Microbiology (2005), **151**, 2551–2561 DOI 10.1099/mic.0.28048-0

Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin

Alexander Bolotin, Benoit Quinquis, Alexei Sorokin and S. Dusko Ehrlich

Génétique Microbienne, Institut National de la Recherche Agronomique, Domaine de Vilvert, 78352 Jouy en Josas CEDEX, France

Correspondence
Alexander Bolotin
bolotine@jouy.inra.fr

Received 17 March 2005
Revised 25 May 2005
Accepted 30 May 2005

Numerous prokaryote genomes contain structures known as clustered regularly interspaced short palindromic repeats (CRISPRs), composed of 25–50 bp repeats separated by unique sequence spacers of similar length. CRISPR structures are found in the vicinity of four genes named *cas1* to *cas4*. *In silico* analysis revealed another cluster of three genes associated with CRISPR structures in many bacterial species, named here as *cas1B*, *cas5* and *cas6*, and also revealed a certain number of spacers that have homology with extant genes, most frequently derived from phages, but also derived from other extrachromosomal elements. Sequence analysis of CRISPR structures from 24 strains of *Streptococcus thermophilus* and *Streptococcus vestibularis* confirmed the homology of spacers with extrachromosomal elements. Phage sensitivity of *S. thermophilus* strains appears to be correlated with the number of spacers in the CRISPR locus the strain carries. The authors suggest that the spacer elements are the traces of past invasions by extrachromosomal elements, and hypothesize that they provide the cell immunity against phage infection, and more generally foreign DNA expression, by coding an anti-sense RNA. The presence of gene fragments in CRISPR structures and the nuclease motifs in *cas* genes of both cluster types suggests that CRISPR formation involves a DNA degradation step.

INTRODUCTION

Genomic sequencing revealed the existence of short direct repeats, 25–50 nucleotides long, interspaced by unique sequences of similar size, in bacterial and archaeal genomes (van Belkum *et al.*, 1998). These elements were proposed to form a family, known as CRISPRs, clustered regularly interspaced short palindromic repeats (Jansen *et al.*, 2002). CRISPR loci show a high level of polymorphism in different strains, and this property had been used for identification of clinical isolates of *Mycobacterium tuberculosis* (Groenen *et al.*, 1993; Kamerbeek *et al.*, 1997), *Streptococcus pyogenes* (Hoe *et al.*, 1999) and *Campylobacter jejuni* (Schouls *et al.*, 2003). Four genes, designated *cas1* to *cas4*, were found adjacent to CRISPR loci in different bacteria, suggesting a functional relationship with repeated sequences (Jansen *et al.*, 2002). Cas3 and Cas4 have motifs characteristic for helicases of superfamily 2, and the *recB* exonuclease family, respectively. However, the mechanism of CRISPR structure formation, and the biological function of CRISPRs are not known. Recently, it has been reported that CRISPR elements in *Yersinia pestis* acquire new repeats by preferential uptake of bacteriophage DNA (Pourcel *et al.*, 2005). We report here that many spacers have homology with known genes, most often derived from extrachromosomal elements such as phages and plasmids. We propose that the formation of CRISPR structures involves DNA fragmentation by *cas* genes, and that the apparent stability and widespread presence of CRISPRs in bacterial genomes may be due their protective function against foreign DNA invasion.

Abbreviations: COG, clusters of orthologous groups; CRISPR, clustered regularly interspaced short palindromic repeat.

The GenBank/EMBL/DDBJ accession numbers for the sequences reported in this paper are DQ072985–DQ073008.

Two tables showing the homology of the spacers with database sequences are available as supplementary material with the online version of this paper.

METHODS

Bacterial strains and DNA preparation. Bacterial strains used in this study are listed in Table 1. The phage-resistance profile of some strains has been obtained from different sources (Fayard 1993; M.-C. Chopin, Génétique Microbienne, Institut National de la Recherche Agronomique, personal communication). *Streptococcus thermophilus* cells were grown in M17 medium (Terzaghi & Sandine, 1975) supplemented with 1% (w/v) lactose at 42 °C in anaerobic conditions. Chromosomal DNA was extracted as described by Simpson *et al.* (1993).

CRISPR locus amplification and sequencing. The CRISPR locus was identified using the complete sequence of *S. thermophilus* CNRZ1066 (Bolotin *et al.*, 2004). The primers for PCR amplification were selected from regions 207 bp upstream (yc70,

Table 1. Homology of CRISPR spacers with microbial genes

CRISPR-carrying strain	Spacer homology*		
	ECE gene	ECE neighbour	ECE unrelated
Methanosarcina acetivorans C2A	–	1	–
Methanobacterium thermoautotrophicum delta H	7	–	–
Pyrobaculum aerophilum IM2	–	–	1
Sulfolobus tokodaii 7	4	–	1
Bacteroides fragilis YCH46	1	–	–
Clostridium tetani E88	1	3	3
Desulfovibrio vulgaris plasmid pDV	1	–	–
Listeria innocua Clip11262	2	–	–
'*Mannheimia succiniciproducens*' *MBEL55E*	1	–	1
Neisseria meningitidis Z2491	1	–	–
Porphyromonas gingivalis W83	–	–	3
Streptococcus agalactiae 2603V/R	1	2	–
Streptococcus agalactiae NEM316	1	–	–
Streptococcus pyogenes M1 GAS	8	–	–
Thermoanaerobacter tengcongensis MB4	1	–	–
Total	29	6	9

*ECE, extrachromosomal element.

TGCTGAGACAACCTAGTCTCTC) and 214 bp downstream (yc31, GCAACGACAGGAAGCGACCAAA) of the identified CRISPR locus. These primers were used for PCR with an annealing temperature 55 °C. The primers yb82, TACTCTCAAGATTTAAGTAACTGTAC, corresponding to the 'direct' strand of CRISPR, and yb83, GTACAGTTACTTAAATCTTGAGAGTA, corresponding to the 'reverse' strand were also used for testing of presence of repeats.

The complete sequences of PCR-amplified fragments corresponding to different CRISPR loci were obtained by primer walking. The sequences of these primers can be provided upon request to A. B. For sequence analysis the loci were divided into direct repeats and corresponding spacer sequences. These were named using acronyms composed of the strain identifier followed by a one letter descriptor ('d' for the repeats and 's' for spacers), and a number referring to the position of the element within the locus.

Spacer sequences were analysed for homology against themselves and the NCBI entrez nucleotide sequence database. CLUSTAL (Higgins & Sharp, 1989) software was used for sequence alignment.

Nucleotide and protein sequences. Nucleotide sequences of CRISPR loci of different bacterial strains were obtained from the NCBI (www.ncbi.nlm.nih.gov) database, corresponding accession nos are given in parentheses after the strain systematic name. cas genes sequences were obtained from NCBI, MBGD (http://mbgd.genome.ad.jp) and ERGO (http://ergo.integratedgenomics.com/ERGO) databases. Assignment of the genes was based on sequence conservation, the MBGD COG (clusters of orthologous groups) database and proximity on the genome, determined in most cases by using the 'pinned region' function of ERGO or the genome comparison facility of MBGD.

GenBank sequence accession nos. Nucleotide sequences were deposited in GenBank under the accession nos DQ072985–DQ073008.

RESULTS

The *S. thermophilus* CRISPR locus and associated genes

Analysis of *S. thermophilus* CNRZ1066 complete genome sequence (Bolotin *et al.*, 2004) revealed a locus of 2·7 kb, having a typical CRISPR organization (Mojica *et al.*, 2000): 42 repeats of 36 bp, GTTTTTGTACTCTCAAGATTTAAGTAACTGTACAAC, separated by unique sequence spacers of 30 bp (Fig. 1). The locus contains two duplications: one of five repeats and five spacers, and the other of one repeat and one spacer.

A gene localized upstream of the locus, *str0658*, encodes a homologue of the Cas1 protein, which is invariably associated with CRISPR repeats (Jansen *et al.*, 2002), but the three other CRISPR-associated genes, *cas2*, *cas3* and *cas4*, are absent. However, our analysis revealed that *str0658* and the two flanking genes, *str0657* and *str0659*, which we term herewith *cas5* and *cas6*, respectively, have homologues that are clustered in numerous bacterial species, belonging to widely divergent phylogenetic groups (Fig. 2a). Remarkably, the homologues of the *cas1* gene are associated either with the *cas2*, *cas3* and *cas4* homologues, or with the *cas5* and *cas6* homologues, but never with both (Fig. 2a). Furthermore, the *cas1* genes group in two clusters, *cas1A* and *cas1B* (Fig. 2b); the former is always associated with *cas2, cas3* and *cas4*, and the latter with *cas5* and *cas6* genes. Both types of gene (*cas1A* and *cas1B*) are present in some species (*S. pyogenes*, *Fusobacterium nucleatum*), but each clusters with other members of the corresponding subgroup

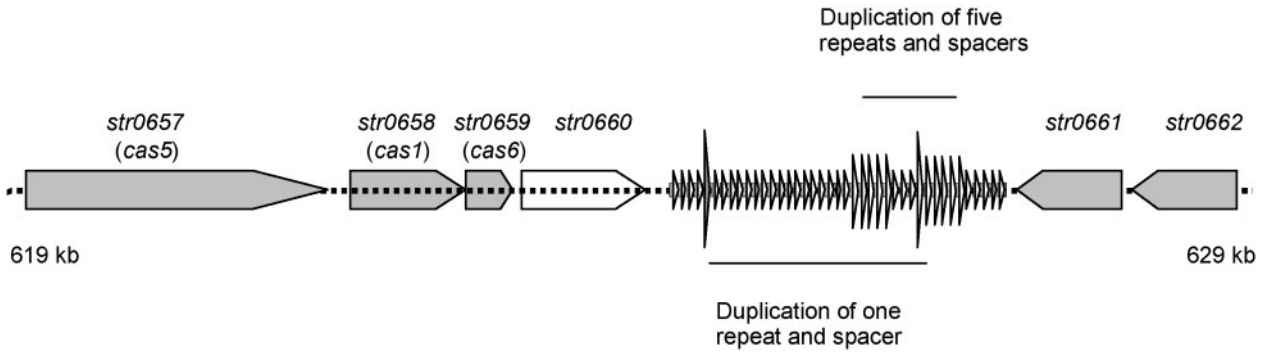


Fig. 1. The CRISPR locus in the *S. thermophilus* strain CNRZ1066. Repeats and spacers are shown as grey boxes and triangles, respectively, the duplicated spacers are shown as larger triangles, ORFs are represented as arrows (shaded for ORFs that have homologues in other genomes) and their designation in the *S. thermophilus* genome is given above the arrows. The numbers indicate the distance from the replication origin (in kb).

(Fig. 2b). Previous analysis of *cas1* genes, referred to as COG 1518 (Makarova *et al.*, 2002), placed the seven members of the *cas1B* subgroup that were available at the time on a clearly separated branch of the phylogenetic tree.

cas1B, *cas5* and *cas6* genes are localized downstream of the CRISPR cluster (illustrated in Fig. 2c for the genomes accessible in the ERGO database). A truncated repeat in the orientation opposite to that found within the CRISPR structure is often present upstream of the cluster (Fig. 2c). Clustering of the *cas1*, *cas5* and *cas6* genes (referred to as COG 1518, COG 3513 and COG 3512, respectively) in four bacterial genomes was previously noticed (Makarova *et al.*, 2002), but the association of the cluster with CRISPR structures was not reported.

The Cas5 family groups large proteins (>1100 aa) that carry an HNH motif present in various nucleases, including colicin E9, which causes cell death by introducing double-stranded breaks into DNA, and a number of restriction enzymes (Walker *et al.*, 2002; Maté & Kleanthous, 2004; Saravanan *et al.*, 2004). The Cas6 family groups short proteins (~100 aa) of high pI, the features found in the Cas2 (short) and Cas1 (high pI) families, but there is no sequence homology between Cas6 and other proteins in the databases.

CRISPR spacers have homology with extant genes

CRISPR loci were found close to *cas1* genes in about 50 of the 198 complete genomes available in the NCBI database. They totalled 2156 spacers, 44 of which are homologous to other genes, in addition to those from *S. thermophilus* and *Streptococcus vestibularis*, described in detail below. The 44 spacers are carried in 4 archaeal and 10 bacterial species, spanning a broad phylogenetic range (Table 1). Most (29 out of 44) are homologous to phage genes, even if they were identified on complete genomes (see Supplementary Table 1 available with the online journal for a detailed analysis). A striking case is that of *S. pyogenes*, which carries two short CRISPR structures close to the *cas1A* and *cas1B* genes, containing three and six spacers, respectively. All spacers of the former, and five of the latter are homologous to *S. pyogenes* prophage genes.

About a third of the 44 spacers are homologous with genes with no obvious extrachromosomal origin. Remarkably, six of these share homology with genes that reside in the vicinity of extrachromosomally derived genes, three share homology with genes of aberrant G+C content (up to 59 mol% G+C, compared to 47 mol% G+C over the entire *Porphyromonas gingivalis W83* genome) and two share homology with genes from regions where the gene order differs from that of phylogenetically close genomes (*Clostridium tetani E88*). Horizontal transfer could lead to the gene organization observed in all these cases.

CRISPR alleles in different *S. thermophilus* strains

To further examine the finding that CRISPR spacers can have phage origin, or more generally extrachromosomal origin, we analysed the CRISPR alleles of 22 *S. thermophilus* and 2 *S. vestibularis* strains. This was prompted by the consideration that phages of lactic acid bacteria are among the best characterized in respect of genome data (Brussow & Hendrix, 2002), and that lactic acid bacteria are exposed frequently to phage attacks under the conditions of dairy fermentations.

PCR reactions with primers homologous to two regions flanking the CRISPR structure (yc70 and yc31) yielded a single band of varying lengths for different strains (Fig. 3). The PCR products were sequenced, and the results are summarized in Table 2. The number of repeats varied between 10 and 51 for different CRISPR alleles. The repeats were strictly identical, with the exception of 38 out of a total of 632 (6 %). The slightly divergent repeats (less than 3 out of 36 bp difference) were mostly situated in the last

(a)

Organism	*cas1*	*cas2*	*cas3*	*cas4*	*cas5*	*cas6*
Aeropyrumperrix (*ape*)	ape1240		ape1232	ape1239		
Aquifex aeolicus (*aae*)	aq_369	aq_xxx	aq_371	aq_370		
Archaeoglobus fulgidus (*afuA*)	af1878	af1876	af1874	af1877		
Archaeoglobus fulgidus (*afuB*)	af2435	af2434		af2436		
Bacillus halodurans (*bha*)	bh0341	bh0342	bh0336	bh0340		
Bacteroides fragilis (*bfa*)	rbx02154				rbx02151	rbx02155
Campylobacter jejuni (*cje*)	cj1522C				cj1523C	cj1521C
Chlorobium tepidum TLS (*cte*)	ct1130	ct1129	ct1135	ct1131		
Chromobacterium violaceum (*cvi*)	cv1229	cv1230	cv1224	cv1228		
Clostridium tetani E88 (*ctcA*)	ctc01148		ctc01146	ctc01147		
Clostridium tetani E88 (*ctcB*)	ctc01463		ctc01465	ctc01464		
Corynebacterium diphtheriae (*cdi*)	dip0037				dip0036	dip0038
Desulfovibrio vulgaris (*dvu*)	dvu0134	dvu0135	cas3	dvu0132		
Fusobacterium nucleatum ATCC 25586 (*fnuA*)	fn1177	fn1176	fn1179	fn1178		
Fusobacterium nucleatum ATCC 49256 (*fnuB*)	fnv1819				fnv1820	fnv1821
Leptospira interrogans (*lic*)	lic10942	lic10940	lic10938	lic10943		
Listeria innocua (*lin*)	lin2743				lin2744	
Listeria monocytogenes (*lmo*)	lmo2714				lmo2713	lmo2712
Methanobacterium thermoautotrophicum (*mth*)	mth1084	mth1083	mth1086	mth1085		
Methanococcus jannaschii (*mja*)	mj0378	mj0386	mj0376	mj0377		
Methanosarcina mazei (*mma*)	mm0559	mm0558	mm0561	mm0557		
Mycoplasma gallisepticum (*mga*)	mga_0523				mga_0519	mga_0525
Mycoplasma mobile (*mmo*)	mmoB0320				mmoB0330	
'Nanoarchaeum equitans' (*neq*)	neq017	neq016	neq022	neq021		
Neisseria meningitidis (*nme*)	nma0630				nma0631	nma0629
Pasteurella multocida (*pmu*)	pm1126				pm1127	pm1125
Porphyromonas gingivalis (*pgi*)	cas1	cas2	cas3	cas4		
Pyrobaculum aerophilum (*pai*)	pae0200	pae0199	pae0198	pae0208		
Pyrococcus furiosus (*pfu*)	pf1118	pf1117	pf1120	pf1119		
Pyrococcus horikoshii (*phoA*)	ph0173	phyyy	ph0176	ph0175		
Pyrococcus horikoshii (*phoB*)	ph1245	phxxx	ph1246	phs033		
Rhodospirillum rubrum (*rru*)	rrr03414				rrr03411	rrr03413
Streptococcus agalactiae (*sag*)	sag0895				sag0894	sag0896
Streptococcus mutans (*smu*)	smu1404C				smu1405C	smu1403C
Streptococcus pyogenes (*spyA*)	spy1562	spy1561	spy1567	spy1563		
Streptococcus pyogenes (*spyB*)	spy1047				spy1046	spy1048
Streptococcus suis (*ssu*)	ssui1059				ssui1060	ssui1061
Streptococcus thermophilus (*sth*)	str0658				str0657	str0659
Sulfolobus solfataricus (*sso*)	sso1405	sso1404	sso1402	sso1392		
Sulfolobus tokodaii (*sto*)	sto026	sto025	sto032			
Thermoanaerobacter tengcongensis (*tte*)	tte2658	tte2657	tte2660	tte2659		
Thermotoga maritima (*tma*)	tm1797	tm1796	tm1799	tm1798		
Thermus thermophilus (*tth*)	ttp0196	ttp0195	ttp0132	ttp0197		
Treponema denticola (*tde*)	tde0328				tde0327	tde0329
Wolinella succinogenes (*wsu*)	ws1444				ws1445	ws1443
Xanthomonas axonopodis (*xac*)	xac3842	xac3843	xac3837	xac3843		

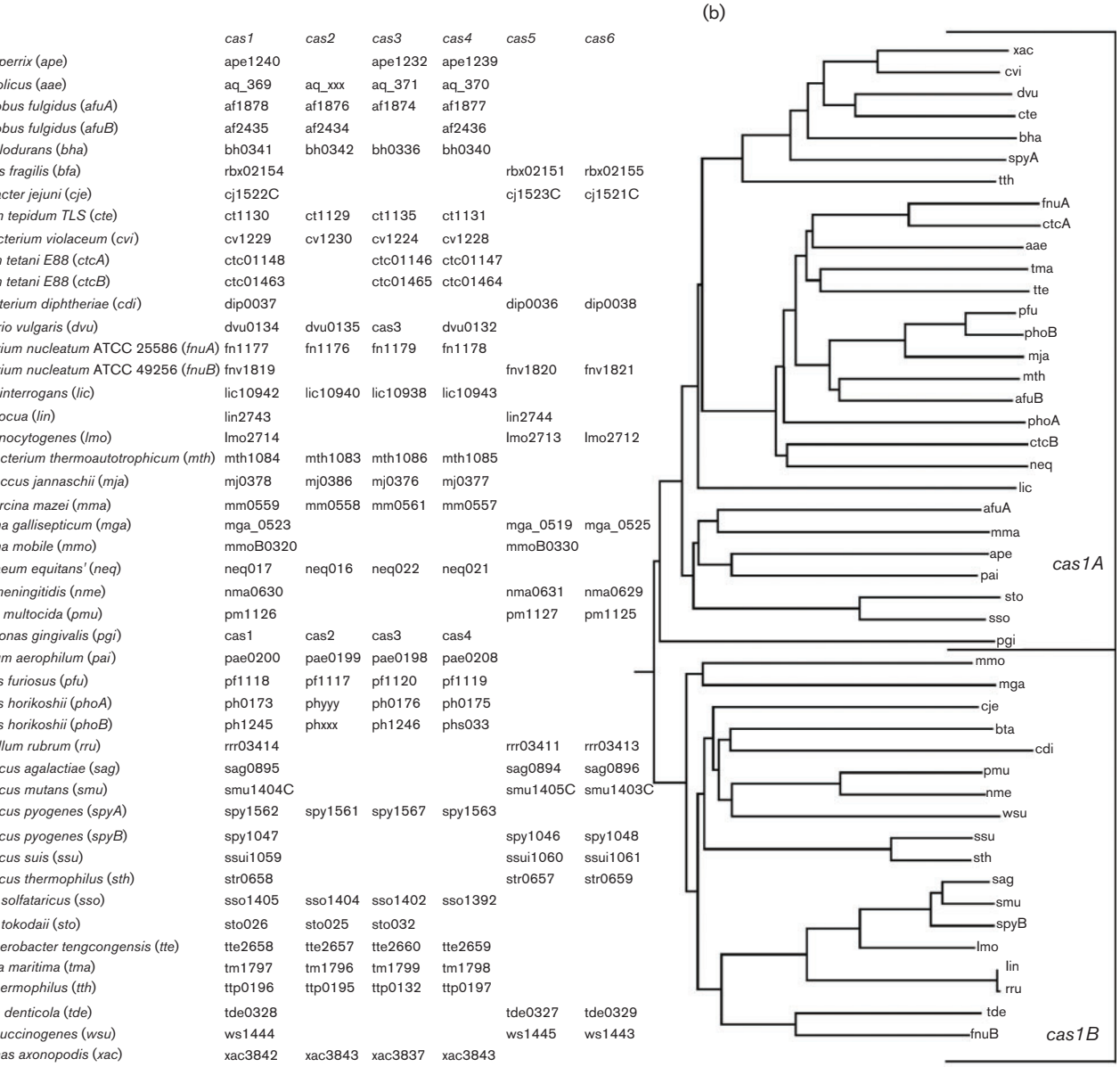


Fig. 2. Distribution of *cas* genes in microbial genomes. (a) Two families of *cas* clusters. The genes were identified in the ERGO and MBGD databases. The *cas2* genes of *Aquifex aeolicus* (*aq_xxx*) and *P. horikoshii* (*phyyy* and *phxxx*) were not annotated in the databases but have been reported (Jansen *et al*., 2002). (b) Clustering of *cas1* genes into two families. The neighbour-joining facility of CLUSTALW was used for tree construction. (c) Association of the *cas1A cas5 cas6* gene cluster with CRISPR structures. Genes are indicated by arrows, CRISPR structures by black boxes and the position of the incomplete repeat is indicated by a triangle, placed above or below the gene line to denote the repeat is in the same or the opposite orientation, respectively, to that of the repeats present in the CRISPR structures. The genes from the ERGO database are shown.

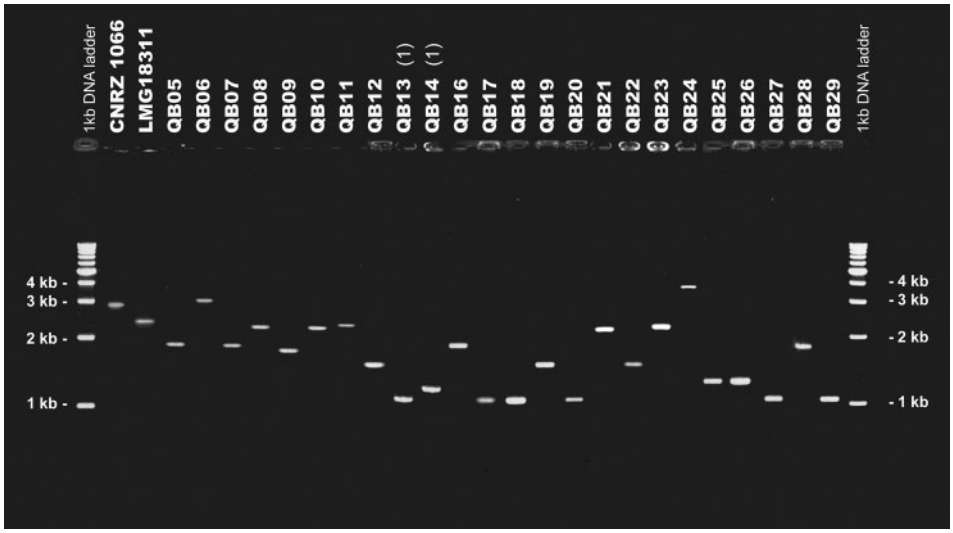


Fig. 3. Identification of the CRISPR locus in *S. thermophilus* and *S. vestibularis*. PCR was carried out with primers yc31 and yc70, and the products were analysed by gel electrophoresis. Strain designation is indicated above the lanes, and details are given in Table 2.

position of an allele. The length of spacers comprised between 28 and 32 bp, being 30 for 556 of the 618 spacers (90 %). The total length of the CRISPR loci (from the first nucleotide of the first repeat to the last nucleotide of the last repeat) varied between 628 and 3404 bp. Three identical CRISPR loci were present in more than one strain (groups A, B and C, found in five, two and two strains, respectively; Table 2). Internal duplications were detected in almost a quarter of the loci, indicating a substantial level of recombination in CRISPR structures.

The 519 spacers of the 20 different CRISPR loci were compared, and 349 (70 % of the total) were found to be unique. At least one example of each unique spacer must have been acquired in a separate event, presumably involving a CRISPR-specific mechanism (see below). In contrast, the identical spacers present in different strains could have been transferred laterally from a donor to a recipient strain, and integrated at a CRISPR locus of the recipient strain by general homologous recombination. The number of spacers common to different strains is summarized in Table 3.

S. thermophilus spacers are homologous to extrachromosomal elements

Out of 349 unique spacers some 124 (36 %) had significant matches ($E < 0{\cdot}001$) with sequences in the NCBI nucleotide database. The best matches were with phages of streptococci (75 %), and plasmids of *S. thermophilus* or *Lactococcus lactis* (20 %). A list of the spacers that have 100 % identity to regions of phages and plasmids is presented in Supplementary Table 2, available with the online journal. Different CRISPR alleles had spacers matching different phages; the number of matching spacers at different alleles was between 0 and 21 (Table 2). No particular phage region was found to match the spacers, as illustrated for the phage Sfi21 (Fig. 4) and also found for other fully sequenced phages (data not shown). The order of spacers within a CRISPR allele did not correspond to the order of matching regions along the phage genome. However, spacers were homologous predominantly with one of the phage strands, as illustrated for Sfi21 (Fig. 4) and also found for all fully sequenced phages, where the bias for 27 unique spacers strictly identical with phage sequences was about 3·5 to 1 (21 to 6). As this strand is predominantly coding, the orientation of the spacers relative to the phage ORFs was biased to a similar extent (3·4 to 1; 17 of the 22 unique spacers homologous to an ORF had an orientation corresponding to that of the ORF; Table 3).

Alignment of 70 bp regions from extrachromosomal elements that comprised 30 bp of 100 % homology with a spacer revealed no similarities other than a 5 bp degenerate sequence, Pu-py-A-A-a, situated downstream from the spacer-matching stretch (Fig. 5). It might be significant that this sequence matches the end of the *S. thermophilus* repeat (...ACAAC), with the exception of the very terminal base, and that it is purine-rich, as are the corresponding ends of CRISPR repeats from many other organisms, having a consensus ...GAAAC (Mojica *et al.*, 2000).

Phage resistance of *S. thermophilus* is correlated with the number of spacers in a CRISPR locus

As many spacers in the CRISPR loci have homology with phage sequences, we searched for a correlation between CRISPR properties and a phage-related phenotype. A previous study reported the phage-resistance profile of a number of *S. thermophilus* strains (Fayard, 1993). The results

Table 2. *Streptococcus* strains and their CRISPR loci

Strain ID	Other ID*	Species	CRISPR length†	No. of spacers	No. of unique spacers‡	Identical allele	Duplication		No. of phage-matching spacers§							Phage sensitivity‖
							No.	No. of spacers involved	Sfi11 (AF158600)	Sfi19 (AF115102)	Sfi21 (AF115103)	DT1 (AF085222)	O1205 (U88974·1)	7201 (AF145054)	Total	
CNRZ	CNRZ1066	*S. thermophilus*	2744	41	0	–	2	7	7	6	7	4	9	0	16	7/59
LMG	LMG18311	*S. thermophilus*	2213	33	15	–	0	–	4	4	3	1	4	5	12	ND
QB05	CNRZ302	*S. thermophilus*	1618	24	3	–	1	1	2	0	0	1	2	1	4	0/59
QB06	CNRZ388	*S. thermophilus*	2808	42	16	–	2	2	2	6	5	5	2	6	14	0/59
QB07	CNRZ389	*S. thermophilus*	1619	24	15	–	0	–	3	2	1	2	2	5	8	4/59
QB08	CNRZ1100	*S. thermophilus*	2016	30	4	–	0	–	2	4	2	2	2	2	11	0/59
QB09	CNRZ1202	*S. thermophilus*	1551	23	7	–	0	–	2	5	8	4	3	2	10	0/59
QB10	CNRZ703	*S. thermophilus*	2013	30	30	–	0	–	1	2	5	1	0	1	6	0/59
QB11	CNRZ1575	*S. thermophilus*	2080	31	31	–	0	–	2	2	1	1	1	0	4	3/59
QB12	CNRZ385	*S. thermophilus*	1354	20	20	–	0	–	0	3	3	2	1	3	6	0/59
QB13	JIM8229	*S. vestibularis*	628	9	8	–	1	1	0	0	0	0	0	0	0	ND
QB14	JIM8230	*S. vestibularis*	762	11	11	–	0	–	1	1	1	0	1	0	1	ND
QB16	JIM1567	*S. thermophilus*	1620	24	14	–	1	3	3	4	1	1	3	2	5	27/59
QB17	JIM1560	*S. thermophilus*	827	12	12	A	0	–	1	1	2	0	2	0	5	32/59
QB18	JIM1575	*S. thermophilus*	827	12	12	A	0	–	1	1	2	0	2	0	5	27/59
QB19	JIM1584	*S. thermophilus*	1289	19	19	C	0	–	1	1	1	1	0	0	2	32/59
QB20	JIM1588	*S. thermophilus*	827	12	12	A	0	–	1	1	2	0	2	0	5	26/59
QB21	JIM70	*S. thermophilus*	1946	29	2	–	0	–	2	1	1	1	1	2	3	2/7
QB22	JIM71	*S. thermophilus*	1289	19	19	C	0	–	1	1	1	1	0	0	2	2/7
QB23	JIM72	*S. thermophilus*	2011	30	3	–	0	–	2	2	2	2	1	3	4	2/7
QB24	JIM76	*S. thermophilus*	3404	51	1	–	4	12	10	8	9	6	12	0	21	ND
QB25	CNRZ1205	*S. thermophilus*	1089	16	0	B	0	–	0	3	6	2	1	2	7	47/59
QB26	1205.3¶	*S. thermophilus*	1089	16	0	B	0	–	0	3	6	2	1	2	7	ND
QB27	4035#	*S. thermophilus*	827	12	12	A	0	–	1	1	2	0	2	0	5	ND
QB28	JIM1293	*S. thermophilus*	1354	20	0	–	0	–	1	0	0	1	1	1	3	1/7
QB29	JIM1518	*S. thermophilus*	827	12	12	A	0	–	1	1	2	0	2	0	5	ND

*CNRZ, INRA collection of micro-organisms; JIM, collection of micro-organisms of Genetique Microbienne; LMG, collection of micro-organisms of University of Louvain la Neuve.
†Locus length is given in bp, from the first nucleotide of the first repeat to the last nucleotide of the last repeat.
‡All spacers of identical alleles are indicated.
§$E < 0{\cdot}001$. GenBank entries are indicated in parentheses.
‖Given as a ratio of propagating phages to total tested. ND, no data.
¶Strain described by Stanley *et al.* (1999).
#Strain described by Le Marrec *et al.* (1997).

Table 3. Relations between CRISPR alleles from different *S. thermophilus* strains

The number of spacers common to the displayed strains is shown.

	CNRZ	LMGs	QB05	QB06	QB07	QB08	QB09	QB10	QB11	QB12	QB13	QB14	QB16	QB17	QB18	QB19	QB20	QB21	QB22	QB23	QB24	QB25	QB26	QB27	QB28	QB29
CNRZ	**41**	0	0	0	0	0	0	0	0	0	0	0	**5**	0	0	0	0	0	0	0	**41**	0	0	0	0	0
LMGs	0	**33**	0	**13**	**1**	**16**	0	0	0	0	0	0	0	0	0	0	0	**3**	0	**3**	0	0	0	0	0	0
QB05	0	0	**24**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**19**	0
QB06	0	**13**	0	**42**	**7**	**23**	0	0	0	0	0	0	0	0	0	0	0	**9**	0	**9**	0	0	0	0	0	0
QB07	0	**1**	0	**7**	**24**	**5**	0	0	0	0	0	0	0	0	0	0	0	**4**	0	**4**	0	0	0	0	0	0
QB08	0	**15**	0	**23**	**5**	**30**	0	0	0	0	0	0	0	0	0	0	0	**7**	0	**7**	0	0	0	0	0	0
QB09	0	0	0	0	0	0	**23**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**16**	**16**	0	0	0
QB10	0	0	0	0	0	0	0	**30**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
QB11	0	0	0	0	0	0	0	0	**31**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
QB12	0	0	0	0	0	0	0	0	0	**20**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
QB13	0	0	0	0	0	0	0	0	0	0	**9**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
QB14	0	0	0	0	0	0	0	0	0	0	0	**11**	0	0	0	0	0	0	0	0	0	0	0	0	0	0
QB16	**5**	0	0	0	0	0	0	0	0	0	0	0	**24**	0	0	0	0	0	0	0	**7**	0	0	0	0	0
QB17	0	0	0	0	0	0	0	0	0	0	0	0	0	**12**	**12**	0	**12**	0	0	0	0	0	0	**12**	0	**12**
QB18	0	0	0	0	0	0	0	0	0	0	0	0	0	**12**	**12**	0	**12**	0	0	0	0	0	0	**12**	0	**12**
QB19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**19**	0	0	**19**	0	0	0	0	0	0	0
QB20	0	0	0	0	0	0	0	0	0	0	0	0	0	**12**	**12**	0	**12**	0	0	0	0	0	0	**12**	0	**12**
QB21	0	**3**	0	**9**	**4**	**7**	0	0	0	0	0	0	0	0	0	0	0	**29**	0	**27**	0	0	0	0	0	0
QB22	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**19**	0	0	**19**	0	0	0	0	0	0	0
QB23	0	**3**	0	**9**	**4**	**7**	0	0	0	0	0	0	0	0	0	0	0	**27**	0	**30**	0	0	0	0	0	0
QB24	**48**	0	0	0	0	0	0	0	0	0	0	0	**7**	0	0	0	0	0	0	0	**51**	0	0	0	0	0
QB25	0	0	0	0	0	0	**16**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**16**	**16**	0	0	0
QB26	0	0	0	0	0	0	**16**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**16**	**16**	0	0	0
QB27	0	0	0	0	0	0	0	0	0	0	0	0	0	**12**	**12**	0	**12**	0	0	0	0	0	0	**12**	0	**12**
QB28	0	0	**19**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**20**	0
QB29	0	0	0	0	0	0	0	0	0	0	0	0	0	**12**	**12**	0	**12**	0	0	0	0	0	0	**12**	0	**12**

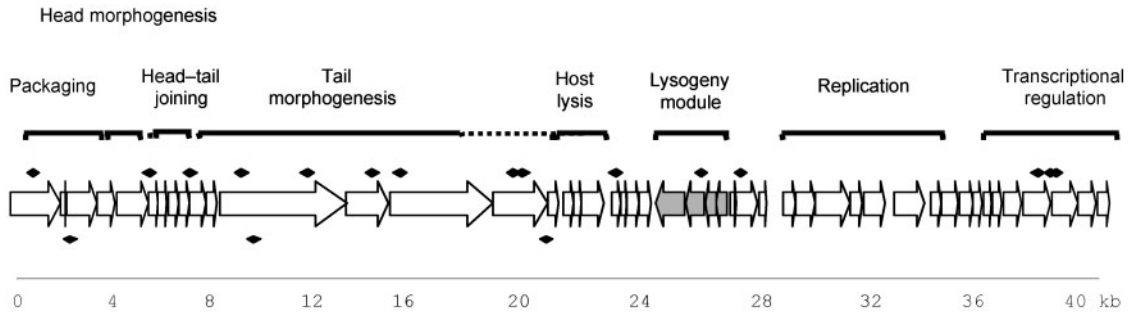


Fig. 4. Localization of spacer-matching sequences along the phage Sfi21 genome. The phage genetic map is drawn after GenBank entry NC_000872 (ORFs are shown as arrows), the regions involved in different stages of phage development, identified by comparative analysis (Desiere *et al*., 2002), are indicated above the map, and the scale (in kb) below it. Phage regions having a BLAST E score $<0{\cdot}001$ with the CRISPR spacers are indicated by the diamonds placed above or below the map, denoting homology with the top or the bottom DNA strand, respectively.

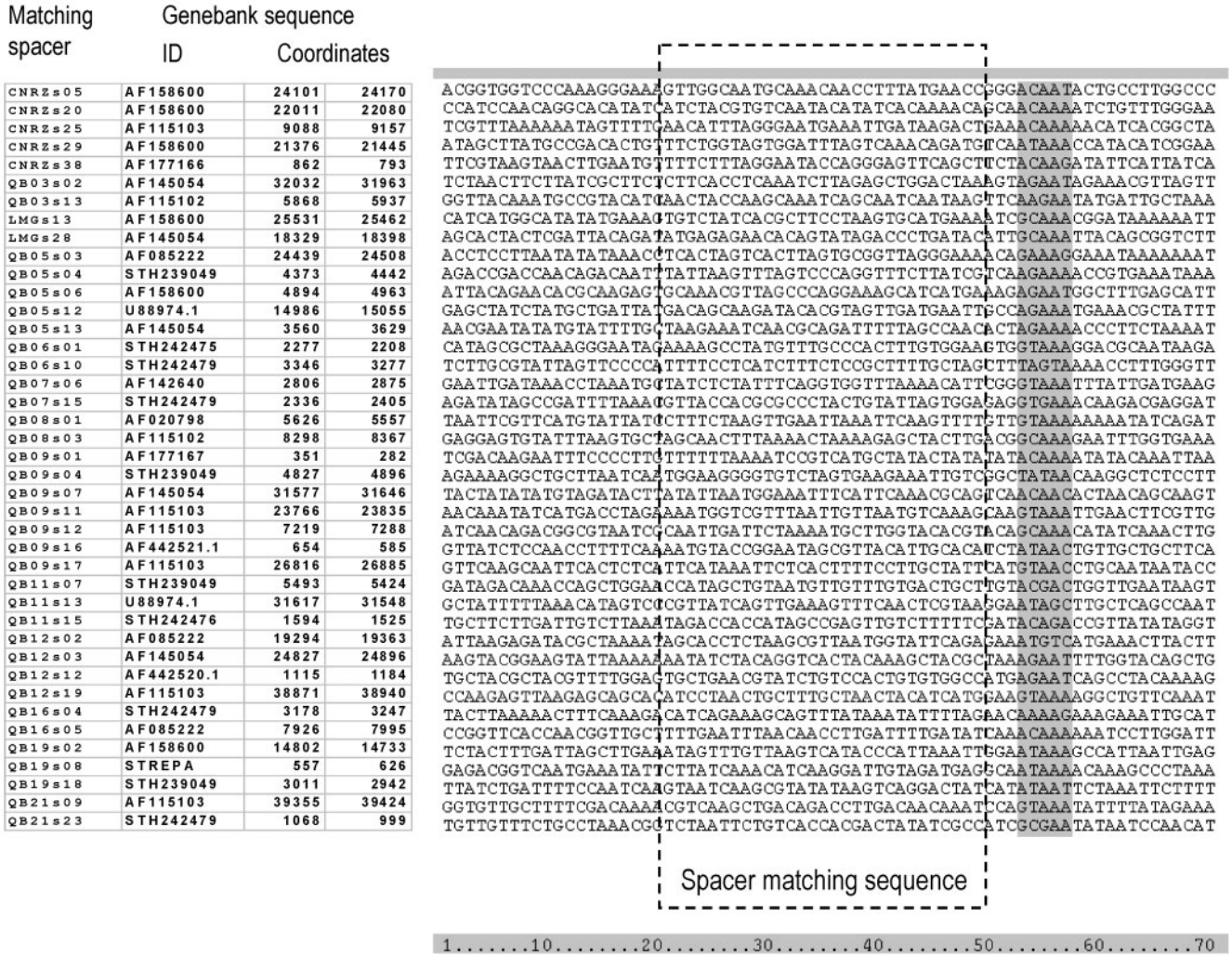
for a panel of 59 phages tested on a number of strains for which we determined the CRISPR locus sequence are summarized in Table 2 and displayed in Fig. 6. A negative correlation between the number of spacers at a locus and the sensitivity of the strain to phage infection (expressed as a proportion of phages able to propagate on a strain) is observed. About half of the variance of phage sensitivity for nine strains infected by at least one phage can be explained by the number of spacers (R^2, 0·51; Fig. 6). The five strains resistant to all phages (Fig. 6) were not included in the former group, as they may encode dominant, CRISPR-independent, phage-resistant determinants. When the data obtained for additional strains, tested with a smaller panel of seven phages, were examined (Table 1; M.-C. Chopin, personal communication) no correlation was seen, possibly due to the small sample size (Fig. 6). However, when all the data were pooled, the correlation was still significant (R^2, 0·43), suggesting that the small panel results may be not very different from those obtained with the large panel.

DISCUSSION

About 40 % of *S. thermophilus* CRISPR spacers show significant homology to sequences in the NCBI database. An overwhelming majority of these are homologous to *S. thermophilus* phages (75 %), and a substantial fraction to *S. thermophilus* and *Lactococcus lactis* plasmids (20 %). This indicates the extrachromosomal origin of many CRISPR spacers in *S. thermophilus*. Are the other spacers in *S. thermophilus* of a similar origin? At present, the complete sequences of 6 *S. thermophilus* phages and 6 other lactic acid bacteria phages have been reported (Desiere *et al*., 2002); these 12 constitute possibly the most thoroughly characterized phage group, with respect to genome data (Brussow & Hendrix, 2002). In addition, the NCBI database contains partial or complete sequences of over 20 *S. thermophilus* and 67 *Lactococcus lactis* plasmids. However, many more *S. thermophilus* phages and plasmids have still to be sequenced, and it is thus tempting to speculate that many more spacers in this organism have extrachromosomal origin. As these elements must have been invading *S. thermophilus* strains during their evolutionary history, we suggest that CRISPR spacers reflect past phage and plasmid infections.

An extrachromosomal origin of spacers is not limited to *S. thermophilus*, as it is found in many other bacteria and archaea. However, some spacers are homologous with genes that are not clearly related to extrachromosomal genes. Remarkably, these genes are frequently (in ~75 % of cases) located in the vicinity of extrachromosomally derived genes, or in regions potentially transferred from unrelated organisms, or in regions where gene order differs from that of the phylogenetically close genomes. Horizontal gene transfer may underlie all these cases, suggesting that incorporation of gene fragments into CRISPR structures might take place upon invasion of a prokaryote cell by foreign DNA. This invasion may be most often mediated by the extrachromosomal elements, but could also be due to other processes, such as DNA transformation. Nevertheless, the overwhelming majority of CRISPR spacers (98 %) have no homology with known genes. The further accumulation of sequences in the databases may reveal the origin of these spacers.

The mechanism of CRISPR generation is not known, but the *cas* genes, which are invariably closely linked to the CRISPR structures, are presumably involved in this process, as was previously pointed out by Jansen *et al.* (2002). The process should involve the formation of segments of a defined size, destined to become spacers, and their linkage to the repeated element. The presence of exonuclease motifs of the *recB* type in the *cas4* genes, and the HNH endonuclease motif in the *cas5* genes prompts us to suggest that the segments are formed by a nucleolytic activity, but in two different ways. The Cas4 exonuclease might act from an end, as does the RecBCD enzyme complex, aided by a Cas3 helicase, which generates oligonucleotides (Singleton *et al*.,



Position	53	54	55	56	57
	Nucleotide frequency				
A	0·61	0·12	0·83	0·90	0·63
G	0·34	0·17	0·15	0·05	0·07
C	0·00	0·34	0·00	0·00	0·15
T	0·05	0·37	0·02	0·05	0·15
Consensus	Pu	py	A	A	a

Fig. 5. The consensus sequence adjacent to a CRISPR spacer-matching DNA stretch. The top part of the figure shows an alignment of DNA regions of different phages and plasmids containing a region identical with a CRISPR spacer. The identity of the matching spacer, GenBank ID of the sequence and the coordinates of the displayed region are given on the left. The box and grey zone identify the spacer-matching sequence and the consensus sequence, respectively. The middle part of the figure shows a CLUSTAL-generated display of the nucleotide SD for each position of the displayed sequence. The bottom part of the figure shows a summary of the nucleotide frequencies at the positions with high SDs and the deduced consensus sequence. Capital and small letters identify the positions containing $>80\,\%$ or $>60\,\%$, respectively, of one or two bases.

2004 and references therein). In contrast, the Cas5 endonuclease might excise the segments by internal DNA cleavage, possibly directed by the short conserved sequence that we identified in the extrachromosomal donor elements at a constant position relative to the spacer-matching region. A precedent for this type of activity is the action of

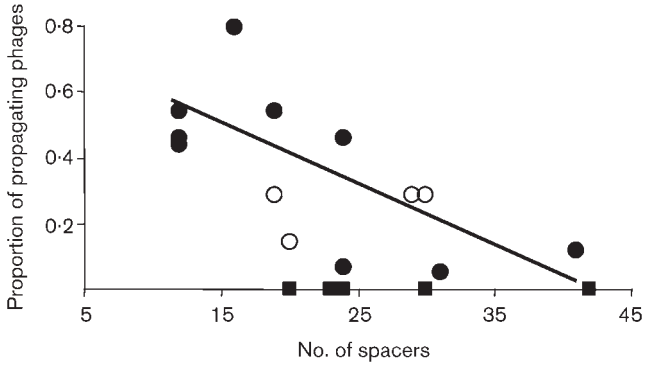


Fig. 6. Correlation of *S. thermophilus* phage resistance and the number of spacers in a CRISPR locus. Filled symbols correspond to data obtained from strains tested with the panel of 59 phages. The line of best-fit refers to strains that were not fully phage resistant (●), and for which $y=-0{\cdot}02x+0{\cdot}77$ and $R^2=0{\cdot}51$. Fully phage-resistant strains (■), were not taken into account for the correlation shown. ○, Strains tested with the panel of seven phages.

type III restriction enzymes, which cut 25 to 27 bases from their recognition site (see Dryden *et al.*, 2001 for a review). The type III restriction enzyme-like endonucleolytic action might be polar, as it involves tracking on the DNA, which could account for the biased orientation of the phage-derived spacers in the *S. thermophilus* CRISPR structures. We envisage that the Cas1 proteins, encoded by the two related genes, *cas1A* and *cas1B*, found in the two types of *cas* gene clusters, may be involved in the process of linking the DNA segments to the repeats. Biochemical study of Cas proteins should allow us to test this model of CRISPR formation.

The biological role of CRISPR elements is not known, although it was suggested that this element plays a role in replicon partitioning (Mojica *et al.*, 1995; She *et al.*, 1998). A protein that binds to the repeats was purified from *Sulfolobus solfataricus*, and it was suggested that it might be involved in DNA condensation of the CRISPR structures (Peng *et al.*, 2003). Here we report a correlation between the number of spacers in a locus and the resistance of *S. thermophilus* to phage infection, suggesting that CRISPRs can have a different biological role, protecting the bacteria against phage attack. How could such protection be mediated? A possible mechanism is via anti-sense RNA inhibition of phage gene expression, which is supported by the following observations. First, spacers that are homologous to phage coding sequences can have either of the two orientations within a CRISPR locus, and thus give rise to anti-sense RNA, irrespective of the direction of locus transcription. Second, CRISPR loci do appear to be transcribed, as reported for *Archeoglobus fulgidus* (Tang *et al.*, 2002) and *Sulfolobus solfataricus* (Tang *et al.*, 2005), and can thus generate the anti-sense RNA. It was proposed that various anti-sense short RNAs might regulate gene expression (Tang *et al.*, 2005). Third, it was shown that anti-sense RNA inhibits phage propagation (Sturino & Klaenhammer, 2002, 2004). Studies combining fully sequenced phages and strains with characterized CRISPR loci should allow further testing of this hypothesis, notwithstanding the fact that, besides the effect of CRISPR, many other factors also contribute to phage resistance (see Coffey & Ross, 2002 for a recent review). Finally, CRISPR spacers could protect bacteria not only against phage infection, but also against invasion by other extrachromosomal elements, inhibiting expression of the genes they carry. Horizontal exchanges between CRISPR elements, which we detected by comparing different loci, could extend the protective range to extrachromosomal elements that have not yet invaded a particular strain. Such a beneficial, protective role could account for the wide spread and the apparent stability of CRISPR structures among prokaryotes.

ACKNOWLEDGEMENTS

We thank Marie-Christine Chopin for communication of unpublished results, Pierre Renault for discussions of the phage-resistance results, and Douwe van Sinderen for providing some of the *S. thermophilus* strains used in this study.

REFERENCES

Bolotin, A., Quinquis, B., Renault, P. & 20 other authors (2004). Complete genome sequence and comparative analysis of the dairy bacterium *Streptococcus thermophilus*. *Nat Biotechnol* **22**, 1554–1558.

Brussow, H. & Hendrix, R. W. (2002). Phage genomics, small is beautiful. *Cell* **108**, 13–16.

Coffey, A. & Ross, R. P. (2002). Bacteriophage-resistance systems in dairy starter strains, molecular analysis to application. *Antonie van Leeuwenhoek* **82**, 303–321.

Desiere, F., Lucchini, S., Canchaya, C., Ventura, M. & Brussow, H. (2002). Comparative genomics of phages and prophages in lactic acid bacteria. *Antonie van Leeuwenhoek* **82**, 73–91.

Dryden, D. T., Murray, N. E. & Rao, D. N. (2001). Nucleoside triphosphate-dependent restriction enzymes. *Nucleic Acids Res* **29**, 3728–3741.

Fayard, B. (1993). *Caractérisation de 69 bactériophages of Streptococcus salivarius subsp. thermophilus incluant 10 bactériophages tempérés*. PhD thesis, University Nancy I, France.

Groenen, P. M., Bunschoten, A. E., van Soolingen, D. & van Embden, J. D. (1993). Nature of DNA polymorphism in the direct repeat cluster of *Mycobacterium tuberculosis*, application for strain differentiation by a novel typing method. *Mol Microbiol* **43**, 1057–1065.

Higgins, D. G. & Sharp, P. M. (1989). Fast and sensitive multiple sequence alignments on a microcomputer. *Comput Appl Biosci* **43**, 151–153.

Hoe, N., Nakashima, K., Grigsby, D. & 7 other authors (1999). Rapid molecular genetic subtyping of serotype M1 group A *Streptococcus* strains. *Emerg Infect Dis* **43**, 254–263.

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

Kamerbeek, J., Schouls, L., Kolk, A. & 8 other authors (1997). Simultaneous detection and strain differentiation of *Mycobacterium tuberculosis* for diagnosis and epidemiology. *J Clin Microbiol* **43**, 907–914.

Le Marrec, C., van Sinderen, D., Walsh, L., Stanley, E., Vlegels, E., Moineau, S., Heinze, P., Fitzgerald, G. & Fayard, B. (1997). Two groups of bacteriophages infecting *Streptococcus thermophilus* can be distinguished on the basis of mode of packaging and genetic determinants for major structural proteins. *Appl Environ Microbiol* **63**, 3246–3253.

Makarova, K. S., Aravind, L., Grishin, N. V., Rogozin, I. B. & Koonin, E. V. (2002). A DNA repair system specific for thermophilic Archaea and bacteria predicted by genomic context analysis. *Nucleic Acids Res* **30**, 482–496.

Maté, M. J. & Kleanthous, C. (2004). Structure-based analysis of the metal-dependent mechanism of H-N-H endonucleases. *J Biol Chem* **279**, 34763–34769.

Mojica, F. J., Ferrer, C., Juez, G. & Rodriguez-Valera, F. (1995). Long stretches of short tandem repeats are present in the largest replicons of the Archaea *Haloferax mediterranei* and *Haloferax volcanii* and could be involved in replicon partitioning. *Mol Microbiol* **43**, 85–93.

Mojica, F. J., Diez-Villasenor, 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* **43**, 244–246.

Peng, X., Brügger, K., Shen, B., Chen, L., She, Q. & Garrett, R. A. (2003). Genus-specific protein binding to the large clusters of DNA repeats (Short Regularly Spaced Repeats) present in *Sulfolobus* genomes. *J Bacteriol* **185**, 2410–2417.

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.

Saravanan, M., Bujnicki, J. M., Cymerman, I. A., Rao, D. N. & Nagaraja, V. (2004). Type II restriction endonuclease R.KpnI is a member of the HNH nuclease superfamily. *Nucleic Acids Res* **32**, 6129–6135.

Schouls, L. M., Reulen, S., Duim, B., Wagenaar, J. A., Willems, R. J. L., Dingle, K. E., Colles, F. M. & van Embden, J. D. (2003). Comparative genotyping of *Campylobacter jejuni* by amplified fragment length polymorphism, multilocus sequence typing, and short repeat sequencing: strain diversity, host range, and recombination. *J Clin Microbiol* **41**, 15–26.

She, Q., Phan, H., Garrett, R. A., Albers, S. V., Stedman, K. M. & Zillig, W. (1998). Genetic profile of pNOB8 from *Sulfolobus*, the first conjugative plasmid from an archaeon. *Extremophiles* **2**, 417–425.

Simpson, C. L., Giffard, P. M. & Jacques, N. A. (1993). A method for the isolation of RNA from *Streptococcus salivarius* and its application to the transcriptional analysis of the *gtfJK* locus. *FEMS Microbiol Lett* **108**, 93–97.

Singleton, M. R., Dillingham, M. S., Gaudier, M., Kowalczykowski, S. C. & Wigley, D. B. (2004). Crystal structure of RecBCD enzyme reveals a machine for processing DNA breaks. *Nature* **432**, 187–193.

Stanley, E., Fitzgerald, G. F. & van Sinderen, D. (1999). Characterisation of *Streptococcus thermophilus* CNRZ1205 and its cured and re-lysogenised derivatives. *FEMS Microbiol Lett* **176**, 503–510.

Sturino, J. M. & Klaenhammer, T. R. (2002). Expression of antisense RNA targeted against *Streptococcus thermophilus* bacteriophages. *Appl Environ Microbiol* **68**, 588–596.

Sturino, J. M. & Klaenhammer, T. R. (2004). Antisense RNA targeting of primase interferes with bacteriophage replication in *Streptococcus thermophilus*. *Appl Environ Microbiol* **70**, 1735–1743.

Tang, T. H., Bachellerie, J. P., Rozhdestvensky, T., Bortolin, M. L., Huber, H., Drungowski, M., Elge, T., Brosius, J. & Huttenhofer, A. (2002). Identification of 86 candidates for small non-messenger RNAs from the archaeon *Archaeoglobus fulgidus*. *Proc Natl Acad Sci U S A* **99**, 7536–7541.

Tang, T. H., Polacek, N., Zywicki, M., Huber, H., Brugger, K., Garrett, R., Bachellerie, J. P. & Huettenhofer, A. (2005). Identification of novel non-coding RNAs as potential anti-sense regulators in the archaeon *Sulfolobus solfataricus*. *Mol Microbiol* **55**, 469–481.

Terzaghi, B. E. & Sandine, W. E. (1975). Improved medium for lactic streptococci and their bacteriophages. *Appl Microbiol* **29**, 807–813.

van Belkum, A., Scherer, S., van Alphen, L. & Verbrugh, H. (1998). Short-sequence DNA repeats in prokaryotic genomes. *Microbiol Mol Biol Rev* **43**, 275–293.

Walker, D. C., Georgiou, T., Pommer, A. J., Walker, D., Moore, G. R., Kleanthous, C. & James, R. (2002). Mutagenic scan of the H-N-H motif of colicin E9: implications for the mechanistic enzymology of colicins, homing enzymes & apoptotic endonucleases. *Nucleic Acids Res* **30**, 3225–3234.