Microbiology (2005), **151**, 653–663 **DOI** 10.1099/mic.0.27437-0

CRISPR elements in *Yersinia pestis* acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies

C. Pourcel,[1] G. Salvignol[1] and G. Vergnaud[1,2]

[1]GPMS, Institut de Génétique et Microbiologie, Université Paris XI, 91405 Orsay cedex, France
[2]Centre d'Etudes du Bouchet, 5 rue Lavoisier, 91710 Vert le Petit, France

Correspondence
G. Vergnaud
Gilles.Vergnaud@igmors.u-psud.fr

Received 28 June 2004
Revised 8 November 2004
Accepted 12 November 2004

The remarkable repetitive elements called CRISPRs (clustered regularly interspaced short palindromic repeats) consist of repeats interspaced with non-repetitive elements or 'spacers'. CRISPRs are present in both archaea and bacteria, in association with genes involved in DNA recombination and repair. In the *Yersinia pestis* genome, three such elements are found at three distinct loci, one of them being highly polymorphic. The authors have sequenced a total of 109 alleles of the three *Y. pestis* CRISPRs and they describe 29 new spacers, most being specific to one isolate. In nine strains of *Yersinia pseudotuberculosis*, 132 spacers were found, of which only three are common to *Y. pestis* isolates. In *Y. pestis* of the Orientalis biovar investigated in detail here, deletion of motifs is observed but it appears that addition of new motifs to a common ancestral element is the most frequent event. This takes place at the three different loci, although at a higher rate in one of the loci, and the addition of new motifs is polarized. Interestingly, the most recently acquired spacers were found to have a homologue at another locus in the genome, the majority of these inside an inactive prophage. This is believed to be the first time that the origin of the spacers in CRISPR elements has been explained. The CRISPR structure provides a new and robust identification tool.

INTRODUCTION

Short regularly spaced repeats (SRSRs) (Mojica *et al.*, 2000), also called clustered regularly interspaced short palindromic repeats (CRISPRs) (Jansen *et al.*, 2002) have been found in all archaea and hyperthermophilic bacteria investigated so far and also in some eubacteria at one or several loci in the chromosome. Their structure is remarkably constant. They consist of repeated sequences, 21–37 bp in length, separated by similarly sized variable sequences or spacers. The biological function of CRISPRs is not known. The repeated sequence possesses characteristic features suggesting that it could be a target for DNA-binding proteins, and in *Haloferax mediterranei* it was proposed to be involved in replicon partitioning (Mojica *et al.*, 1995). The structure of CRISPRs, and the fact that they are located near the replication origin in *Pyrococcus* species, support this hypothesis (Zivanovic *et al.*, 2002). The existence of several conserved genes (called *cas* genes by Jansen *et al.*, 2002) in the vicinity of CRISPRs, potentially involved in DNA recombination and repair, is an additional argument for a physiological role. CRISPRs are thought to increase in size by duplicating the constant sequence and adding at least one new spacer by a still unknown mechanism. Indeed, in the previously described CRISPRs the origin of the spacers could not be elucidated. The most-studied CRISPR locus is the direct repeat (DR) region of *Mycobacterium tuberculosis*. The high degree of polymorphism generated by the variable spacers forms the basis of the spoligotyping method (Kamerbeek *et al.*, 1997). Recently, a detailed analysis of the spacers by sequencing the DR of 26 strains of the *M. tuberculosis* complex has led to the proposal that a common ancestor, bearing a large number of units, has evolved by the interstitial deletion of motifs (Groenen *et al.*, 1993; Sola *et al.*, 2003; van Embden *et al.*, 2000). This explains the hundreds of different combinations seen in contemporary strains. No indication of insertion of new spacers has been found; however Fabre *et al.* (2004) recently described a strain from the *M. tuberculosis* complex with a DR locus containing a completely different set of spacers.

In *Yersinia pestis*, Jansen *et al.* (2002) reported the existence of three CRISPR elements with the same repeat sequence. *Y. pestis* strains are classified into three biovars according to their ability to reduce nitrate and to ferment glycerol (Devignat, 1951). Since *Y. pestis* was first linked to plague by

Abbreviations: CRISPR, clustered regularly interspaced short palindromic repeat; DR, direct repeat; VNTR, variable number of tandem repeats; MLVA, multiple-locus VNTR analysis.

Yersin (1894), strains of biovar Antiqua have been generally isolated from Asia and Africa; Medievalis was found in Central Asia, and Orientalis worldwide. Pestoides are particular strains isolated in Central Asia, and have never been found associated with disease in humans (Anisimov *et al.*, 2004). The complete genomic sequences of *Y. pestis* CO92 (Parkhill *et al.*, 2001), biovar Orientalis, and of strain KIM (Deng *et al.*, 2002), biovar Medievalis, have been determined, as well as the sequence of a so-called 'microtus' strain (Zhou *et al.*, 2004). The deficiency of glycerol fermentation in CO92 was found to result from a microdeletion in the *glpD* gene, and all Orientalis strains investigated so far (Motin *et al.*, 2002; Pourcel *et al.*, 2004) have been shown to harbour the same defect. This confirms the initial proposition that the Orientalis phenotype is derived from the Antiqua phenotype. In contrast, the Medievalis phenotype has been associated with different mutation events in the *napA* gene (Pourcel *et al.*, 2004).

In the present work, the three CRISPR elements were analysed in a collection of *Y. pestis* strains from three biovars and different geographical origins and in nine *Yersinia pseudotuberculosis* strains. Members of the species *Y. pseudotuberculosis*, from which *Y. pestis* is thought to have recently evolved some 1500 to 20 000 years ago (Achtman *et al.*, 1999), also possess the three CRISPRs. A total of 109 CRISPR alleles (mostly of the most polymorphic locus, CRISPR YP1) were sequenced. We describe a collection of new spacers, some of them probably recently acquired by *Y. pestis* strains in a clearly polarized fashion. The majority of these spacers correspond to fragments of a prophage.

METHODS

Bacterial strains and isolates. Most of the strains are from the collection maintained by the French Ministry of Defence at Centre d'Études du Bouchet (CEB). Some originated from the CIP (Collection Institut Pasteur) and others came from French medical military institutions (Hernandez *et al.*, 2003). The *Y. pestis* strains were isolated mostly from patients in Dalat, Vietnam, between 1964 and 1967 (Orientalis O2a biovar, Pourcel *et al.*, 2004), and in Africa (Kenya, Congo), Kurdistan and Madagascar (Table 1). Additional reference strains and DNA isolates were from the Institute of Microbiology Federal Armed Forces, Munich (Germany), or were kindly provided by Dr H. Mollaret (Paris, France), Professor F. Allenberger (Vienna, Austria), Professor H. Tschäpe (Wernigerode, Germany), Dr E. Carniel (Paris, France) and Dr A. Rakin (Munich, Germany). The genotypes, biovar and geographical origin of the strains are indicated in Table 1. Thermolysates were prepared by heating a bacterial suspension in water for 30 min at 95 °C. The larger collection of 182 strains and isolates has been previously genotyped using multiple-locus variable-number-of-tandem-repeat (VNTR) analysis (MLVA) (Le Flèche *et al.*, 2001; Pourcel *et al.*, 2004). Sixty-one different genotypes are resolved, one corresponding to a Pestoides, 3 to African Antiqua strains, 3 to Asian Antiqua strains, 43 to Orientalis strains and 11 to Medievalis strains. An *in silico* MLVA typing of the recently sequenced microtus strain 91001 (Zhou *et al.*, 2004) was achieved using the tool described by Denoeud & Vergnaud (2004) and accessible at http://minisatellites.u-psud.fr/blast/blast_primers_multi.htm. The corresponding new MLVA pattern is represented by genotype 62.

CRISPR PCR amplification and sequencing. PCR reactions and analyses were performed as described by Le Flèche *et al.* (2001). One of the CRISPR elements, CRISPR YP1, corresponds to marker yp2769ms06 previously used as a VNTR marker (Le Flèche *et al.*, 2001; Pourcel *et al.*, 2004). Primer pairs were designed to amplify the second and third CRISPR locus, respectively yp2895ms76 (CRISPR YP2) L (5′-ATATCCTGCTTACCGAGGGT-3′) and R (5′-AATCAGCCACGCTCTGTCTA-3′), and yp1773ms77 (CRISPR YP3) L (5′-GCCAAGGGATTAGTGAGTTAA-3′) and R (5′-TTTACGCATTTTGCGCCATTG-3′).

For sequencing, a 60 µl PCR reaction was performed. Amplicon quality was checked on a 2 % agarose gel. Then the product was purified by PEG precipitation as described by Embley (1991). Sequencing was performed by MWG Biotech (Germany).

Sequence similarity analyses were performed using the NCBI BLAST with microbial genomes database (http://www.ncbi.nlm.nih.gov) and GenBank.

Data management. The data produced were stored using the BioNumerics software package version 3.5 (Applied-Maths).

RESULTS

CRISPR polymorphism in *Y. pestis*

In all three sequenced genomes, three CRISPR elements are found. One of them, called here YP1, is associated with the *cas1* and *cas3* genes and is located in the vicinity of the replication terminus (Fig. 1). The relative position of YP1, YP2 and YP3 varies in the three genomes due to large-scale DNA rearrangements. The sequences of the three CRISPR elements present in the CO92 *Y. pestis* genome are shown in Fig. 2(a). The same 28 bp repeated sequence is found interspaced with spacers, 32 or 33 bp in length, in all three elements. They start with a truncated (19 bp for YP1 but 22 bp for YP2) repeat and end with a perfect 28 bp repeat, followed by a conserved sequence (Fig. 2b). This sequence, containing a stretch of A/T, was called the 'leader' by Jansen *et al.* (2002), and is generally found in all the CRISPR-containing bacterial genomes analysed, at the end of every CRISPR locus.

CRISPR YP1 was previously used as a VNTR marker in our MLVA assay (Le Flèche *et al.*, 2001; Pourcel *et al.*, 2004), showing a high degree of size polymorphism with 10 different allele sizes. Antiqua and Medievalis strains had rather short alleles with PCR product sizes compatible with the presence of three to six motifs, and even a lack of PCR product in three strains (Table 1, genotypes 4 and 5). In contrast, Orientalis strains had allele sizes compatible with the presence of 7–11 motifs, with the exception of four strains possessing four spacers. Sequencing of the CRISPR YP1 locus was performed on 91 samples, including at least one isolate of each MLVA genotype. The alleles all show the same organization, i.e. the repetition of a conserved 28 bp sequence and a spacer (32–34 bp long, depicted as boxes in Fig. 3). The length of the locus as measured by PCR correlates exactly with the number of repeats and spacers (data not shown). The spacers, and consequently

Table 1. Origin and characteristics of the *Y. pestis* strains

Genotype	Isolates	Origin	Biovar	Strains used (including common names or aliases)
1	1	Georgia	Pestoides	8786-G
2	3	Kenya	Antiqua	CEB87-020 (Margaret), CEB87-026, CEB87-031
3	1	Congo	Antiqua	CEB87-021 (343)
4	1	Kenya	Medievalis	CEB87-027 (129M)
5	2	Congo, Kenya	Antiqua	CEB03-523 (Lita)
6	1	Russia	Antiqua	AR142 (5761)
7	1	Russia	Antiqua	AR320 (735)
8	2	China, Japan	Antiqua	KUMA
9	1	Indonesia	Orientalis	CEB02-285 (Java10)
10	1	Indonesia	Orientalis	Java9
11	2	Unknown	Orientalis	GB
12	1	India	Orientalis	195/P
13	1	South Africa	Orientalis	CEB87-030 (K120)
14	2	Brazil	Orientalis	CEB02-289 (Exu38)
15	1	United States	Orientalis	CO92
16	1	Senegal	Orientalis	CEB87-029 (M20)
17	1	Unknown	Orientalis	G32pgm+
18	2	Turkey, Madagascar	Orientalis	CEB87-033 (EV76)
19	1	Madagascar	Orientalis	CEB02-107 (6/69)
20	1	Unknown	Orientalis	CEB02-111
21	2	Germany, Madagascar	Orientalis	Hambourg 9, CEB02-250
22	5	Unknown	Orientalis	CEB02-113
23	1	India	Orientalis	CEB02-287 (Barabanki4)
24	1	India	Orientalis	CEB02-288 (Barabanki3)
25	1	Senegal	Orientalis	Thierno
26	1	Senegal	Orientalis	Fay
27	4	Vietnam	Orientalis	CEB02-450
28	17	Vietnam	Orientalis	CEB02-451, 02-455, 02-110, 87-022
29	3	Vietnam	Orientalis	CEB02-447
30	3	Vietnam	Orientalis	CEB03-030
31	3	Vietnam	Orientalis	CEB02-424
32	2	Vietnam	Orientalis	CEB87-028
33	2	Vietnam	Orientalis	CEB02-457
34	1	Vietnam	Orientalis	CEB02-425
35	1	Vietnam	Orientalis	CEB02-524
36	2	Burma	Orientalis	CEB02-286 (548)
37	1	Vietnam	Orientalis	CEB02-458
38	2	Vietnam	Orientalis	CEB02-540, 03-020
39	2	Vietnam	Orientalis	CEB02-418
40	34	Vietnam	Orientalis	CEB02-419, 02-406, 02-467, 02-299
41	1	Vietnam	Orientalis	CEB02-465
42	1	Vietnam	Orientalis	CEB03-014
43	1	Vietnam	Orientalis	CEB02-405
44	3	Vietnam	Orientalis	CEB02-227, 02-423
45	1	Vietnam	Orientalis	CEB02-247
46	6	Vietnam	Orientalis	CEB02-417
47	1	Vietnam	Orientalis	CEB02-449
48	40	Vietnam	Orientalis	CEB02-529, 02-468, 02-453, 02-527, 02-423, 03-016, 03-027, 02-446
49	1	Vietnam	Orientalis	CEB02-430
50	1	Vietnam	Orientalis	CEB02-460
51	1	Vietnam	Orientalis	CEB02-428
52	1	Kurdistan	Medievalis	CEB02-109 (PKR24)
53	1	Kurdistan	Medievalis	CEB02-502 (PKR65)
54	1	Kurdistan	Medievalis	CEB87-025 PKH-III-M13

Table 1. cont.

Genotype	Isolates	Origin	Biovar	Strains used (including common names or aliases)
55	1	Kurdistan	Medievalis	PKH4
56	1	Kurdistan	Medievalis	PKR25
57	1	Iran	Medievalis	CEB02-284 (PAR3)
58	1	Iran	Medievalis	CEB02-404 (PAR13)
59	1	Kurdistan	Medievalis	KIM
60	1	Kurdistan	Medievalis	CEB02-296
61	1	Kurdistan	Medievalis	CEB87-024 (PKR288)

the motifs, were given a letter code starting with the reference strains KIM (genotype 59) and CO92 (genotype 15), whose organization is 'abc' and 'abcdefgh' respectively (Fig. 3). Additional spacers found in the other isolates were given letter codes i to z. Most of the Medievalis isolates are 'abc' except for genotype 54 ('abci') and 61 ('abct'). Asian Antiqua strains KUMA and Yokohama (genotype 8) and two Russian strains (genotypes 6 and 7) are also 'abc', whereas the African Antiqua are 'abcdj' (genotype 2) or 'abcdjk' (genotype 3). Strains of genotypes 4 and 5 lack a CRISPR YP1 element [indicated by (del) in Fig. 4], as demonstrated by repeated attempts to amplify the locus with different primer pairs (data not shown). In the Orientalis isolates, the most frequent allele is 'abcdefgh', distributed into 27 genotypes representing 100 isolates. The frequent genotypes 28 and 40 are among this group. Eight alleles seen in one or two isolates possess an additional unique motif, added after motif h. In contrast, allele 'abcdefgho' is found in genotype 48, comprising 40 isolates, and genotypes 44 and 50. The only difference between MLVA genotypes 46–49 is the CRISPR YP1 size, genotypes 49 and 47 having motifs p or vw respectively added after o. In genotypes 11, 12 and 13, which possess motifs 'efgh', the first constant repeat is a normal (truncated) first repeat (Fig. 2a). The genotype 37 allele is missing motif e. The Pestoides strain is 'abcdem' and the microtus strain (genotype 62) is 'adf'.

At least one strain from each genotype was amplified at loci CRISPR YP2 and YP3. The CRISPR YP2 size polymorphism

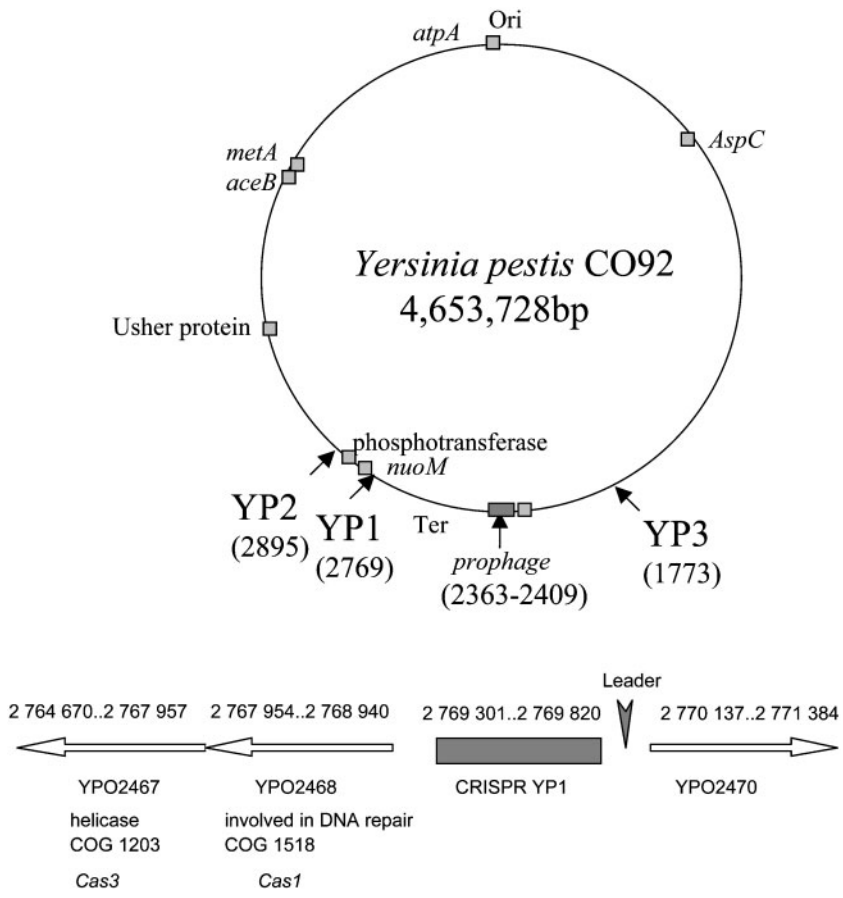

Fig. 1. Position of the CRISPRs and spacer homologues on the CO92 *Y. pestis* genome. The positions of the three CRISPRs and of the prophage are indicated in kb. Also shown are the origin (Ori) and termination (Ter) of replication. The genes on both sides of CRISPR YP1 and their exact position are indicated in the lower part of the figure.

(a) *Y. pestis* CRISPRs

```
Constant region 28bp                  spacers 32-33bp
VNTRyp2769ms06 CRISPR YP1
tttgattatTGCCTGTGCGGCAGTGAAC TCAGGGGACTGGCGAACAATGTCTTTCATGAT     a
TTTCTAAGCTGCCTGTGCGGCAGTGAAC GAAAAGGTAAGATGGGCAAGCTTCTAGTAGTT     b
TTTCTAAGCTGCCTGTGCGGCAGTGAAC ATTATCTGAATGGCATTTTCTTTGGCGCAGAT     c
TTTCTAAGCTGCCTGTGCGGCAGTGAAC TCGCCATTCCGTGAACCTGAGCGCGTTCGCGA     d
TTTCTAAGCTGCCTGTGCGGCAGTGAAC ATATTCTCGAGCGATAGCAATAGCCATTCCAC     e
TTTCTAAGCTGCCTGTGCGGCAGTGAAC TCGGTCAAACAAATTTAGGCGACGATTTAACA     f
TTTCTAAGCTGCCTGTGCGGCAGTGAAC AAAAAGAATTTGGGATTAAAGTTACCCATCAG     g
TTTCTAAGCTGCCTGTGCGGCAGTGAAC TCAATGCCTGAATCTCTGGCGTGATAGCTGCGG    h
TTTCTAAGCTGCCTGTGCGGCAGTGAAC

yp2895 CRISPR YP2
tctaTAAGCTGCCTGTGCGGCAGTGAAC TCTGTACGCATACCGCCATCTTGCATCAGTCT     a2
TTTCTAAGCTGCCTGTGCGGCAGTGAAC AGCAAAAATCTTAATTACATCTGATGATTTCGG    b2
TTTCTAAGCTGCCTGTGCGGCAGTGAAC TTTACGGCACGGCGAAAGATTCGGTTCTTGTC     c2
TTTCTAAGCTGCCTGTGCGGCAGTGAAC TTCTGGATAGGACAAATAGGATGATTGTATCAG    d2
TTTCTAAGCTGCCTGTGCGGCAGTGAAC AACGAACCCACGTAGAATTGCCATCACCGCCGG    e2
TTTCTAAGCTGCCTGTGCGGCAGTGAAC

yp1773 CRISPR YP3
ttattggGCTGCCTGTGCGGCAGTGAAC GTTATACCCCGCGCAGGGAGTGAAGCGTTGAC     a3
TTTCTAAGCTGCCTGTGCGGCAGTGAAC TTAAGTTCTTTTTGTCAGCATCTTTAATAAATA    b3
TTTCTAAGCTGCCTGTGCGGCAGTGAAC CTGAAATACAAATAAAATAAATCGTCGAACATA    c3
TTTCTAAGCTGCCTGTGCGGCAGTGAAC
```

(b) Conserved CRISPR leader sequence

```
YP1 TTTCTAAGCTGCCTGTGCGGCAGTGAACAGTAAGATAATACGA-TAACATCCTGTTTGTA
YP2 TTTCTAAGCTGCCTGTGCGGCAGTGAACAGTAAGATAATACGGGTAACAGACTGTTTGTA
YP3 TTTCTAAGCTGCCTGTGCGGCAGTGAAC-GTAAGATAATACGGATAACCCGATGTTTATC
    **************************** *************  ****    ***** *

YP1 AAAT-ACTTATTTCGCTAATGGGGAAAAAACCCTTTTTTTA---GACCACCGATAACCAC
YP2 AAAT-AATTCTTTCGCCAAAGGGTAAAAAATGATTTTTTT---AACCCTCGGTAAGCAG
YP3 AAATGAGCAATGGCGCAAAATGCGTAAAAACCCTTTTTTTAGTGAAATACC--TGAGTAG
    **** *    *  *** **  *   *****   *******     *    *  * *  *

YP1 AATGTAAAATCAATGAGTTAGCAGTAGCTAAAAAAATAGGGTCAGAACATAACTCATAAT
YP2 GATATAAAATCAATGAGTTAGCCATAGCTAAAAAAATAGGGTCAAAAAATGATTC-CCCT
YP3 CATA-AAAATCAATACGTTAGTCATAGTGATAAAAAGAGGGTCACAAGA--ATCGGGGGG
     **  *********  *****   ***  * ***** ******* ** *  *
```

Fig. 2. Sequences of the three CRISPR loci in CO92. (a) Sequence organization of the three CRISPR loci present in the CO92 strain. Eight spacer elements, named a to h, each one flanked by a repeat element (in upper-case italic), are observed at the CRISPR YP1 locus. Only five and three such spacers are observed at loci 2 and 3, respectively. The 28 bp repeats are the same at all three sites, except for the very first repeat, which lacks a few base-pairs. (b) On one side of the CRISPR structure, a conserved region of less than 200 bp called the leader sequence is observed. The leader sequences from the three sites are aligned; asterisks indicate identity in all three sequences. An A/T-rich stretch is underlined.

was much lower than that seen at locus YP1, with only four sizes corresponding to 3–6 motifs. We sequenced alleles of the Antiqua strains and those of Orientalis and Medievalis strains with a size which differs from, respectively, CO92 (5 motifs) and KIM (4 motifs). Nine types were found among the 17 alleles sequenced (Fig. 3). Four Orientalis isolates have an additional motif as compared to CO92 (genotypes 19, 40, 48, 49). The Pestoides strain and the Antiqua strains clearly differ, motifs j2 and k2 (interstitial) being present only in Pestoides. In Asian Antiqua isolates and in genotypes 4 and 5 lacking the CRISPR YP1 locus, YP2 is 'a2b2c2d2' as in the microtus strain (genotype 62) and in the group of Medievalis isolates including the KIM strain.

At the CRISPR YP3 locus, KIM and CO92 strains have the same three motifs 'a3b3c3'. No size polymorphism was observed in any of the *Y. pestis* strains tested by PCR. In contrast, the microtus strain possesses only motif a3, and the Pestoides strain has, in addition to motifs a3b3c3, motifs d3 and e3. In summary, partial sequence analysis reveals the existence of respectively 21, 9 and 3 different CRISPR YP1, YP2 and YP3 alleles.

CRISPR polymorphism in *Y. pseudotuberculosis*

Y. pseudotuberculosis represents an older, more diverged species from which *Y. pestis* probably emerged as a clone (Achtman *et al.*, 1999). The three CRISPRs are present at least in some members of this species, and can be amplified using the *Y. pestis* primer pairs. We analysed the CRISPRs of nine strains selected for their relative genetic proximity to *Y. pestis* as estimated by MLVA (data not shown). An important polymorphism was seen in CRISPR YP1, the alleles being in general larger than in *Y. pestis*, except for strain IP32802, which only possesses three motifs. The sequencing of these alleles produced a large collection of new spacers, 132 for nine isolates (data not shown). In strain IP32952 the first three motifs are a, b, c and they are the only examples of motifs shared with *Y. pestis*. Five additional motifs were shared by two or three *Y. pseudotuberculosis* isolates, the rest being unique to a given strain. The polymorphism of CRISPR YP2 and YP3 is also high, and some alleles are very large. Interestingly, the *Y. pseudotuberculosis* strain IP32802 has small alleles at these two loci as well, with only one spacer at CRISPR YP2 and CRISPR YP3.

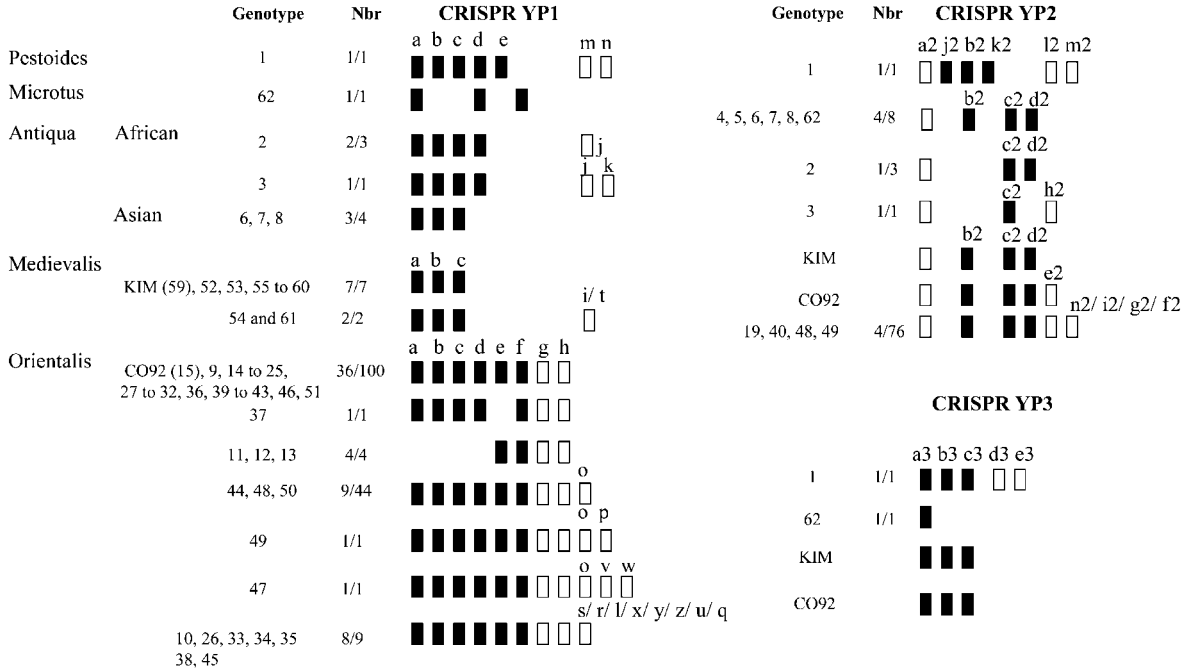

Fig. 3. The organization of CRISPRs. Arrangement of motifs (shown as boxes) in the different alleles, showing the polarized addition of new units. White boxes include spacers for which a homologue was found at another position in the genome. The genotype number refers to the classification by MLVA analysis (Table 1). 'Nbr' indicates the number of alleles sequenced over the total number of strains with the corresponding MLVA genotype.

The origin of new spacers

We analysed all the spacers by BLAST against the sequences of the CO92 and KIM genomes, and against the complete GenBank database. In Fig. 3 the different CRISPR alleles are shown as a succession of boxes, where white boxes indicate spacers for which a homologous sequence was found at a locus different from the three CRISPR loci in the *Y. pestis* genome. For the first spacers of CRISPR YP1, a, b, c, d, e and f, no sequence homology was found, whereas with motifs g to z, one copy of a homologous sequence, consisting of a portion of a gene or intergenic sequence, could be found in another region of the *Y. pestis* genome (Table 2). Some spacers in CRISPR YP2 and CRISPR YP3 also had matches elsewhere in the genome. Interestingly, 24 out of 31 *Y. pestis* spacers were found at a single locus, inside a 46 kb region corresponding to a defective lambdoid prophage. Spacers p, l and f2 were found in two contiguous genes, *aceB* and *metA*, separated by only 400 bp. Spacers j, r, t, y and z were found in four additional regions. In *Y. pseudotuberculosis*, only three spacers had a homologue in the *Y. pestis* genome at another locus, spacers psf and psg2 in the prophage, and spacer psc in the *atpA* gene.

DISCUSSION

We have analysed a collection of alleles at three CRISPR loci in *Y. pestis* and *Y. pseudotuberculosis*. By chance a number of features specific to the *Y. pestis* model proved to be key assets for the present investigation. They include first, the fact that the *Y. pestis* CRISPR loci are sufficiently small to be easily detected by PCR and sequenced; and secondly, the fact that a relatively large number of isolates associated with an epidemic outbreak, of highly similar genetic background, were available. By comparing the structure of the different loci, the motif arrangement and the origin of some of these motifs, we can propose a model for the evolution of CRISPR. This model can then be used to make predictions on the CRISPR YP1 organization of a putative ancestral *Y. pestis* strain.

CRISPR YP1 was investigated in detail by measuring its length in 26 isolates of the Orientalis O1 group and in 134 isolates of the Orientalis O2a group (Pourcel *et al.*, 2004), followed when relevant by sequencing of the locus. The vast majority of isolates have the YP1 allele 'abcdefgh', with six exceptions in the O1 group and 53 exceptions in the O2a group (44 of which are 'abcdefgho', observed in three genotypes, Fig. 4). All these strains show the same *glpD*, Orientalis-specific deletion compared to the Antiqua phenotype (Motin *et al.*, 2002; Pourcel *et al.*, 2004), which inactivates the gene. Therefore they must result from a clonal expansion which subsequently produced the two groups, O1 and O2a, represented here. The 'abcdefgh' allele is the only allele detected in both groups, and must have been present in a common ancestor. The other alleles differ from the 'abcdefgh' structures by simple deletion or insertion events. 'abcd' is missing in the closely related

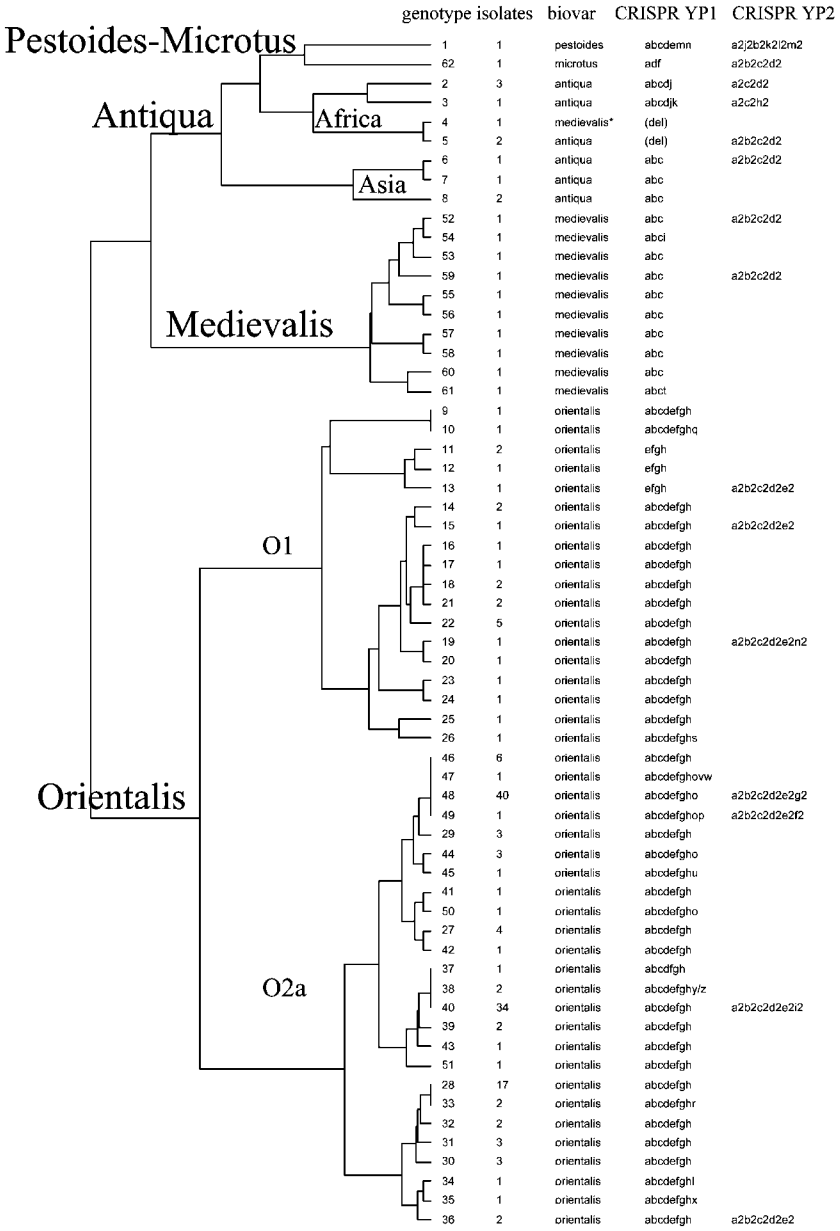

Fig. 4. Arrangement of motifs in CRISPRs compared to 24-markers MLVA. Clustering analysis was done using the categorical and Ward options. In contrast to the analysis of Pourcel *et al.* (2004), the ms06 (CRISPR YP1) typing data were not used for the clustering analysis. Consequently, genotypes which differ only by their CRISPR YP1 allele size are now identical. The biovars are indicated as Pestoides-Microtus (i.e. including the new microtus strain genotype 62), African or Asian Antiqua, Medievalis and Orientalis. The genotype 4 strain marked medievalis* differs from the classical Medievalis strains (Pourcel *et al.*, 2004). From left to right, the columns designate the genotype number, the number of isolates with an identical genotype, the biovar, the letter code for the CRISPR YP1 locus, and the letter code for the CRISPR YP2 locus.

genotypes 11, 12, 13, and 'e' is missing in genotype 37 (Fig. 4). All eleven insertion events are terminal additions after motif 'h' near the leader sequence. In two instances, new motifs were added to the frequent 'abcdefgho' variant, in strains with an otherwise identical MLVA genotype (genotypes 46, 47, 48, 49; Fig. 4). As in the DR of *M. tuberculosis* the order of the motifs is always the same and there is no duplication (van Embden *et al.*, 2000). In *Y. pestis*, the fact that the Orientalis O2a strains investigated here are of extremely limited geographical (Dalat, Vietnam) and temporal (years 1964–1967) origin clearly indicates that addition of spacers is an ongoing process.

These observations suggest the simple following rules for CRISPR evolution in general: (1) random deletions of one or more spacers and repeats may occur; (2) in contrast, the addition of new motifs is polarized and requires that the last constant region near the leader sequence is duplicated and that a piece of DNA about 32 bp in length (the spacer) is simultaneously copied and added; (3) the presence of identical spacers in a CRISPR allele reflects shared ancestry and does not result from independent events.

These rules have a predictive value. In the Pestoides strain, which can be considered as an outgroup according to MLVA analysis and current knowledge (Anisimov *et al.*, 2004), motifs a, b, c, d and e are present, indicating that a common ancestor of all *Y. pestis* biovars and Pestoides had already acquired these five motifs. Similarly, the sequence of the microtus genome reveals the presence of motifs 'adf'. Consequently, one might expect to find ancestor (Antiqua phenotype) strains with spacers a to f (Fig. 5). Because most

Table 2. Sequence of newly characterized spacers and their position outside the CRISPR loci

Chr. region	Spacer	Sequence	Position in CO92	YPO gene	Gene or product
I	z	AGATCGTGATGATAAACACACTTCCAACACAT	684 074–684 043	Between YPO0622 and YPO0623	
	y	TTGTAAGCCTTGGAACTCTGGACGCATTTTTC	684 222–684 191	YPO0623	*aspC*
II	g2	TAACGAGAAAGTCTTAATCTTTAATTCATCAG	2 357 592–2 357 561	Between YPO2075 and YPO2076	
		Between 2 363 016 and 2 409 384 putative defective lambdoid bacteriophage			
	n	TCACCAATGAGGGCGACCATGCCGAGGACTTG	2 370 263–2 370 232	Between YPO2094 and YPO2095	
	h2	TTAACGTAGCCAGGGCGTGTGGAACATGCCTAGT	2 374 566–2 374 597	YPO2101	Phage protein
	g	AAAAAGAATTTGGGATTAAAGTTACCCATCAG	2 374 731–2 374 762	YPO2102	Phage protein
	o	ACGTCATCCTGAAGGCTAGGCAGCTCGGCTTC	2 375 250–2 375 281	YPO2103	Unknown protein
	i2	TGGGACGCTTTACAGTCTGCACGTCTCTGAGT	2 375 515–2 375 546	YPO2103	
	i	GATGAGTAATGCCTTCAGCGCATTTCTCTTCA	377 2 7512 377 720	Between IS*285* and YPO2106	
	n2	ACATCTGGCCCACGACAAACATCGCGAACCGT	2 378 009–2 377 978	YPO2106	Phage protein
	d3	TTGGCAATCATGTTTGGCTGCGCTTGGTTAAAC	2 378 885–2 378 853	Between YPO2106 and YPO2108	
	s	CAGGAATGTTGGCCGCGATTGTTGCAGCTTGG	2 379 140–2 379 171	Between YPO2106 and YPO2108	
	e3	TGTCAGGCTGGGACTCTGATTTTTCAATTCGT	2 379 295–2 379 263	Between YPO2106 and YPO2108	
	k	TCAGTCCCGTTATGGTGCTGGTGTTGCCCGTAAG	2 379 357–2 379 389	YPO2108	Phage protein
	m2	TATGAGTGACAGCCGTTTTCACCACCGCCGTG	2 379 919–2 379 888	YP02108	Phage protein
	m	TTATCCGTGACCGACTCAAATACACGCTGGAACG	2 380 022–2 280 053	YPO2108	Phage protein
	a2	TCTGTACGCATACCGCCATCTTGCATCAGTCT	2 380 328–2 380 297	YPO2108	Phage protein
	u	AGCAATAAATCCCAAGGGGACAGCATGCTATT	2 380 514–2 380 545	Between YPO2108 and YPO 2109	
	h	TCAATGCCTGAATCTCTGGCGTGATAGCTGCGG	2 380 655–2 380 611	YPO2109	Phage protein
	l2	CGGACGCGGTGAAAACTATCCTGCAACGATTC	2 380 755–2 380 791	YP02109	Phage protein
	v	GAAATTGTGGGTGTAGATGTTGCAGACGCCTC	2 381 664–2 381 695	YPO2110	Phage protein
	q	TTGTTGCTAGTTGCATGTTTTTTCCAGCTATT	2 382 232–2 382 201	YPO2110	Phage protein
	e2	AACGAACCCACGTAGAATTGCCATCACCGCCGG	2 383 015–2 382 983	YPO2111	Phage protein
	w	TCTGACGTTGCCTGTGTTGCCGCTCTCGTATT	2 386 534–2 386 503	YPO2119	Phage tail protein
	x	CATTCTTAACGCCCCGCTCTGTTAGTGACAAA	2 389 780–2 389 749	YPO2119	Phage tail protein
	psf	GCTATATTCTCTCCGAGAAATCATTAGATAGTGG	2 393 501–2 393 468	YPO2125	Phage protein
	psg2	TATTGATGTAAGTGCGGCACCGGATATCGAGT	2 403 013–2 403 044	YPO2134	Phage tail protein
III	r	AGTACACCAAGTAGGCCGGTTAGCACCACCAT	2 857 781–2 857 812	YPO2544	*nuoM*
IV	j	ACATCACTAGCGGTATCGAACAATTGAGCGAG	2 885 306–2 885 274	YPO2569	Phosphotransferase
V	t	AGAGGGTATACGTATTTCCCAAACCATAATGG	3 286 069–3 286 100	YPO2943	Usher protein
VI	p	TCGCTTTGTATTTCTACCATAACTATCAGCAA	4 172 356–4 172 325	YPO3726	*aceB*
	l	CGGTCATCCCGAGTATGATGTCGATACGTTAG	4 173 576–4 173 545	YPO3727	*metA*
	f2	TGGAGTGGCGAGTTAGAGAGTAAACGTAGGAA	4 174 084–4 174 116	YPO3727	*metA*
VII	psc	GTAGAATACGTCGCCAGGATAGGCTTCACGAC	4 646 789–4 646 820	YPO4123	*atpA*
	j2	TGTTCTGAGAACGCTGCATGTCATTGCCTGGGG	Absent		
	k2	AGTGACTAACACGTCACAAATGTCGCCGTTCT	Absent		

Orientalis isolates investigated contain 'abcdefgh', it is tempting to speculate that the Antiqua strain, progenitor of the Orientalis strains, had the same allele. The shorter alleles 'abcdj' and 'abc' observed in the Antiqua and Medievalis strains investigated here would be the result of deletion events. Additional Antiqua strains, from Africa and more importantly Central Asia, will need to be investigated in order to prove these predictions, summarized in Fig. 5.

In the nine alleles of *Y. pseudotuberculosis* that were sequenced, 132 different motifs were found, only three of them, a, b, c, being present in *Y. pestis*. Work is in progress

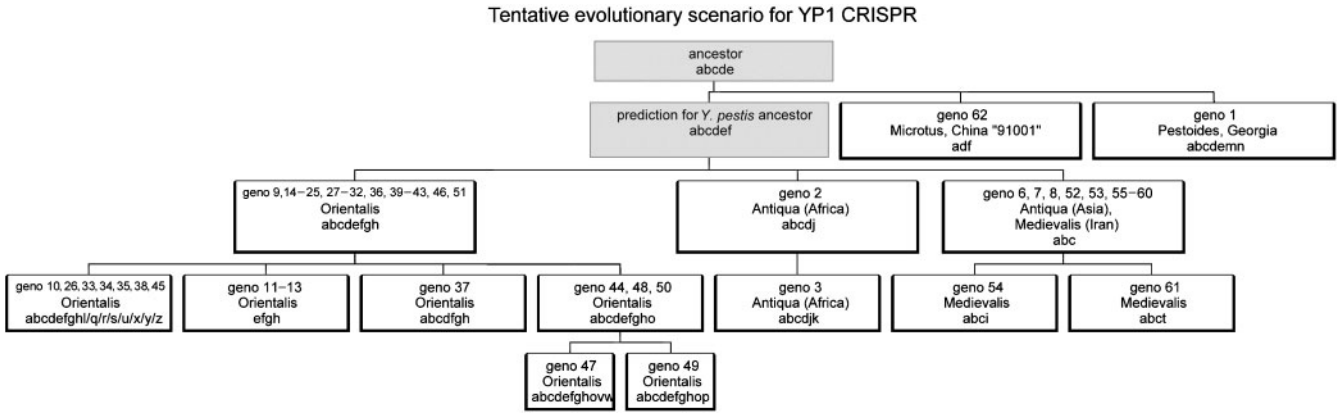

Fig. 5. Tentative evolutionary scenario for the CRISPR alleles. Shaded boxes are proposed missing links which have not been observed in our strain collection.

to analyse the CRISPR loci in a larger collection of *Y. pseudotuberculosis* isolates, with an emphasis on serotype O1 : b strains which are thought to be most closely related to *Y. pestis* (Skurnik *et al.*, 2000). The large number of unique motifs found in the *Y. pseudotuberculosis* isolates is reminiscent of the situation in *Campylobacter jejuni* (Schouls *et al.*, 2003). Similarly, in '*Mycobacterium canettii*', considered to be the ancestor of the *M. tuberculosis* complex, more than 50 spacers have been described that are not found in members of the *M. tuberculosis* complex (Fabre *et al.*, 2004; van Embden *et al.*, 2000). The addition of new spacers has not been observed in *M. tuberculosis* (van Embden *et al.*, 2000). The presence of insertion elements may indeed indicate that the DR locus is inactive.

Jansen *et al.* (2002) described a group of genes they called *cas*, present near one CRISPR of all the sequenced thermophilic archaea and in some bacteria. The *cas1* and the *cas3* genes are present at one side of the CRISPR YP1 (Fig. 1). Cas1 belongs to the family of proteins COG1518 predicted to be involved in DNA repair and is probably a DNase. Cas3 is believed to be a helicase and belongs to COG1203 (Jansen *et al.*, 2002). The *cas* genes are members of a cluster that has been extensively described by Makarova *et al.* (2002) independently of the study of CRISPRs, despite the fact that one CRISPR was always present close to this cluster. Makarova *et al.* (2002) believe that these genes are part of a DNA repair system particularly important in thermophilic archaea. Mojica *et al.* (2000) proposed that the CAS proteins might also be responsible for adding new repeats to the CRISPR loci. The fact that CRISPR YP1, the most active of the three elements, is the one associated with the *cas* genes supports this hypothesis. The adjacent leader sequence probably plays a role in this process.

More than two-thirds of the new spacers have a homologue in a region extending over 46 kb and corresponding to a prophage. A second preferential site extending over 2 kb contains the homologues of six spacers and corresponds to genes *aceB* and *metA*. The other spacers have homologues at different positions in the genome. This is believed to be the first time that the potential source of the CRISPR spacers has been identified. None of the spacers were found in the regions of instability described previously (Hinchliffe *et al.*, 2003; Radnedge *et al.*, 2002). The observation that *Y. pestis* CRISPR loci acquire new spacers from a prophage DNA is quite striking. Most of these phage sequences are also present in the *Yersinia enterocolitica* genome, although having only 80–90 % homology. It is believed that *Y. pestis* and *Y. enterocolitica* separated 41–186 million years ago (Achtman *et al.*, 1999). This suggests that the prophage was present in an ancestor but that the CRISPRs in *Y. pestis* have acquired new motifs only recently. It is interesting to note that the CRISPR YP1 and the prophage are located near the replication terminus (Fig. 1), a region in which levels of recombination are high.

We have found in the published sequence data additional evidence for a mechanism by which CRISPRs could acquire phage DNA. Hoe *et al.* (1999) described spacer sequences and organization in a CRISPR of *Streptococcus pyogenes*. Only one of the five sequenced *S. pyogenes* genomes, MIGAS, possesses a CRISPR. We performed a BLAST search with the *S. pyogenes* spacer sequences against the five sequenced genomes and found that seven out of the nine spacers described correspond to a phage-associated sequence, present in at least one of the genomes except that of MIGAS. Phage DNA constitutes up to 12·4 % of the *S. pyogenes* genome (Beres *et al.*, 2002) and is involved in recombination and horizontal transfer of new genes. There may be, in the case of strain MIGAS, a relationship between the presence of a CRISPR and the lack of a particular prophage. One possible explanation for that finding could be that CRISPRs are structures able to take up pieces of foreign DNA as part of a defence mechanism. In this view, it is tempting to further speculate that CRISPRs may represent a memory of past 'genetic aggressions'. The fact that most of the spacers described in other bacteria have no

homologue in the databases could still be explained by such a phage origin, as only a very small number of the existing bacteriophages have so far been sequenced.

The way in which the CRISPR loci appear to evolve in *Y. pestis*, and the frequency at which they acquire new motifs, at least within the Orientalis group of strains, are such that these loci may provide powerful and easy-to-use phylogenetic tools in complement to MLVA. It may be that the picking up of new spacers is not occurring at a uniform rate across the *Y. pestis* species, but rather that some unknown conditions are able to trigger an increased activity. In spite of the very limited number of Antiqua and Medievalis strains which could be investigated, MLVA data suggested the existence of two groups of Antiqua strains (Pourcel *et al.*, 2004). The first group, from Asia, represented by genotypes 6, 7 and 8, holds an intermediate position between the Medievalis and the Orientalis group; the second group comprises the African Antiqua strains. This grouping is supported by sequence analysis of the CRISPR loci (Fig. 4).

In addition to the biological relevance of their curious behaviour, analysis of the CRISPRs of *Y. pestis* provides a tool, comparable to spoligotyping for *M. tuberculosis* (Kamerbeek *et al.*, 1997), which might open the way to strain typing of ancient, degraded DNA, since a perfectly conserved multi-copy sequence (the repeat) can be used to amplify a variable library of very short spacers and since this kind of assay is not sensitive to the occasional misincorporation of an incorrect nucleotide during the amplification process.

ACKNOWLEDGEMENTS

Work on the typing and molecular epidemiology of dangerous pathogens is supported by the French Ministry of Defence. We thank H. Tschäpe, F. Allerberger, E. Carniel and A. Rakin for their generous gifts of purified genomic DNA. We are grateful to H. Neubauer and B. Holland for critical reading of the manuscript.

REFERENCES

Achtman, M., Zurth, K., Morelli, G., Torrea, G., Guiyoule, A. & Carniel, E. (1999). *Yersinia pestis*, the cause of plague, is a recently emerged clone of *Yersinia pseudotuberculosis*. *Proc Natl Acad Sci U S A* **96**, 14043–14048.

Anisimov, A. P., Lindler, L. E. & Pier, G. B. (2004). Intraspecific diversity of *Yersinia pestis*. *Clin Microbiol Rev* **17**, 434–464.

Beres, S. B., Sylva, G. L., Barbian, K. D. & 13 other authors (2002). Genome sequence of a serotype M3 strain of group A *Streptococcus*: phage-encoded toxins, the high-virulence phenotype, and clone emergence. *Proc Natl Acad Sci U S A* **99**, 10078–10083.

Deng, W., Burland, V., Plunkett, G., 3rd & 18 other authors (2002). Genome sequence of *Yersinia pestis* KIM. *J Bacteriol* **184**, 4601–4611.

Denoeud, F. & Vergnaud, G. (2004). Identification of polymorphic tandem repeats by direct comparison of genome sequence from different bacterial strains: a Web-based resource. *BMC Bioinformatics* **5**, 4.

Devignat, R. (1951). Variétés de l'espèce *Pasteurella pestis*. Nouvelle hypothèse. *Bull W H O* **4**, 247–263.

Embley, T. M. (1991). The linear PCR reaction: a simple and robust method for sequencing amplified rRNA genes. *Lett Appl Microbiol* **13**, 171–174.

Fabre, M., Koeck, J. L., Le Flèche, P., Simon, F., Hervé, V., Vergnaud, G. & Pourcel, C. (2004). High genetic diversity revealed by variable-number tandem repeat genotyping and analysis of *hsp65* gene polymorphism in a large collection of '*Mycobacterium canettii*' strains indicates that the *M. tuberculosis* complex is a recently emerged clone of '*M. canettii*'. *J Clin Microbiol* **42**, 3248–3255.

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* **10**, 1057–1065.

Hernandez, E., Girardet, M., Ramisse, F., Vidal, D. & Cavallo, J. D. (2003). Antibiotic susceptibilities of 94 isolates of *Yersinia pestis* to 24 antimicrobial agents. *J Antimicrob Chemother* **52**, 1029–1031.

Hinchliffe, S. J., Isherwood, K. E., Stabler, R. A. & 7 other authors (2003). Application of DNA microarrays to study the evolutionary genomics of *Yersinia pestis* and *Yersinia pseudotuberculosis*. *Genome Res* **13**, 2018–2029.

Hoe, N., Nakashima, K., Grigsby, D. & 7 other authors (1999). Rapid molecular genetic subtyping of serotype M1 group A *Streptococcus* strains. *Emerg Infect Dis* **5**, 254–263.

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.

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* **35**, 907–914.

Le Flèche, P., Hauck, Y., Onteniente, L. & 7 other authors (2001). A tandem repeats database for bacterial genomes: application to the genotyping of *Yersinia pestis* and *Bacillus anthracis*. *BMC Microbiol* **1**, 2.

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.

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* **17**, 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* **36**, 244–246.

Motin, V. L., Georgescu, A. M., Elliott, J. M. & 8 other authors (2002). Genetic variability of *Yersinia pestis* isolates as predicted by PCR-based IS*100* genotyping and analysis of structural genes encoding glycerol-3-phosphate dehydrogenase (*glpD*). *J Bacteriol* **184**, 1019–1027.

Parkhill, J., Wren, B. W., Thomson, N. R. & 32 other authors (2001). Genome sequence of *Yersinia pestis*, the causative agent of plague. *Nature* **413**, 523–527.

Pourcel, C., André-Mazeaud, F., Neubauer, H., Ramisse, F. & Vergnaud, G. (2004). Tandem repeats analysis for the high resolution phylogenetic analysis of *Yersinia pestis*. *BMC Microbiol* **4**, 22.

Radnedge, L., Agron, P. G., Worsham, P. L. & Andersen, G. L. (2002). Genome plasticity in *Yersinia pestis*. *Microbiology* **148**, 1687–1698.

Schouls, L. M., Reulen, S., Duim, B., Wagenaar, J. A., Willems, R. J., 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.

Skurnik, M., Peippo, A. & Ervela, E. (2000). Characterization of the O-antigen gene clusters of *Yersinia pseudotuberculosis* and the cryptic O-antigen gene cluster of *Yersinia pestis* shows that the plague bacillus is most closely related to and has evolved from *Y. pseudotuberculosis* serotype O : 1b. *Mol Microbiol* **37**, 316–330.

Sola, C., Filliol, I., Legrand, E., Lesjean, S., Locht, C., Supply, P. & Rastogi, N. (2003). Genotyping of the *Mycobacterium tuberculosis* complex using MIRUs: association with VNTR and spoligotyping for molecular epidemiology and evolutionary genetics. *Infect Genet Evol* **3**, 125–133.

van Embden, J. D., van Gorkom, T., Kremer, K., Jansen, R., van Der Zeijst, B. A. & Schouls, L. M. (2000). Genetic variation and evolutionary origin of the direct repeat locus of *Mycobacterium tuberculosis* complex bacteria. *J Bacteriol* **182**, 2393–2401.

Yersin, A. (1894). La peste bubonique à Hong-Kong. *Ann Inst Pasteur* **2**, 428–430.

Zhou, D., Tong, Z., Song, Y. & 14 other authors (2004). Genetics of metabolic variations between *Yersinia pestis* biovars and the proposal of a new biovar, microtus. *J Bacteriol* **186**, 5147–5152.

Zivanovic, Y., Lopez, P., Philippe, H. & Forterre, P. (2002). *Pyrococcus* genome comparison evidences chromosome shuffling-driven evolution. *Nucleic Acids Res* **30**, 1902–1910.