat 2.1-1 shows over seven mutations, and repeat 2.3-1 invariably starts with T instead of C. Aside from 2.1-1 and 2.3-1, repeat versions that differ at one or two positions from the consensus are found in CRISPR2.1 and CRISPR2.3 with a lower degree of polymorphism, identical for both arrays (23 variants each). Interestingly, CRISPR2 repeat variants in arrays of the same strain are not necessarily linked, suggesting that their sequences are somehow influenced by the context of the repeat.

With respect to spacers, the frequency of incidence in the panel of strains analysed varies depending on the array considered and their relative location within it. We will refer to particular spacers as 'CRISPR array No.' – 'spacer No. as identified in Fig. 2'. Spacer 2.1-1 is present, with high identity (over 90 %), in 64 out of 68 CRISPR2.1$^+$ strains, located in all cases at the first position of the array (distal to the leader). In general, CRISPR2.1 and CRISPR2.3 spacers situated farther from the leader are the most frequently encountered, whilst those found in just one strain (strain-specific) are biased to the proximal end, and to a lesser extent to intermediate positions within the array (Fig. 2). The two spacers of CRISPR2.2 are present, with high identity (one occasional mutation), in all CRISPR2.2$^+$ strains.

CRISPR4/CAS-Y system

In addition to CRISPR type 2, we have detected up to two arrays of type 4 repeats in *E. coli* (Kunin *et al.*, 2007), located in the region between the *clpA* and *infA* genes (Fig. 1, Table 1). If only one array is present, we will refer to it as CRISPR4.1-2. When two are found, we will refer to the one adjacent to *clpA* as CRISPR4.1, and to the *infA* proximal one as CRISPR4.2. A set of typical Ypest-subtype (Haft *et al.*, 2005) CAS genes (CAS-Y) is situated between the two arrays. We amplified and sequenced the *clpA–infA* region of 71 ECOR strains. Four of them and five sequenced genomes have CAS-Y genes (CAS-Y$^+$), always flanked by CRISPR4.1 and CRISPR4.2, with the number of units varying from four to 18 and from three to 20, respectively. Interestingly, only one CAS-Y$^+$ strain (B7A) has CAS-E genes. Of 153 spacers analysed in these CAS-Y$^+$ strains, 100 are unique (65.4 %), with a similar proportion for both arrays (47 out of 73 in CRISPR4.1 and 53 out of 80 in CRISPR4.2). This situation contrasts with that of CAS-Y$^-$ strains, where the single array CRISPR4.1-2 has from one to five repeats, with a total of 10 unique spacers,

and whereas no variant has been found for CAS-Y$^+$ spacers, mutations are frequent in CAS-Y$^-$ strains (Fig. 2).

The majority of CRISPR4 repeats have the sequence TTTCTAAGCTGCCTGTACGGCAGTGAAC. Eighteen variants were found with up to six mismatches (Supplementary Fig. S1). The most different repeat within each CRISPR4 array lies at the distal end with respect to *clpA*, suggesting the existence of a leader at the opposite edge. Indeed, alignments of regions adjacent to the arrays revealed an AT-rich sequence with a 53 % identity at this side, comprising about 70 bp (data not shown). Similarly to CRISPR2, the highest diversity of CRISPR4 spacers was in the leader region (Fig. 2).

Spacer diversity

About half of the unique CRISPR2 spacers found (160 out of 303 in CRISPR2.1 and 173 out of 332 in the *ygcE–ygcF* locus) are present in several strains, located in equivalent relative positions. In contrast, of 100 unique CRISPR4 spacers in CAS-Y$^+$ strains, 85 (85 %) are strain-specific (Fig. 2 and Supplementary Fig S2).

There are significant differences between CRISPR arrays with respect to the ratio of spacers with homologues as compared with the number of unique spacers: 10.9 % (33/303) for CRISPR2.1, 7.9 % (24/332) for CRISPR2.3/CRISPR2.2-3, 27 % (27/100) for CRISPR4 arrays of CAS-Y$^+$ strains, and 90 % (9/10) for CRISPR4.1-2 arrays (CAS-Y$^-$ strains). Globally, of the 745 unique spacers found in the two CRISPR systems, 93 (12.5 %) are homologous to sequences (proto-spacers) in non-mobile elements (14), plasmids (27) and phages (52; see Supplementary Fig. S2). It is worth noting that among viral proto-spacers, 34 (65.4 %) correspond to a prophage in the genomes of E24377A and SE11 (Fig. 3). Although the general bias to phages is manifest for CRISPR2 (80.7 % of spacers matching sequences in the databases), CRISPR4 proto-spacers have a different prevalence, which also depends on the presence of CAS-Y genes: CAS-Y$^+$ strains have a preference for plasmid proto-spacers (20 out of 27; 74.1 %), and nine out of 10 CRISPR4 spacers of CAS-Y$^-$ strains are homologous to sequences in the absent CAS-Y genes (see Discussion).

Also noteworthy is the heterogeneous distribution of spacers with homologues throughout each CRISPR array, usually closer to the leader (Supplementary Fig. S2), which means that matches are found more frequently for the most recently acquired spacers.

In agreement with previous reports that support DNA molecules as both spacer donors and interference targets (Barrangou *et al.*, 2007; Brouns *et al.*, 2008; Lillestøl *et al.*, 2006; Marraffini & Sontheimer, 2008; Mojica *et al.*, 2009; Semenova *et al.*, 2009; Vestergaard *et al.*, 2008), given a common point of reference for orientation of spacers, proto-spacers match both strands of the carrier DNA molecule and any direction of transcription, and are found even in non-coding regions (see Fig. 3).

Owing to the large number of spacers that match the above-mentioned E24377A-SE11 prophage, distribution of proto-spacers was considered in this element (Fig. 3). Gene homology and genetic organization analysis of the region containing the proto-spacers revealed that the integrated prophage is a siphovirus–myovirus recombinant that combines P2-like and λ-like gene modules, expanding about 45 kb. Most proto-spacers (30 out of 34) fall within genes encoding proteins with conserved domains, even though such genes only form about 50 % of this region. It is also noticeable that the aforementioned spacers were found in 29 strains pertaining to the five MLEE groups, signifying a general propensity of the species to gain spacers from such sequences or elements.

CRISPR/CAS content versus strain features

We investigated possible connections between CRISPR/CAS content and characteristics of the carrier strain. The sensitivity analysis of ECOR strains to a set of 59 coliphages from worldwide sources (Kutter, 2009) did not correlate with CRISPR or CAS content (Supplementary Fig. S3). In general, there is no clear link between the number of repeats and source or host identity (data not shown). However, when focusing on strains without functional CRISPR/CAS systems, we found that none of the 15 ECOR strains of B2 has CAS-E genes, and the seven available *Shigella* spp. genomes lack any complete CAS operon.

DISCUSSION

Diversity and activity of the CRISPR/CAS systems

Two different CRISPR/CAS systems are found in *E. coli*. The presence of repeats and CAS genes of each type conform to 10 main layouts, apart from minor variations due to insertion of mobile elements (Fig. 1). The absence of CAS-Y genes, with the subsequent degeneration of the CRISPR4 arrays, and the presence of CAS-E together with three CRISPR2 arrays are the most frequent. Polymorphism is largely increased by the variety of CRISPR intervening sequences. More than 1500 spacers, including some previously described (Mojica *et al.*, 2009), have been analysed in this work. These spacers are arranged in a number of alleles that varies from one, in the case of the most conserved array (CRISPR2.2) to 58 (CRISPR2.1), leading to 77 combinations of CRISPR2 and CRISPR4 alleles. Moreover, among CAS$^+$ strains, about half of the detected spacers are unique, which is evidence of high activity for the two CRISPR systems. This activity is greater for CRISPR4/CAS-Y, given that the proportion of unique spacers is much higher in this system than that of CRISPR2/CAS-E (65 and 40 %, respectively). Diversity parameters for CRISPR2.1 and CRISPR2.3 are also very

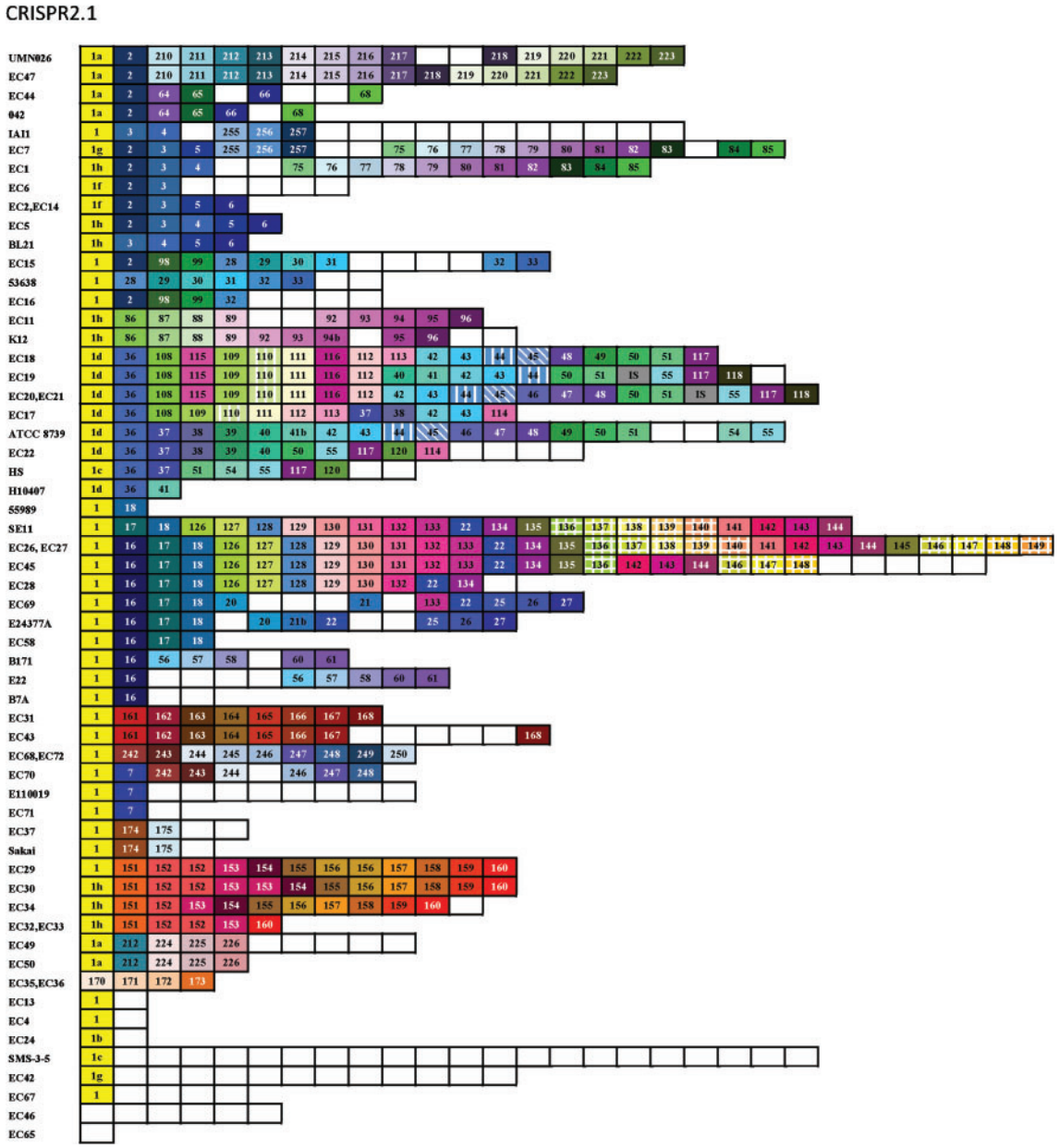
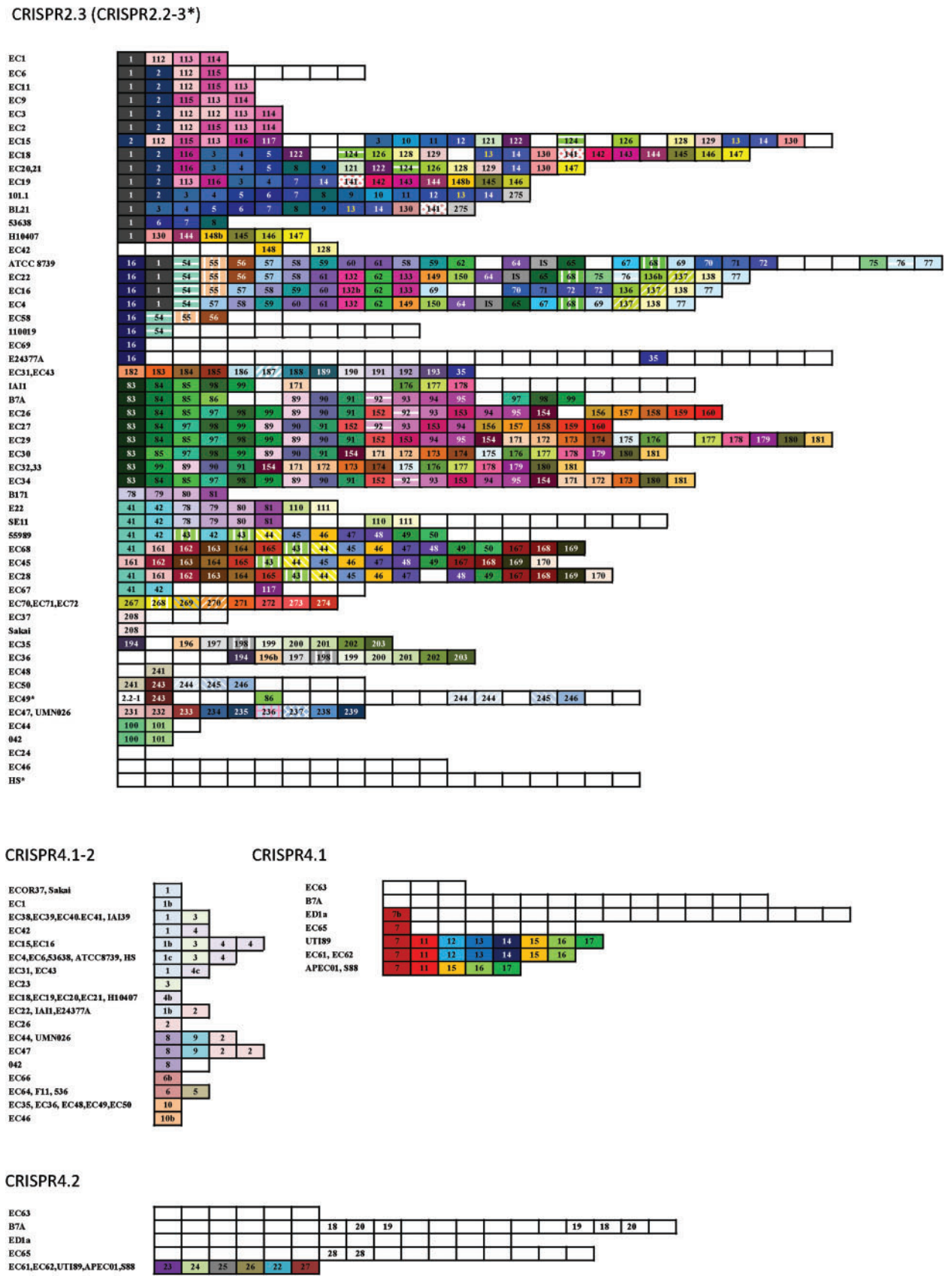


Fig. 2. Graphic representation of spacers in CRISPR arrays of *E. coli*. Arrays are equally oriented with respect to the leader (right). Spacers are represented by boxes, and those duplicated in the same strain or present in more than two strains are identified by a distinct number. The latter are also highlighted with a specific colour/pattern (white boxes correspond to strain-specific spacers). Homologous spacers (same origin and different sequence) have the same number but a different accompanying letter. Strains with two repeats in CRISPR2.2-3 have the same distinct spacer and, like those strains without spacers, have been omitted in the corresponding array. Strains with CRISPR2.2-3 instead of CRISPR2.3 are labelled with an asterisk. When more than five strains share a given allele, only one is quoted: CRISPR2.1 of EC5 is identical to that of EC8, EC10, EC12, EC25 and 101.1; CRISPR2.3 of EC2 is identical to that of EC5, EC7, EC8, EC10, EC12, EC13, EC14, EC25 and K12; CRISPR4.1-2 of EC1 is identical to that of EC2, EC3, EC5, EC7, EC9, EC10, EC11, EC12, EC13, EC14, EC17, EC25, 101.1, K12, BL21 and 110019; CRISPR4.1-2 of EC23 is identical to that of EC51, EC52, EC53, EC54, EC55, EC56, EC57, EC59, EC60 and CFT073; CRISPR4.1-2 of EC26 is identical to that of EC27, EC28, EC29, EC30, EC32, EC33, EC34, EC45, EC58, EC67, EC68, EC69, EC70, EC71, EC72, B171, E22, SE11 and 55989.

Fig. 2. cont.



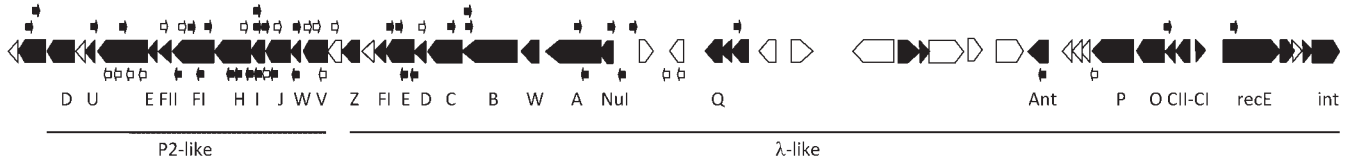


Fig. 3. Location of proto-spacers along the E24377-SE11 prophage. Genes are shown as boxes pointing in the direction of transcription, and proto-spacers as small arrows pointing towards the corresponding leader. Genes encoding proteins with conserved domains are filled, and those with high identity to known P2 or λ ORFs are identified underneath. Proto-spacers are labelled according to the degree of identity with spacers: filled when the identity is at least 90 % and empty for lower percentages.

close, indicating a similar rate of spacer turnover. In contrast, the strict conservation of CRISPR2.2 shows a lack of activity (Horvath *et al.*, 2008), which could be related to the absence of a leader and/or to the degeneration of its repeats. In this context, it is remarkable that, as a general rule, the degree of divergence between adjacent CRISPR repeats correlates with the conservation of the intervening spacers among strains. Indeed, CRISPR2.2 spacers and those adjacent to the terminal degenerate repeat of CRISPR2.1 and CRISPR4 arrays are the most frequently encountered. Moreover, the spacer next to the less degenerate 2.3-1 repeat is conserved to a lesser extent. These data indicate that repeats play a fundamental role in spacer turnover. Specific nuclease or integrase activities could recognize the canonical repeat sequence. Alternatively, a base-pairing mechanism could be involved. In this sense, recombination between repeats occurs as suggested for other species (Horvath *et al.*, 2008; Lillestøl *et al.*, 2006). Apart from duplications, there are rearrangements of spacers in CRISPR4.2 and CRISPR2.3 of strain B7A (see Fig. 2), and at least some CRISPR2.2-3 arrays could have been generated by recombination between CRISPR2.2 and CRISPR2.3 (Supplementary Fig. S4). For instance, the first repeat and spacer of CRISPR2.2-3 in EC49 are identical to those of CRISPR2.2 in EC50, its closest ECOR relative, and four CRISPR2.3 spacers of the latter strain are in the CRISPR2.2-3 array of EC49 (see Fig. 2).

CAS depletion

In good agreement with a functional correlation between CAS and CRISPR, the lack of CAS-E genes in B2 and D strains is invariably linked to the absence of CRISPR2.1 repeats. Moreover, the number of CRISPR2 repeats in CAS-E$^-$ strains of group A and that of CRISPR4 in CAS-Y$^-$ strains is reduced with respect to their CAS$^+$ closest relatives.

Remarkably, in all strains where CRISPR4-associated genes are absent, at least one spacer matching CAS-Y sequences is present in the CRISPR4.1-2 array, and moreover, no CAS-Y$^+$ strain contains spacers homologous to such sequences. Spacers specifically determine the targets of CRISPR-mediated immunity (Barrangou *et al.*, 2007; Brouns *et al.*, 2008; Marraffini & Sontheimer, 2008), most likely after a base-pairing recognition of the complementary sequence. This strongly suggests that the acquisition of CAS-Y-derived spacers in an ancestor with a functional system led to an eventual selection of derivative cells deleted for the corresponding CAS-Y targets, as the result of a CRISPR4/CAS-Y self-interference guided by the new spacers.

The absence of CRISPR2/CAS-E is notable in *Shigella* and B2 strains. However, such a similarity cannot be due to a common origin, since *Shigella* species derive from lineages that are independent of B2 (Fricke *et al.*, 2008; Ogura *et al.*, 2009; Pupo *et al.*, 2000; Turner *et al.* 2006). Moreover, the closely related ECOR strains carry a complete CRISPR2/CAS-E system (our unpublished results) and the deletions in CRISPR2.1/CAS-E are unrelated, entailing distinct genes and sequences (Supplementary Fig. S5). These data indicate that B2 and the different *Shigella* species have lost CRISPR2/CAS-E activity as the result of convergent evolution driven by some common circumstance that makes the activity unnecessary. In this context, it is notable that both groups reach higher population levels in restricted habitats, i.e. the colonic and rectal mucosa of humans in the case of *Shigella*, and the meninges and urogenital tract in the case of most B2 strains (Bingen *et al.*, 1998; Boyd & Hartl, 1998; Picard *et al.*, 1999). The reduced diversity and number of phages in these environments could explain the lack of CRISPR activity. In good agreement with a low incidence of challenging phages, there is an outstanding prevalence of plasmids as donors of CRISPR4 spacers in B2 strains with CAS-Y genes. Moreover, this bias does not relate to a possible preference of the CRISPR4 system, as the four CRISPR4 proto-spacers found in the only non-B2 Cas-Y$^+$ strain analysed are in phages, and the only homologue to CRISPR2 spacers of B2 strains is within a plasmid (Supplementary Fig. S2). Thus, it seems that bacteriophages are the main spacer source for most *E. coli* strains, which is expected as a result of the positive selection of immunized populations exposed to frequent

infections. However, when the challenge by viruses decreases, CRISPR become less fundamental for survival, although still relevant for limiting transmission of other foreign elements.

Insertion of new spacers

Preferential insertion of new spacers at the leader-proximal end of CRISPR arrays and sporadic replacements at the central region have been demonstrated in *Streptococcus thermophilus* (Barrangou *et al.*, 2007; Deveau *et al.*, 2008; Horvath *et al.*, 2008). Accordingly, we have found strain-specific spacers (expected to be the most recently acquired) mainly at the leader edges of CRISPR2 and CRISPR4 cassettes, and also, to a lesser extent, in inner positions (Fig. 2). Insertion of new spacers at the leader terminus is further supported by the occurrence in this region of homogeneous tracts of repeat variants (see Supplementary Fig. S1), suggesting that the pre-existing CRISPR units adjacent to the insertion site determine the sequence of the incoming repeat. This observation concurs with the duplication of the terminal CRISPR, perhaps by a transposition-like mechanism, upon adjacent insertion of a new spacer, as suggested by van der Oost *et al.* (2009).

Targets of CRISPR/CAS systems

After the reported interference of the *E. coli* CRISPR2/CAS-E system against target λ virus (Brouns *et al.*, 2008), and similar to the correlation found between the number of spacers and the phage resistance of *S. thermophilus* (Bolotin *et al.*, 2005), a lower susceptibility to infection would be expected in those ECOR strains that harbour the most complex CRISPR systems. However, no such connection was seen when susceptibility to a set of coliphages was considered. This could be explained by the requirement for specific spacers that match the invader DNA (Barrangou *et al.*, 2007; Brouns *et al.*, 2008; Marraffini & Sontheimer, 2008), and also by the existence in *E. coli* of alternative defence mechanisms that mask the action of CRISPR. In this context, after extensive screening, we have not detected any spacer incorporation in survivors of susceptible *E. coli* strains exposed to phages (our unpublished results), indicating that the proportion that become resistant to infection by the insertion of new spacers is several orders of magnitude lower than the incidence of resistance by other means.

CRISPR action on DNA implies that the identity of the target regions will be unrelated to their expression or the relevance of encoded products (Brouns *et al.*, 2008; Marraffini & Sontheimer, 2008; Semenova *et al.*, 2009). Conversely, E24377A-SE11 prophage proto-spacers are mainly within genes that encode proteins with conserved motifs, a situation also reported for the archaeal virus SIRV1 (Vestergaard *et al.*, 2008). This could be explained by an origin of the corresponding spacers from closely related phages. The lack of sequences from such donors hinders finding spacers that match the less conserved genes. Our data indicate that the E24377A-SE11 chimeric prophage is the closest known relative to the phages that more frequently challenge *E. coli* CRISPR systems. In the case of a similar induction of the spacer uptake process by any cell invader (Mojica *et al.*, 2009), such viruses would correspond to the most abundant coliphages in nature, although the possibility of a favoured interplay of certain spacers, phages and CRISPR systems cannot be dismissed (Vestergaard *et al.*, 2008).

The higher incidence of matches found for the most recently acquired spacers (Supplementary Fig. S2) could be explained by degeneration of the oldest sequences (spacers or proto-spacers), or by a CRISPR-driven selection of genetic elements that lack sequences identical to spacers. Several studies (Andersson & Banfield, 2008; Heidelberg *et al.*, 2009; Held & Whitaker, 2009) concur with the latter alternative, showing a substantial variation in viral sequences targeted by CRISPR. Moreover, a gradual variation in the sequence of spacers sharing a common origin was not detected, with the exception of sporadic differences. Consequently, according to our results, sequences in databases correspond to relatively modern genetic elements that differ greatly, at least in their proto-spacers, from ancient spacer donors. This would partially explain the low proportion of spacers with homologues, whereas the existence in nature of a great diversity of unknown mobile elements would account for the remaining non-matching sequences.

Conclusion

Although two different CRISPR can be found in *E. coli*, only one of the analysed strains harbours both sets of associated CAS genes, suggesting that, in general, one active system suffices to meet the CRISPR-based immunity demands of the species. P2-like and λ-like, or related chimeric phages, are the most frequent CRISPR targets among known sequences. As in other species, new spacers seem to be incorporated mainly at the leader-proximal end of functional arrays. Spacer maintenance correlates with degeneration of the adjacent repeats, likely to be involved in spacer turnover. In agreement with *E. coli* diversity, the heterogeneity of CRISPR/CAS content and spacer identity is substantial. Nevertheless, conservation is also observed at different levels, allowing the use of CRISPR in epidemiology, typing and evolution studies of the species.

ACKNOWLEDGEMENTS

The sequence data for *E. coli* strains 042 and H10407 were generated by the Wellcome Trust Sanger Institute Pathogen Sequencing Unit, and can be downloaded from http://www.sanger.ac.uk/Projects/Escherichia_Shigella/. This work was financed by research grants from the Conselleria de Cultura, Educació i Ciència, Generalitat Valenciana (CTIDIB/2002/155) and the Ministerio de Educación y Ciencia (BIO2004-00523).

REFERENCES

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389–3402.

Andersson, A. F. & Banfield, J. F. (2008). Virus population dynamics and acquired virus resistance in natural microbial communities. *Science* **320**, 1047–1050.

Barrangou, R., Fremaux, C., Deveau, H., Richards, M., Boyaval, P., Moineau, S., Romero, D. A. & Horvath, P. (2007). CRISPR provides acquired resistance against viruses in prokaryotes. *Science* **315**, 1709–1712.

Bingen, E., Picard, B., Brahimi, N., Mathy, S., Desjardins, P., Elion, J. & Denamur, E. (1998). Phylogenetic analysis of *Escherichia coli* strains causing neonatal meningitis suggests horizontal gene transfer from a predominant pool of highly virulent B2 group strains. *J Infect Dis* **177**, 642–650.

Bolotin, A., Quinquis, B., Sorokin, A. & Ehrlich, S. D. (2005). Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. *Microbiology* **151**, 2551–2561.

Boyd, E. F. & Hartl, D. L. (1998). Chromosomal regions specific to pathogenic isolates of *Escherichia coli* have a phylogenetically clustered distribution. *J Bacteriol* **180**, 1159–1165.

Brouns, S. J., Jore, M. M., Lundgren, M., Westra, E. R., Slijkhuis, R. J., Snijders, A. P., Dickman, M. J., Makarova, K. S., Koonin, E. V. & van der Oost, J. (2008). Small CRISPR RNAs guide antiviral defense in prokaryotes. *Science* **321**, 960–964.

Deveau, H., Barrangou, R., Garneau, J. E., Labonté, J., Fremaux, C., Boyaval, P., Romero, D. A., Horvath, P. & Moineau, S. (2008). Phage response to CRISPR-encoded resistance in *Streptococcus thermophilus*. *J Bacteriol* **190**, 1390–1400.

Don, R. H., Cox, P. T., Wainwright, B. J., Baker, K. & Mattick, J. S. (1991). "Touchdown" PCR to circumvent spurious priming during gene amplification. *Nucleic Acids Res* **19**, 4008.

Fricke, W. F., Wright, M. S., Lindell, A. H., Harkins, D. M., Baker-Austin, C., Ravel, J. & Stepanauskas, R. (2008). Insights into the environmental resistance gene pool from the genome sequence of the multidrug-resistant environmental isolate *Escherichia coli* SMS-3–5. *J Bacteriol* **190**, 6779–6794.

Haft, D. H., Selengut, J., Mongodin, E. F. & Nelson, K. E. (2005). A guild of 45 CRISPR-associated (Cas) protein families and multiple CRISPR/Cas subtypes exist in prokaryotic genomes. *PLOS Comput Biol* **1**, e60.

Hale, C., Kleppe, K., Terns, R. M. & Terns, M. P. (2008). Prokaryotic silencing (psi)RNAs in *Pyrococcus furiosus*. *RNA* **14**, 2572–2579.

Heidelberg, J. F., Nelson, W. C., Schoenfeld, T. & Bhaya, D. (2009). Germ warfare in a microbial mat community: CRISPRs provide insights into the co-evolution of host and viral genomes. *PLoS One* **4**, e4169.

Held, N. L. & Whitaker, R. J. (2009). Viral biogeography revealed by signatures in *Sulfolobus islandicus* genomes. *Environ Microbiol* **11**, 457–466.

Horvath, P., Romero, D. A., Coûté-Monvoisin, A. C., Richards, M., Deveau, H., Moineau, S., Boyaval, P., Fremaux, C. & Barrangou, R. (2008). Diversity, activity, and evolution of CRISPR loci in *Streptococcus thermophilus*. *J Bacteriol* **190**, 1401–1412.

Ishino, Y., Shinagawa, H., Makino, K., Amemura, M. & Nakata, A. (1987). Nucleotide sequence of the *iap* gene, responsible for alkaline phosphatase isozyme conversion in *Escherichia coli*, and identification of the gene product. *J Bacteriol* **169**, 5429–5433.

Jansen, R., Embden, J. D., Gaastra, W. & Schouls, L. M. (2002). Identification of genes that are associated with DNA repeats in prokaryotes. *Mol Microbiol* **43**, 1565–1575.

Kunin, V., Sorek, R. & Hugenholtz, P. (2007). Evolutionary conservation of sequence and secondary structures in CRISPR repeats. *Genome Biol* **8**, R61.

Kutter, E. (2009). Phage host range and efficiency of plating. *Methods Mol Biol* **501**, 141–149.

Lillestøl, R. K., Redder, P., Garrett, R. A. & Brügger, K. (2006). A putative viral defence mechanism in archaeal cells. *Archaea* **2**, 59–72.

Lillestøl, R. K., Shah, S. A., Brügger, K., Redder, P., Phan, H., Christiansen, J. & Garrett, R. A. (2009). CRISPR families of the crenarchaeal genus *Sulfolobus*: bidirectional transcription and dynamic properties. *Mol Microbiol* **72**, 259–272.

Makarova, K. S., Grishin, N. V., Shabalina, S. A., Wolf, Y. I. & Koonin, E. V. (2006). A putative RNA-interference-based immune system in prokaryotes: computational analysis of the predicted enzymatic machinery, functional analogies with eukaryotic RNAi, and hypothetical mechanisms of action. *Biol Direct* **1**, 7.

Marraffini, L. A. & Sontheimer, E. J. (2008). CRISPR interference limits horizontal gene transfer in staphylococci by targeting DNA. *Science* **322**, 1843–1845.

Mojica, F. J. M., Juez, G. & Rodríguez-Valera, F. (1993). Transcription at different salinities of *Haloferax mediterranei* sequences adjacent to partially modified *Pst*I sites. *Mol Microbiol* **9**, 613–621.

Mojica, F. J. M., Díez-Villaseñor, C., Soria, E. & Juez, G. (2000). Biological significance of a family of regularly spaced repeats in the genomes of Archaea, Bacteria and mitochondria. *Mol Microbiol* **36**, 244–246.

Mojica, F. J. M., Díez-Villaseñor, C., García-Martínez, J. & Soria, E. (2005). Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. *J Mol Evol* **60**, 174–182.

Mojica, F. J. M., Díez-Villaseñor, C., García-Martínez, J. & Almendros, C. (2009). Short motif sequences determine the targets of the prokaryotic CRISPR defence system. *Microbiology* **155**, 733–740.

Nakata, A., Amemura, M. & Makino, K. (1989). Unusual nucleotide arrangement with repeated sequences in the *Escherichia coli* K-12 chromosome. *J Bacteriol* **171**, 3553–3556.

Ochman, H. & Selander, R. K. (1984). Standard reference strains of *Escherichia coli* from natural populations. *J Bacteriol* **157**, 690–693.

Ogura, Y., Ooka, T., Iguchi, A., Toh, H., Asadulghani, M., Oshima, K., Kodama, T., Abe, H., Nakayama, K. & other authors (2009). Comparative genomics reveal the mechanism of the parallel evolution of O157 and non-O157 enterohemorrhagic *Escherichia coli*. *Proc Natl Acad Sci U S A* **106**, 17939–17944.

Picard, B., García, J. S., Gouriou, S., Duriez, P., Brahimi, N., Bingen, E., Elion, J. & Denamur, E. (1999). The link between phylogeny and virulence in *Escherichia coli* extraintestinal infection. *Infect Immun* **67**, 546–553.

Pourcel, C., Salvignol, G. & Vergnaud, G. (2005). CRISPR elements in *Yersinia pestis* acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies. *Microbiology* **151**, 653–663.

Pupo, G. M., Lan, R. & Reeves, P. R. (2000). Multiple independent origins of *Shigella* clones of *Escherichia coli* and convergent evolution of many of their characteristics. *Proc Natl Acad Sci U S A* **97**, 10567–10572.

Selander, R. K., Caugant, D. A., Ochman, H., Musser, J. M., Gilmour, M. N. & Whittam, T. S. (1986). Methods of multilocus enzyme

electrophoresis for bacterial population genetics and systematics. *Appl Environ Microbiol* **51**, 873–884.

Semenova, E., Nagornykh, M., Pyatnitskiy, M., Artamonova, I. I. & Severinov, K. (2009). Analysis of CRISPR system function in plant pathogen *Xanthomonas oryzae*. *FEMS Microbiol Lett* **296**, 110–116.

Sorek, R., Kunin, V. & Hugenholtz, P. (2008). CRISPR – a widespread system that provides acquired resistance against phages in bacteria and archaea. *Nat Rev Microbiol* **6**, 181–186.

Turner, S. M., Chaudhuri, R. R., Jiang, Z. D., DuPont, H., Gyles, C., Penn, C. W., Pallen, M. J. & Henderson, I. R. (2006). Phylogenetic comparisons reveal multiple acquisitions of the toxin genes by enterotoxigenic *Escherichia coli* strains of different evolutionary lineages. *J Clin Microbiol* **44**, 4528–4536.

van der Oost, J., Jore, M. M., Westra, E. R., Lundgren, M. & Brouns, S. J. (2009). CRISPR-based adaptive and heritable immunity in prokaryotes. *Trends Biochem Sci* **34**, 401–407.

Vestergaard, G., Shah, S. A., Bize, A., Reitberger, W., Reuter, M., Phan, H., Briegel, A., Rachel, R., Garrett, R. A. & Prangishvili, D. (2008). Stygiolobus rod-shaped virus and the interplay of crenarchaeal rudiviruses with the CRISPR antiviral system. *J Bacteriol* **190**, 6837–6845.

Edited by: D. W. Ussery