Chapter 2
Occurrence, Diversity of CRISPR-Cas Systems and Genotyping Implications

Christine Pourcel and Christine Drevet

Abstract This chapter describes the overall variability of CRISPR-Cas systems as observed in publicly available genomes and how this can be used to draw hypotheses on phylogenetic relationships between species and between strains of a given species. The fact that spacers are added sequentially at the leader end and that a given spacer is rarely acquired twice or duplicated are key elements for building hierarchical relationships between strains. Presence/absence of a given CRISPR locus and variability in the number of direct repeats and spacers in that locus between strains have been frequently reported, providing in some cases phylogenic information, but this polymorphism was extensively used for genotyping in only a few instances. The observation that not all strains possess a CRISPR locus in a given species precludes its use as a general typing tool. However, in some species, the degree of variability is a powerful marker of the species diversity and evolution. Through examples found among 1,434 published genomes of bacteria and archaea, different features of the CRISPR-Cas systems diversity will be highlighted.

Contents

C. Pourcel (✉) · C. Drevet
Institut de Génétique et Microbiologie, Université Paris-Sud, Bât 400,
91405 Orsay cedex, France
e-mail: christine.pourcel@u-psud.fr

C. Drevet
e-mail: christine.drevet@igmors.u-psud.fr

R. Barrangou and J. van der Oost (eds.), *CRISPR-Cas Systems*,
DOI: 10.1007/978-3-642-34657-6_2, © Springer-Verlag Berlin Heidelberg 2013

2.1 Introduction

CRISPRs are remarkable structures found in bacterial and archaeal genomes and known to interact with a set of genes called *cas* (Haft et al. 2005; Horvath and Barrangou 2010; Makarova et al. 2011a). Functional analysis of CRISPR-Cas systems in different species showed that it can play a number of roles, including defense against foreign genetic elements, regulation of lysogeny, regulation of biofilm formation, and others (Barrangou et al. 2007; Edgar and Qimron 2010; Zegans et al. 2009), as discussed in Chap. 10. A CRISPR locus is typically made of a succession of direct repeats (repeats) separated by spacers. *Cas* gene products interact with various CRISPR sequences and the target sequences to mediate the interference pathway (van der Oost et al. 2009). Spacers provide the specificity of the defense mechanism and mostly originate from phages or plasmids (Bolotin et al. 2005; Mojica et al. 2005; Pourcel et al. 2005). The occurrence of self-targeting spacers in some CRISPRs (1 in every 250 spacers in average) might lead to autoimmunity or be a part of a regulatory mechanism (Cui et al. 2008; Stern et al. 2010).

Investigation of publicly available genome sequences shows that CRISPRs are present in about 48 % of bacteria and 80 % of archaea, mostly on chromosomes but also on plasmids (Grissa et al. 2007a). *Cas* genes are found in the majority of CRISPR-containing genomes and when several CRISPRs of the same CRISPR-Cas system are present in a single genome, a single set of *cas* genes is generally clustered with one of the CRISPRs. Little is known on the mechanisms that drive multiplication of CRISPRs within a genome and acquisition and loss of spacers. New spacers are acquired by insertion next to the leader and are lost by internal deletion (Pourcel et al. 2005; Lillestol et al. 2006). It was proposed that when a CRISPR locus reaches a certain length, spacers must be lost and the older ones are preferably and more frequently lost first (Tyson and Banfield 2008). Although this may be true for certain CRISPRs in which the total number of spacer seems limited, in some extreme cases, several hundreds of spacers have been observed. Thus, the equilibrium between acquisition and loss appears to be highly different from one system to the other and this must be related to the ecology of the organism, its reliance on CRISPR-mediated immunity, and the pressure applied by foreign elements. A large body of information indicates that horizontal transfer of CRISPR and *cas* genes takes place between strains and between occasionally distant species and genera (Godde and Bickerton 2006; Horvath et al. 2008; Chakraborty et al. 2010; Shah and Garrett 2011). As a consequence, not all strains of a given species systematically possess the same sets of CRISPRs and *cas* genes. Owing to the huge amount of diversity observed in some CRISPR-Cas systems, examination of their elements (repeats, spacers, flanking sequences, and associated genes) provides important phylogenetic information (Grissa et al. 2008a).

With the advent of new sequencing technologies, more and more genomes are made available including multiple strains of a given species. Next-generation sequencing methods are well adapted to the investigation of CRISPRs and facilitate metagenomic analysis, which is an interesting source of sequences for both microorganisms and the viruses that infect them. In this evolving context, bioinformatic tools are needed to confidently and reliably identify and characterize CRISPRs and their elements (Grissa et al. 2009).

2.2 Assessing the Overall Diversity of CRISPRs

2.2.1 Bioinformatic Tools

2.2.1.1 Identification of CRISPRs

One important concern when trying to identify CRISPRs in a genome sequence is the exact definition of these structures. This is challenging, given the extensive sequence diversity of the CRISPR repeats, and the relative paucity of CRISPR-Cas systems that have been thoroughly characterized and shown to be active in the laboratory. A few specific programs have been developed for this purpose, the most used being CRISPRfinder (Grissa et al. 2007b), PILER-CR (Edgar 2007), and CRT (Bland et al. 2007). Additional programs were employed in different studies such as Pygram (Durand et al. 2006), LUNA (Lillestol et al. 2006), or Dotter (Sonnhammer and Durbin 1995). All programs perform as expected on typical CRISPR structures showing one or more of the following characteristics: five or more spacers, *cas* genes located nearby, homogeneous spacer length, and perfectly conserved repeats. Unfortunately, many CRISPRs do not typically meet these criteria. In this chapter, we will essentially discuss the performance of CRISPRfinder [which relies on the REPuter program (Kurtz et al. 2001)] which is at the basis of several other tools aimed at comparing and classifying CRISPRs.

CRISPR finder is based on specific features that were common to the well-characterized CRISPRs at the time of the program implementation and includes a tolerance margin: 23–55 bp repeats interspaced by sequences of 25–60 bp (spacer) and with spacer lengths between 0.6 and 2.5 the repeat length (Grissa et al. 2007b). The repeats are remarkably conserved in the majority of CRISPRs including those with a very large number of repeat-spacer units, but in some structures they show a high degree of heterogeneity such as in *Streptococcus sanguinis* SK36 (Genbank

ID CP000387), in several *Clostridium sp.* strains, or in *Amycolatopsis mediterranei*, for example. The parameters for defining the consensus repeat were chosen in order to cope with these unusual structures. CRISPRfinder returns all compatible structures and classifies them into "confirmed" (more than three units) and "questionable" (1–3 units) CRISPRs. In CRISPRdb, additional filters have been added to validate or exclude some structures, including a comparison of short CRISPRs' repeats to previously identified repeats and restriction on the spacer allowed length when the corresponding repeat has no classical flanking nucleotides such as GTTT or GAAC (Grissa et al. 2007a). Manual curation and further characterization often alleviate the issues inherent to false positives and occasionally false negatives. However, some of the CRISPR-like structures may correspond to other types of genetic elements such as, for example, portions of genes encoding proteins with repeated amino acid segments. Conversely, some of the shortest CRISPR-like structures containing one or two spacers may be true CRISPRs and they need to be evaluated using additional parameters. Therefore, a critical inspection of the results must still be made to discard sequences that are not true CRISPRs and validate short candidates. The presence of *cas* genes in the vicinity, or the identification of the source of one spacer or more [the proto-spacer (Deveau et al. 2008)] are probably the best criteria to fully validate a CRISPR structure. An indirect strong proof is the presence of an identical repeat in a fully validated CRISPR. Some tools are available for this purpose after running CRISPRfinder: spacers BLAST at NCBI to identify proto-spacers, search for *cas* genes using BLASTX, and search for CRISPRs with a significantly similar repeat in the database.

2.2.1.2 Tools to Analyze CRISPR Loci and Components

When comparing two strains with several CRISPR-Cas systems and/or several CRISPRs with the same repeat, it is necessary to individually identify each locus before listing the spacers. CRISPRcompar (Grissa et al. 2008b) has been developed to help in this classification by comparing sequences flanking CRISPRs that have similar repeats. The program is set to consider as similar, two loci with strictly identical repeats and flanking sequences showing 90 % identity over 200 base pairs. In some species, the accumulation of mutations may hide the common origin of two loci. Another important challenge concerns the identification and numbering of spacers especially when hundreds of unique sequences are to be compared and classified. Graphical representation of spacers have been used which can help to visually assess similarities between alleles but which will show limits when processing very large amounts of sequences (Horvath et al. 2008). CRISPRtionary was specifically developed to produce a catalog of spacers for a given CRISPR locus from submitted alleles, to number them and show their order in each allele (Grissa et al. 2008b). A detailed procedure to use these tools has been described (Grissa et al. 2009). Work is now in progress to provide a database of spacers that can be queried online.

2.2.1.3 CRISPR Databases

The challenge in building a database of CRISPRs is to faithfully identify these loci in order to be both exhaustive (reduce false negatives) and correct (eliminate false positives). This relies on the efficiency and quality of the program used to detect CRISPRs in a sequenced genome but also on manual curation since there is no perfect solution due to the diversity of CRISPR structures. At present two databases exist, CRISPRdb [http://crispr.u-psud.fr/ (Grissa et al. 2007a)] and CRISPI [http://crispi.genouest.org/ (Rousseau et al. 2009)], respectively listing 48.4 % (880/1,817) and 47.3 % (755/1,594) of bacterial genomes, and 83.7 % (105/123) and 80 % (96/120) of archeal genomes as possessing a CRISPR. Although these percentages are very similar, the identified CRISPR-like structures are often different, due to the parameters used to define the repeat sequence and also the repeat and spacer lengths.

CRISPRdb is a repertoire of the characteristics and locations of CRISPRs identified by the CRISPRfinder program in published bacteria and archaea genome sequences (chromosomes and associated plasmids) recovered from the RefSeq database released at the NCBI ftp site (ftp://ftp.ncbi.nih.gov/genomes/). Each sequence is submitted to the CRISPRfinder program and the resulting data is further analyzed by making use of the data previously stored in the database, in particular to validate some of the questionable CRISPRs. In addition, a manual curation step is performed after an initial automatic import to eliminate structures that possess typical characteristics of CRISPRs but correspond to tandem repeats.

CRISPI makes use of PYGRAM to identify CRISPRs, and apparently does not apply restrictions to the repeat and spacer lengths. Consequently, many CRISPRs present in CRISPI are not labeled as CRISPRs in CRISPRdb. For example, at the time of this publication, the CRISPI highlights displayed a CRISPR structure in *Xylella fastidiosa* M23 with five 8 bp-long repeats (70 % identity between repeats) and spacers 446–574 bp-long (the longest observed spacer in this database). This genome does not possess any *cas* gene, and it is very likely that the aforementioned structure is not a CRISPR. In the same highlights, the CRISPR with the longest repeats (five 92 bp-long repeats and four 27–45 bp-long spacers) found in *Shewanella putrefaciens* CN-32 corresponded to the tRNA-Asn locus. Many other differences exist between the CRISPRs identified by the two programs, in the number of CRISPRs, but also the sequence of repeats in a given locus. For example in *Cyanothece* sp. ATCC 51142, CRISPRfinder extracts a 37 bp repeat from CRISPR NC 01056-4, whereas Pygram finds a 45 bp repeat. CRISPRdb identifies 5 “confirmed” CRISPRs in *Sorangium cellulosum* “So ce 56”, CRISPI finds 7 CRISPRs, including one which is composed of four 17 bp“repeats” separated by 4 bp-long “spacers”. Our current knowledge of CRISPR-Cas systems clearly indicates that such structures cannot be legitimate CRISPR candidates since a four-nucleotide spacer cannot provide any specificity to the interference mechanism.

The parameters used in CRISPRdb to label a CRISPR-like structure as "confirmed" most probably accommodate the vast majority of existing CRISPRs. The smallest repeat recorded to date corresponds to the lower limit of 23 bp and was found only once in the archaeon *Ferroglobus placidus* DSM 10642 (four CRISPRs which contain 26, 21, 18, and 9 spacers, respectively). In a few instances, the repeat of CRISPRs containing a single spacer was wrongly estimated to be 23 bp-long but this could be corrected by comparison with longer CRISPR alleles in other strains of the same species. The largest repeat identified to date is 50 bp-long in *Weeksella virosa* (1 CRISPR harboring 20 spacers). This suggests that the higher limit of the program (set at 55 bp) is acceptable. Last, among the numerous "questionable" CRISPRs usually possessing one or two spacers, some might be indeed real CRISPRs and may be confirmed later when new genomes containing larger CRISPRs with identical repeats will be processed.

A recent addition to CRISPRdb is the possibility to view annotated *cas* genes in genomes harboring a CRISPR and to perform a BlastX analysis using a local database of Cas proteins extracted from the Uniprot database (http://www.uniprot.org/). Similarly, CRISPI displays a detailed list of CRISPR-associated genes with in the vicinity of each CRISPR.

CRISPRdb provides a list of repeats and spacers from published sequenced genomes. However, there is still a need for databases containing all the spacers that have been identified to date, including sequenced alleles as part of intraspecies diversity studies and spacers extracted from metagenomes. Specific databases are being constructed to record spacers of a given species and variations in CRISPR alleles. This is the case of the SpolDB4 database dedicated to *Mycobacterium tuberculosis*, and of the *Salmonella enterica* and *Legionella pneumophila* databases held at the Pasteur Institute in Guadeloupe http://www.pasteur-guadeloupe.fr:8081/SITVITDemo/ or in Paris http://www.pasteur.fr/recherche/genopole/PF8/crispr/CRISPRDB.html. Of note, the program iSpacer has been created by Aaron White (http://epilityblog.com/blog/) to compare large spacer libraries to the NCBI sequence database in order to search for proto-spacers. It was used to analyze a collection of spacers from *Pseudomonas aeruginosa* CRISPRs (Cady et al. 2011).

2.2.2 Diversity of CRISPRs Found in Published Genomes

As of June 2012, more than 1,800 bacterial and 120 archeal genomes have been publicly released allowing an assessment of the CRISPR diversity. New genomes are released on a continuous basis. Notwithstanding the current bias(es) in the currently sequenced bacterial and archaeal genomes, some key observations can be made which help in tracing the origin of CRISPR-Cas systems.

2.2.2.1 Repeat and Spacer Features

The available data in CRISPRdb were submitted to global analysis in order to investigate and characterize CRISPRs variability. The largest CRISPR was observed in *Haliangium ochraceum* DSM 14365 with 588 repeats. One set of *cas* genes and two other CRISPRs were found in this bacterium (with 190 and 37 repeats), spanning 75 kb. A second group of *cas* genes was found at another location in this genome but no CRISPR seems present in the vicinity of these genes. Interestingly, archaea and thermophilic bacteria as well as others living in extreme habitats frequently have 2 CRISPR-Cas systems and a large number of CRISPRs. The six members of the genus *Caldicellulosiruptor,* which contains the most thermophilic bacteria, have multiple CRISPRs and very large sets of *cas* genes (33 *cas* genes clustered at a single locus in *C. kristjansoni*). This suggests that CRISPR-Cas systems are an essential element for survival in these organisms. The largest number of CRISPRs is observed in the bacterium *Thermomonospora curvata* strain DSM 43183 with 15 CRISPRs and the archeon *Methanocaldococcus* sp. strain FS406-22 with 23 CRISPRs. *Thermincola* sp JR (Genbank CP002028) possesses three CRISPR-Cas systems made of three clusters of 2, 1, and 4 CRISPRs with different repeats and a different set of *cas* genes at each locus. *Truepera radiovictrix* DSM17093 possesses 4 different CRISPR-Cas systems for 9 CRISPRs at 7 different genetic loci. Some genera and/or species with multiple genome sequences available seem to completely lack CRISPRs such as *Chlamydia* sp. and *Chlamydophila* sp. or *Streptococcus pneumoniae*.

When analyzing the distribution of repeats into size groups, clear differences are seen between archaea and bacteria. Both show two main peaks at 29–30 bp and 36–37 bp but the smaller class of 24–25 bp is seen essentially in archaea, whereas the large repeats of 44 bp and more are only seen in bacteria (Fig. 2.1a). The diagram in Fig. 2.1b shows the length difference between repeats and spacers (average of spacers lengths) in relation to the repeat length. It suggests that a selective pressure exists for total repeat-spacer sizes of 60–75 bp. The diagram in Fig. 2.1c shows a tendency for archaeal CRISPRs to possess the larger number of repeats. When submitting the repeats list to UCLUST de novo clustering with option–id 0.50 at http://drive5.com/usearch (Edgar 2010), 185 clusters were found. 42 % of the sequences cluster into 10 groups containing more than 20 similar repeats (max 79). An analysis by Kunin et al. of 561 repeats from 195 genomes based on their folding score led to the definition of 33 clusters, 12 of which contained 10 or more members (Kunin et al. 2007). In our analysis, 13 % of the repeats belong to small groups (containing less than 5 sequences) including 7 % of single sequences. Among the latter, there are CRISPRs with a high number of spacers. Most of the repeats within a particular group show similar repeat length but there are exceptions. Indeed, internal insertions and deletions (INDELs) are frequently observed in the alignment between short and long repeats within a group. Of note, there is no repeat between 39 and 43 bp (Fig. 2.1a). The 44–50 bp repeats are clustered in a specific group. When the similarity is low, one side of the repeat is often better conserved.

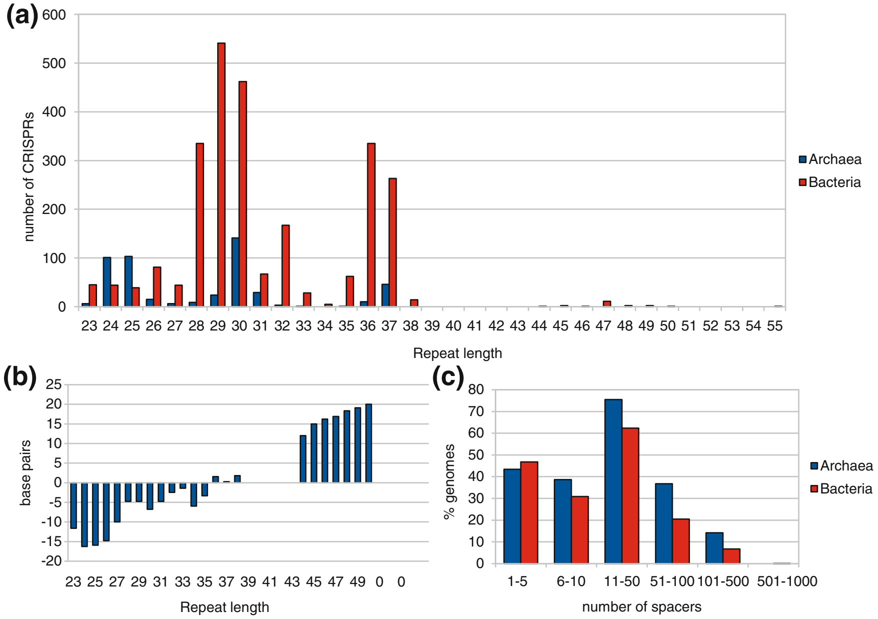


Fig. 2.1 Characteristics and distribution of repeats and spacers as observed in CRISPRdb. **a** Repeat length variability, **b** difference between repeats and average spacers length, **c** number of spacers

Works by Carte et al. (2008), Brouns et al. (2008), and Deltcheva et al. (2011) have shown that repeat sequences are targets for the cleavage by endoribonucleases. The large diversity of repeat sequences suggests that only part of the sequence is recognized by the Cas machinery and that the secondary structure is essential, in agreement with previous observations (Mojica et al. 2000; Jansen et al. 2002). Kunin et al. (2007) showed that among their 12 larger clusters, some but not all repeats were able to form stem-loop structures. To test whether a limited number of short sequences could be recognized in repeats, we performed a search for motifs using MEME (http://meme.sdsc.edu/meme/) with default parameters (zero or one motif per sequence, 3 maximum number of motifs to find). We found that 922 out of 1,041 repeats possessed one of three motifs and these motifs were differently localized over the repeat sequence (Fig. 2.2) (Bailey and Elkan 1994). In cluster 3 described by Kunin et al., the repeat possesses motif 1 forming the loop and motif 2 responsible for the formation of the stem. It may be of interest to note that the ten 44 bp and longer repeats (44, 46, 47, 48, 49, and 50 bp-long) show important similarity over 23 bp on one side (containing motif 1) and are associated with the csn1/Cas9 of Type II CRISPR-Cas systems (Fig. 2.3) (Makarova et al. 2011b). In 9 out of 10 cases, the corresponding CRISPRs are present in members of each of the three classes of the phylum *Bacteroidetes*.

The mechanism of acquisition of new spacers has been shown to involve insertion of a new repeat and a new spacer at the leader end (Barrangou et al. 2007;

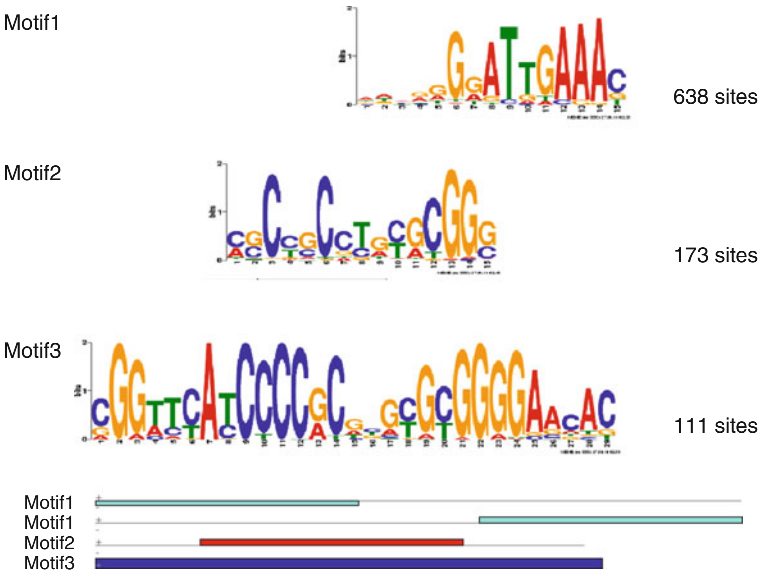
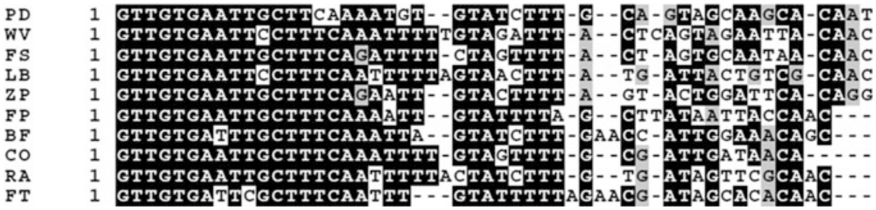


Fig. 2.2 Identification of common motifs in repeats. Three sequence logos produced by MEME at http://meme.sdsc.edu/, Motif1, Motif2, and Motif3, are observed in 638, 173, and 111 repeat types, respectively. At the *bottom* of the figure the diagram shows the most frequent position of the three motifs in the repeat sequences



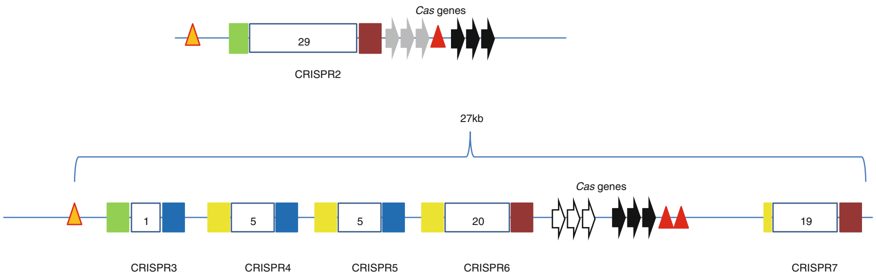
```
PD  1  GTTGTGAATTGCTTCAAAATGT--GTATCTTT-G--CA-GTAGCAAGCA-CAAT
WV  1  GTTGTGAATTCCTTTCAAATTTTTGTAGATTT-A--CTCAGTAGAATTA-CAAC
FS  1  GTTGTGAATTGCTTTCAGATTTT-CTAGTTTT-A--CT-AGTGCAATAA-CAAC
LB  1  GTTGTGAATTCCTTTCAATTTTTAGTAACTTT-A--TG-ATTACTGTCG-CAAC
ZP  1  GTTGTGAATTGCTTTCAGAATT--GTACTTTT-A--GT-ACTGGATTCA-CAGG
FP  1  GTTGTGAATTGCTTTCAAAATT--GTATTTTA-G--CTTATAATTACCAAC---
BF  1  GTTGTGATTTGCTTTCAAATTA--GTATCTTT-GAACC-ATTGGAAACAGC---
CO  1  GTTGTGAATTGCTTTCAAATTTT-GTAGTTTT-G--CG-ATTGATAACA-----
RA  1  GTTGTGAATTGCTTTCAATTTTTACTATCTTT-G--TG-ATAGTTCGCAAC---
FT  1  GTTGTGATTCGCTTTCAATTT---GTATTTTTAGAACG-ATAGCACACAAC---
```

Fig. 2.3 Alignment of the 10 long consensus repeat sequences. PD: *Prevotella denticola*, WV: *Weeksella virosa*, FS: *Fibrobacter succinogenes*, LB: *Leadbetterella byssophila*, ZP: *Zunongwangia profunda*, FP: *Flavobacterium psychrophilum*, BF: *Bacteroides fragilis*, CO: *Capnocytophaga ochracea*, RA: *Riemerella anatipestifer*, and FT: *Fluviicola taffensis*

Deveau et al. 2008). Several spacers can be added during the adaptation process. In the majority of CRISPR structures all the spacers are unique, but duplication of single or groups of spacers can be observed principally in long CRISPRs. For example, in *Spirochaeta caldaria* DSM7334 NC-015732-3, 28 different sequences are observed out of 52 spacers and only 14 are present once. Another example is found in *Myxococcus fulvus* HW-1: adjacent CRISPRs NC-015711-8 and NC-015711-9 with 41 unique sequences out of 101 spacers; 14 spacers are present once while others occur up to 9 times. Although it is possible that some spacers are acquired several times independently, the most probable mechanism for spacer duplication is via recombination or replication slippage. Also, given the challenges inherent to assembly of CRISPR loci, it might be necessary to validate some of the observed patterns.

2.2.2.2 Creation of New CRISPRs and Transfer of the System

Several CRISPRs with the same repeat and conserved flanking leader sequences can be found in some genomes, often next to each other, but the set of *cas* genes is present in a single copy without any spacer shared between these structures. The smallest CRISPR identified by CRISPRfinder in such genomes consists of two repeats surrounding a single spacer. To generate such a structure one must imagine a mechanism that copies the leader and the last repeat possibly by transcription from an adjacent promoter and reverse transcription. The frequent presence of transposase genes near the CRISPR-Cas loci suggests a role in the translocation process. As an example of a complex arrangement, Fig. 2.4 shows the schematic representation of six CRISPRs with very similar repeats (1 or 3 mismatches over 30 bp) and two sets of associated *cas* genes, in the bacterium *Flexistipes sinusarabici* DSM 4947.



CRISPR_id	Start Position	End Position	Number of spacers	DR consensus	System type	Cas
NC_015672_1	1011494	1014319	43	CTTTAAATCCCAGAATTGGGAACTAAAAC	I	Cas1 to Cas6
NC_015672_2	1428495	1430685	29	GTCGTAATCCCTTCAAATCAGGTCATTAATTCCAAT	III-A	Cas1, Cas2, Cas6, Csm3, Csm5
NC_015672_3	1709462	1709573	1	GTCGTAATCCCTTCTAATCAGGTCATGGATTCCAAT		
NC_015672_4	1709819	1710226	5	GTCGTAATCCCTTCTAATCAGGTCATGGATTCCAAT		
NC_015672_5	1710472	1710885	5	GTCGTAATCCCTTCTAATCAGGTCATGGATTCCAAT		
NC_015672_6	1711131	1712652	20	GTCGTAATCCCTTCTAATCAGGTCATGGATTCCAAT	III-B	Cas1, Cas2, Cas6, Crm2, Cmr3, Cmr4, Cmr5, Cmr6,
NC_015672_7	1734298	1735756	19	GTCGTAATCCCTTCAAGTCAGGTCATGAATTCCAAT		

Fig. 2.4 Schematic representation of CRISPR-Cas clusters with similar repeats in the bacterium *Flexistipes sinusarabici.* Six CRISPRs are represented as *boxes* showing the number of spacers. Flanking sequences are depicted with *colored boxes*. *Triangles* represent transposases genes. *Cas* genes are shown by *small white, gray, or black arrows*. Below are shown the position of the 7 CRISPRs found in the genome, the number of spacers, the sequence of the consensus repeat, the *cas* names, and the CRISPR-Cas system type

Analysis of published sequences clearly shows that different genera share similar CRISPR-Cas systems although these genera are not phylogenetically linked when using other genetic markers (Haft et al. 2005; Godde and Bickerton 2006; Chakraborty et al. 2010). A most intriguing observation is the presence of CRISPR-Cas in archae and in bacteria but not in Eukaryota, even the monocellular ones. It has been suggested that CRISPR-Cas systems which are present in the majority of archaea have been transferred to thermophilic bacteria and subsequently spread to other bacterial species. Indeed although the repeats seem to show specific characteristics in archaea, the *cas* gene systems are shared by members of the two domains (Makarova et al. 2011a). Plasmids may be vectors for CRISPRs-Cas systems as some of them have been found to possess complete systems. A total of 121 CRISPRs are carried by 58 plasmids out of 1,269 present in 50 taxons in CRISPRdb. For example, the *L. pneumophila* strain Lens possesses on a plasmid two CRISPRs and a complete set of *cas* genes (subtype I–F) similar to those found in the chromosome of the same strain but also in *P. aeruginosa*, *Yersinia pestis* and *Escherichia coli*.

Most of the *Staphylococcus aureus* sequenced genomes are devoid of CRISPR-Cas except for the livestock-associated ST398 lineage (Golding et al. 2010) and an ST75 early branching lineage (Holt et al. 2011). It is interesting to note that the CRISPR-Cas system is present next to the staphylococcal cassette chromosome (SCC) *mec*V subtype. A similar CRISPR-Cas system is present in some strains of *Staphylococcus epidermidis* (Gill et al. 2005) and *Staphylococcus lugdunensis*.

2.3 The Historical Use of CRISPR Polymorphism for Genotyping

2.3.1 Why and How to Perform Intra-Species Typing

A species may be defined as the sum of numerous strains that are classically differentiated by phenotypic and genetic characteristics, the precision of which depends on the question asked. When performing epidemiological investigations during outbreaks for example, it is critical to be able to trace the source of an infection similarly to forensics investigations in humans. Likewise, it is also important to evaluate the genetic complexity of a species and the speed at which it is evolving. The ultimate genotyping is the determination of the complete genome sequence of an organism. Alternatively, specific genetic polymorphisms can be used to compare strains, such as presence/absence of insertion elements (IS), single nucleotide polymorphisms (SNP), and variable number of tandem repeats (VNTR). The frequency of the genetic changes at these loci and the level of homoplasia (the independent occurrence of identical mutations) influence the informational value of the method and the possibility to use it to infer phylogenetic relationships between strains. In addition, the simplicity and cost of the method is of key importance when numerous samples must be simultaneously processed. Finally in order to be able to compare results between laboratories and to store the data into shared databases, genotypes in the form of a numerical code must be favored over gel/picture-based fingerprints. The characteristics of CRISPRs make them intriguing genetic markers for genotyping and population structure analysis but there is still a lot to be understood with regard to the molecular mechanisms that induce polymorphism in these sequences.

Several techniques have been used to assess the variability at a given CRISPR locus. Each of them will be described in the following paragraphs while discussing species for which a CRISPR-based genotyping scheme has been developed. Sequencing is the most straightforward but is not easily applicable to large alleles. In this case only portions may be amplified and sequenced (perhaps using primers designed in selected spacers). Because the evolution of an active CRISPR occurs via insertion of new spacers at the leader end, sequencing of this portion can be particularly informative. In contrast, sequencing the opposite end, which can contain spacers conserved across various strains, can be useful to cluster related phylogenetic group of strains. Hybridization to spacer-derived oligonucleotides, called “spoligotyping” had been used for some bacterial species but this will only investigate the presence of a pre-established selection of known spacers. Finally it is possible to rapidly differentiate alleles by high-resolution DNA melt curve analysis.

2.3.2 Intra-Species CRISPR Variations

2.3.2.1 *Mycobacterium tuberculosis* and *Mycobacterium canettii*

The first use of CRISPR polymorphism for diagnosis and genotyping was described in the *M. tuberculosis* complex (MTBC) which encompasses different species including *M. tuberculosis*, *M. africanum*, *M. bovis,* and the *M. cannettii* taxon (Kamerbeek et al. 1997). Groenen et al. (1993) were the first to analyze an MTBC CRISPR locus they called “DR” which showed polymorphism between different strains. They initially applied a PCR-based method called direct variable repeat PCR (DVR-PCR) derived from the minisatellite variant repeat PCR technique (MVR-PCR) (Jeffreys et al. 1991). This method, which is not suitable for routine use and high-throughput genotyping was replaced by a very elegant PCR and hybridization-based method called spacer oligotyping or “spoligotyping” (Kamerbeek et al. 1997). Oligonucleotides corresponding to 37 spacers present in the genome of *M. tuberculosis* strain H37Rv and 6 spacers of *M. bovis* BCG are bound to a membrane which is hybridized to amplification products generated by PCR between two repeats. Later, the addition of 25 new spacers improved the discriminatory power of the technique (van der Zanden et al. 2002).

Spoligotyping provides a pattern which can easily be coded and shared between laboratories. In the spolDB4 version of the international spoligotyping database, 1939 shared types (observed twice or more) were identified among 39,295 strains (Brudey et al. 2006).

Spoligotyping and derived methods such as the microbead-based hybridization assay (Zhang et al. 2010) can only indicate the presence/absence profile of known spacers. Sequencing of many MTBC isolates showed that this CRISPR locus is not acquiring new spacers and that polymorphism is only generated by loss of spacers (van Embden et al. 2000), some of which may be the result of IS element insertions (Groenen et al. 1993; Warren et al. 2002). Because all the members of the MTBC appear to possess a CRISPR locus and since the number of spacers remains low and is not increasing, spoligotyping is perfectly adapted to the analysis of this complex. The situation is different for members of the *M. cannettii* taxon. Indeed the first strains analyzed in detail appeared to possess a CRISPR with the same repeat as in *M. tuberculosis* but with a different set of spacers (van Embden et al. 2000). Later analysis of a larger collection of *M. canettii* isolates showed that many did not possess any CRISPR and others had new set of spacers (Fabre et al. 2004, 2010). This confirms the higher degree of diversity within the *M. canettii* taxon which is believed to be the most probable source species of the whole complex (Fabre et al. 2004).

Overall, spoligotyping has been central in the identification of clades in the MTBC, and it is a useful approach for phylogenetic studies (Filliol et al. 2003) but it has a limited value for evolutionary studies (Comas et al. 2009). Yet it remains the cheapest assay to rapidly classify strains.

2.3.2.2 *Yersinia pestis*

Yersinia pestis is a rather monomorphic species, highly pathogenic, and recently emerged (less than 20,000 years and may be not more than a few thousand years) from the more diverse *Yersinia pseudotuberculosis* species (Morelli et al. 2010; Bos et al. 2011). In the eight currently publicly available genomes, 1–3 CRISPRs (initially called Yp1, Yp2, Yp3, and subsequently renamed Ypa, Ypb, Ypc) have been observed with an identical repeat and with a single set of *cas* genes near one of the loci (Ypest I–F subtype). In 2005, the analysis of CRISPR polymorphism in a large collection of isolates mostly from a single epidemic episode provided key information on the mechanism of acquisition of new spacers and the origin of these spacers while opening the way to a new genotyping approach for epidemiological and phylogenetic studies (Pourcel et al. 2005).

This study was based on the analysis of amplicon size for three CRISPR loci in 182 isolates of which 142 originated from Dalat, Vietnam, during the 1964–1967 epidemic, and sequencing of 109 different alleles. Twenty-six unique spacers were observed for CRISPR Ypa, 14 for CRISPR Ypb, and 5 for CRISPR Ypc (Pourcel et al. 2005). The most variable locus is CRISPR Ypa, which is perhaps linked to the presence of *cas* genes in the immediate vicinity. When alleles were compared it appeared that common spacers were found at one end of the locus near the incomplete last repeat, whereas unique spacers were found at the other end, near the leader sequence. This is clearly shown by comparing the 8 CRISPR1 loci from published genomes, using CRISPRcompar. Spacers 8–12 at the leader end are unique, whereas spacers 1–7 at the trailer are shared by at least two strains. Genotyping by multiple locus VNTR analysis (MLVA) and CRISPR confirmed that strains from the Dalat epidemic in Vietnam with an almost identical MLVA type, could be distinguished by the presence of unique spacers located at the leader end of CRISPR Ypa alleles. The only plausible explanation was that they had been recently added to the CRISPR (Pourcel et al. 2004). This was the first evidence that addition of new spacers in CRISPR was polarized, a distinctive feature of CRISPR locus evolution, which was later confirmed in studies by Lillestol et al. (2006) and Barrangou et al. (2007). This observation is not only essential for understanding the mechanism of spacer acquisition but also to infer phylogenetic relationship between strains. Deletions of spacers on the contrary appear to be randomly distributed.

Since the first report on CRISPR polymorphism in *Y. pestis*, almost four hundred additional isolates have been studied (Cui et al. 2008; Riehm et al. 2012). More than 130 *Y. pestis* spacers have been identified so far among 600 isolates representing almost all known *Y. pestis* foci and including both subspecies *pestis* and *microtus* (Cui et al. 2008; Riehm et al. 2012). Apart from 14 ancestral spacers (6 in CRISPR1, 5 in CRISPR2 and 3 in CRISPR3) a proto-spacer can be found for all the others, majoritarily corresponding to a single prophage sequence, but also to a non-viral region in the chromosome. In most instances 100 % identity between the spacer and the proto-spacer is observed which raises the question of potential autoimmunity. It was suggested that autoimmunity is prevented by mutations in the CRISPR-Cas or in adjacent CRISPR motifs and that it does not constitute a regulatory mechanism (Stern et al. 2010). Interestingly, the proto-spacer-adjacent motif (PAM) (Horvath et al. 2008; Deveau et al. 2008; Mojica et al. 2009) shows a very weak conservation in *Y. pestis* proto-spacers (Cui et al. 2008; Mojica et al. 2009). This however may not be important for self versus non-self discrimination as demonstrated in *S. epidermidis* (Marraffini and Sontheimer 2010). Thus, it is possible that in *Y. pestis*, the CRISPR-Cas system serves another function apart from defense against foreign DNA.

2.3.2.3 *Yersinia pseudotuberculosis*

The three *Y. pestis* CRISPRs can also be found in most *Y. pseudotuberculosis* strains but the diversity of spacers is tremendously higher, reflecting the ancestral nature of the loci in this species and the position of *Y. pestis* within the much larger *Y. pseudotuberculosis* species. Actually, the whole *Y. pestis* species represents a single multilocus sequence type (ST) among 90 other STs uncovered so far in *Y. pseudotuberculosis* (Laukkanen-Ninios et al. 2011). In the initial study by Pourcel et al. (2005) 132 different spacers could be observed in the CRISPR Ypa alleles from 9 *Y. pseudotuberculosis* strains. The sequencing of 20 additional alleles identified 160 new spacers (Pourcel, unpublished results).

2.3.2.4 *Streptococcus pyogenes*

S. pyogenes, also called group A *Streptococcus,* is a species in which phages are the major source of genome diversification, constituting up to 12.4 % of the genome (Beres et al. 2002). Ten out of fifteen strains in available genomes possess one or two CRISPRs. Out of 41 unique spacers, 25 [CRISPR1 (11/17) CRISPR2 (14/24)] match with prophage sequences. Interestingly, a prophage is absent from a strain when a corresponding spacer is present in a CRISPR (Pourcel et al. 2005; Nozawa et al. 2011).

Early on, Hoe et al. (1999) investigated the interest of CRISPR variations for genotyping of *S. pyogenes* by sequencing 31 alleles from serotype M1 strains. Although deletion polymorphism was demonstrated, they showed that the informational value of the assay was lower than sequencing of the streptococcal inhibitor of complement (*sic*) gene. Since then no report of the use of CRISPR for typing of this species has been published.

2.3.2.5 *Campylobacter jejuni*

Campylobacter species, notably *C. jejuni* and *C. coli* are the leading cause of gastroenteritis worldwide. *Campylobacter* populations are characterized by high genetic diversity, weak clonality, and high level of recombination. Many genotyping techniques have been developed, of which multi locus sequence typing (MLST) is currently the leading method since its development by Dingle et al. (2001).

The genome sequence of 5 out of 6 strains contain a single CRISPR at the same locus as shown by CRISPRcompar. In 2003 Schouls et al. genotyped 184 strains with three different techniques, amplified fragment length polymorphism (AFLP), MLST, and sequencing of the CRISPR locus (Schouls et al. 2003a). They showed that 19 out of 184 tested strains did not yield a PCR product and 28 contained a CRISPR locus carrying a single repeat and thus no spacer. In the remaining strains 2–8 repeat-spacer units were found, yet 170 different spacers were detected which represents a high degree of polymorphism. There was a large inter-strain variability and the congruence between MLST, AFLP, and CRISPR typing was good. Because 26 % of strains were not typable by CRISPR sequencing it was concluded that this was not the method of choice for typing, but could be useful rather for subtyping of strains with similar AFLP or MLST profiles. Later, Price et al. (2007) developed a high-resolution DNA melt curve (HRM) analysis of the *C. jejuni* and *C. coli* CRISPR locus. They analyzed the CRISPR locus of 138 isolates containing between 1 and 13 spacers. Sequencing of 32 alleles produced 55 novel unique spacers. The CRISPR HRM genotype was determined for 29 isolates, producing highly reproducible and specific melt profiles. Further 125 isolates were then analyzed, demonstrating the power of the HRM method for discriminating "same" or "different" CRISPR genotypes.

2.3.2.6 *Streptococcus thermophilus* and Other Lactic Acid Bacteria

S. thermophilus is a lactic acid bacterium (LAB) widely used in milk fermentation processes as a starter culture. A deep investigation of CRISPR-Cas systems in 102 LAB genomes revealed the presence of eight distinct families in 46.1 % of strains (Horvath et al. 2009). A large diversity in *cas* genes, repeat and spacers content reflects the lateral origin, and the rapid evolution of CRISPR-Cas systems.

The genetic diversity of *S. thermophilus* has been investigated by different fingerprinting techniques, mostly by Random Amplified Polymorphic DNA (RAPD) and more recently by AFLP but these techniques produce genotypes that cannot be easily compared (Lazzi et al. 2009). Because of the commercial importance of this species it is necessary to generate comparable genotypes to identify genetic signatures that characterize specific strains. In that respect, the use of CRISPR polymorphism could be relevant. *S. thermophilus* genome sequences possess 1–4 CRISPR-Cas systems. CRISPR1 was found in all 124 strains analyzed whereas CRISPR2 was found in 59 out of 65 strains and CRISPR3 in 53 out of 66 strains (Horvath et al. 2008). A total of 39.5 % isolates carried all three loci. CRISPR1 shows the highest spacer diversity, followed by CRISPR3, due to internal deletions of spacers and additions at the leader end. Clustering of strains according to their spacer content can help in reconstructing phylogeny in this species. The authors suggest that the dynamic nature of CRISPR loci is potentially valuable for typing and comparative analysis of strains. Furthermore, the fact that *S. thermophilus* acquires new spacers at a high frequency upon challenge by phage infection allows the selection of multiresistant strains that show new and easily detectable genetic elements (Deveau et al. 2008). CRISPR-Cas systems have been analyzed in other species of LAB, notably in *Lactobacillus* (*Lb. acidophilus*, *Lb. casei*, *Lb. delbrueckii*, *Lb. paracasei*, *Lb. rhamnosus*, *and Lb. salivarius*), and *Bifidobacteria.*

2.3.2.7 *Corynebacterium diphtheriae*

Genotyping of *C. diphtheriae* by different methods and more recently by MLST has revealed a significant intraspecies diversity and the existence of clones although recombination hinders the structure of the population (Bolt et al. 2010). Strain NCTC 13129 which genome has been sequenced, possesses two CRISPRs, one with a 36 bp repeat and 7 spacers and another with a 29 bp repeat and 27 spacers, each locus being associated with a set of *cas* genes. Mokrousov et al. first described a spoligotyping method for this species making use of the polymorphism at the two CRISPR loci called DRA and DRB (Mokrousov et al. 2005, 2007). A reverse hybridization macroarray-based assay similar to the *M. tuberculosis* spoligotyping method was developed to study both DRA and DRB. A total number of 27 spacers (21 from DRB and 6 from DRA) were investigated, allowing to subdivide 156 strains of the 1990s ‘Russian epidemic clone’ into 45 spoligotypes. Later, 20 *C. diphteriae* biotype gravis strains collected in Belarus in 2005 in a suspected epidemic foci and showing the same ribotype were investigated by this method, displaying three different spoligotypes (Mokrousov et al. 2009). This confirmed that spoligotyping provides additional discrimination as compared to MLST. To generalize the method, it would be necessary to sequence alleles from more strains of diverse origin, in order to assess the polymorphism of existing loci and to determine whether there are evidences of spacer acquisition at one end of the loci (Mokrousov 2009).

2.3.2.8 *Escherichia coli*

Two CRISPR-Cas systems can be found in *E. coli,* one I-E (Ecoli) subtype (CRISPR2) and one I-F (Ypest) subtype (CRISPR4) (Haft et al. 2005; Makarova et al. 2011a). The presence and diversity of several CRISPRs belonging to the two systems was investigated by Diez-Villasenor et al. in a total of 100 strains representative of the species (including 28 sequenced genomes and 72 strains of the reference collection ECOR) (Diez-Villasenor et al. 2010). Sequencing of CRISPR2.1 and CRISPR2.3 spacers defined 58 and 52 alleles, respectively. Of 153 spacers analyzed in strains possessing the Type I-Ft CRISPR-Cas system, 100 were unique (47 out of 73 in CRISPR4.1 and 53 out of 80 in CRISPR4.2). Comparison of alleles allowed the clustering of strains possessing common spacers, but in the absence of data from another genotyping technique it was not possible to evaluate the informativity of CRISPR typing.

Another study (Touchon et al. 2011) investigated 263 strains and 27 sequenced genomes. The diversity of several Type I-E (CRISPR2) loci was assessed and compared to the phylogeny derived from MLST. A complete lack of CRISPR was observed in strains of the phylogenetic group B2, a major source of extra intestinal infection. CRISPRs shared common spacers within MLST groups and diversity was observed for example within clonal group C, although it appears that deletion rather than acquisition of new spacers was the source of polymorphism. Because there is no exact correlation between CRISPR arrangement and MLST grouping, probably related to horizontal transfer, CRISPR typing cannot be used as a general typing method for *E. coli,* but it could be useful in association with MLST to differentiate strains from a single clonal group, as illustrated for the C group.

2.3.2.9 *Pseudomonas aeruginosa*

The population structure of *P. aeruginosa* has been described as panmictic/epidemic to reflect the fact that only a few clones can be identified, related to antibiotic resistance or linked to specific clinical conditions such as cystic fibrosis (Romling et al. 2005). According to available genome sequence data, two CRISPR-Cas systems are observed: one Type I-F (Ypest) in the reference strain

UCBPP-PA14 and a Type I-E (Ecoli) in reference strain 2,192. The prevalence of these two subtypes was determined in collections of clinical isolates from different countries [unpublished and (Cady et al. 2011)]. In the work of Cady et al., 122 clinical isolates were investigated by amplification of *csy1* (Type I-F) and *cse3* (Type I-E) using PCR primers derived, respectively from strains PA14 and PA2192. Forty out of 122 isolates putatively harbor Type I-F and 6 % Type I-E. In all instances a single localization was found in the complete genome showing that the locus has not been inserted several times independently. Sequencing of all the loci resulted in 656 unique spacers. Among the spacers that showed 100 % identity to non-CRISPR sequences, 65 independent spacers were identical to lysogenic *P. aeruginosa* bacteriophages. We observed similar percentages of the two subtypes in a collection of 200 French isolates and also found a majority of spacers that match with lysogenic phage DNA. To determine whether CRISPR polymorphism could be used for genotyping, we compared the distribution of isolates possessing CRISPR-Cas systems to the clustering obtained using MLVA. In isolates genetically linked to strain PA14 and in all isolates from clone C found in cystic fibrosis patients (Romling et al. 2005), the Type I-FCRISPR-Cas system was found. The CRISPRs polymorphism in these clones allowed fine subtyping and also some phylogenetic reconstruction.

2.3.2.10 *Salmonella enterica*

Multiple serovars of *Salmonella enterica* subsp. *enterica* are associated with foodborne infection. Molecular techniques with high discriminatory power are necessary to investigate outbreaks. In the 15 available genome sequences 1–3 CRISPRs are detected. In a study of 28 sequenced genomes Fricke et al. (2011) investigated the polymorphism of these structures and observed a considerable variability which in part reflected the phylogeny of the species.

Liu et al. (2011a) described an “MLST” scheme in which they combined the sequence analysis of virulence genes *sseL* and *fimH* with that of two CRISPR loci. This assay was applied to the genotyping of 171 clinical isolates from nine *Salmonella* serovars. CRISPR profiles were converted into a CRISPR type and treated as an allele into the MLST scheme. Outbreak strains/clones could be differentiated by addition of CRISPR sequences as compared to using virulence genes only. Investigation of CRISPR polymorphism provided better discrimination of *Salmonella* serovar Enteritidis than PFGE and showed high epidemiologic concordance for all serovars screened except Muenchen. Later, these authors characterized 168 *Salmonella* serovar Enteritidis isolates using the assay now called CRISPR-MVLST to differentiate it from classical MLST (based on 7 housekeeping genes) leading to 27 sequence types (Liu et al. 2011b).

At present, several teams are investigating the polymorphism of CRISPRs in *Salmonella* and developing new hybridization-based assays for genotyping, which could complement the currently used methods. A database of *S. enterica* spacers is available for Blast at the Pasteur Institute http://www.pasteur.fr/recherche/genopole/PF8/crispr/CRISPRDB.html (Fabre et al. 2012).

The CRISPRs in *S. typhimurium* also show a high level of polymorphism which is being used for genotyping (Fabre et al. 2012).

2.3.2.11 *Erwinia amylovora*

E. amylovora, a phytopathogenic bacterium causing fire blight, has relatively low genetic diversity within the species. Commonly used genotyping methods provide poor discrimination of strains within local infested region (Rezzonico et al. 2011). Three CRISPR loci are present in the genome sequence of strain CFBP1430. A total number of 454 unique spacers were identified from the three CRISPR loci among 37 *E. amylovora* isolates (Rezzonico et al. 2011). The shortest CRISPR locus with 5 spacers was almost invariant. When combining the result for all three loci, 18 genotypes were identified. In this work, MEGA version 4.0 was used to infer phylogenetic relationships based on spacers present in strains (Tamura et al. 2007). McGeeh et al. identified 588 individual spacers among 85 isolates within the three CRISPR arrays present in *E. amylovora* (McGhee and Sundin 2012) and defined 28 distinct genotypes. The shortest locus with 5 spacers was invariant as shown in the study by Rezzonico et al. (2011), whereas variability was observed with the other two loci. CRISPR genotyping enabled the differentiation of strains that were shown, in previous studies, to belong to the same genotype using other methods. Furthermore, cluster analysis revealed the similarities and differences among isolates related to geographic source and host isolation.

2.3.2.12 Other Species

In *Mycoplasma gallisepticum* (Delaney et al. 2012), *S. agalactiae*, (Lopez-Sanchez et al. 2012), and *Microcystis aeruginosa* (Kuno et al. 2012), CRISPR locus variability offers new possibilities to perform population structure analysis. Based on our current understanding of CRISPR implementation for typing purposes, CRISPR polymorphism is being investigated in *L. pneumophila* and *Acinetobacter baumannii* to subdivide strains with similar MLST or MLVA genotype. Indeed some of the major *L. pneumophila* clonal complexes including Paris, Lens, and Corby possess one or two of Type I-E (Ecoli) or Type I-F (Ypest) CRISPR-Cas systems and spacer polymorphism can be observed that provide additional discriminatory power to the current genotyping methods (Ginevra et al. 2012). Likewise, clonal complex AYE in *A. baumannii* CRISPR shows spacer polymorphism which might be useful for subtyping (Hauck et al. PLoS one 2012).

2.3.3 Follow-up of CRISPR Diversity in Complex Microbiomes

The development of metagenome analyses provides increasing information about virus population dynamics and interaction with bacterial population, such as for example in acidophilic microbial biofilms (Tyson and Banfield 2008; Andersson and Banfield 2008), a microbial mat in hotsprings (Heidelberg et al. 2009), the oral cavity of a rat (van der Ploeg 2009), the ocean (Sorokin et al. 2010), the human gut (Minot et al. 2011), or in the rumen microbiome (Berg Miller et al. 2011). The results of these studies showed that CRISPR polymorphism reflect virus encounters, acquisition of new spacers, and locus evolution. As more metagenomic studies get underway, we anticipate that investigating CRISPR polymorphism will provide insights into microbial population structures, and their interplay with predatory viruses.

2.4 Discussion

The diversity of repeats, spacers and *cas* genes is amazing considering that the primary function of the system seems to be resistance against invasive DNA (Horvath and Barrangou 2010).

The presence of CRISPR-Cas immune systems in members of two of the three domains of life questions its origin and evolution. An evolutionary scenario based on the analysis of Cas protein families proposes that the system originated in thermophilic Archaea, and spread horizontally to bacteria, but numerous unanswered questions remain (Makarova et al. 2007, 2011b). It is also possible that such a primitive system existed in the last universal common ancestor LUCA (Glansdorff et al. 2008). Because the CRISPR-Cas systems evolve in response to pressures from invasive DNA and are deeply affected by horizontal transfer, it has been suggested that they function like a *bona fide* Lamarckian mechanism (Koonin and Wolf 2009).

Investigation of intra-species CRISPR-Cas systems polymorphism provides some clues on their evolution while sometimes constituting new genotyping tools. The different examples described above show that CRISPR polymorphism is elevated in some species, and can be exploited for rapid genotyping of even closely related strains, but cannot be the sole source of genetic diversity for bacterial genotyping. In some cases, it provides a rapid means to assign a strain to a phylogenetic group, or to identify a new branch. Analysis of CRISPR diversity in strains of species with a long history of evolution identifies large collections of spacers. On the contrary, inside a clonal complex, it appears that CRISPR variability may provide additional information for genotyping. For recently emerged species such as *M. tuberculosis* or *Y. pestis,* which have the complexity of a clonal complex, CRISPR typing provides important phylogenetic information. In many species in which only a fraction of the strains possess a CRISPR, it may be a valuable marker to identify subgroups of strains. Overall, it is necessary to increase the knowledge of intraspecies diversity to better understand the evolution rate of these structures, both by deletion or gain of spacers. In certain species this will depend on the selection forces applied by invasive DNA. In others, on the contrary, a CRISPR may just be slowly losing its spacers via internal deletion(s). When several CRISPRs are present, the locus next to a cluster of *cas* genes may be more active in terms of spacer acquisition, whereas loss of spacers by deletion may be similar (Pourcel et al. 2005; Horvath et al. 2008). This will have important consequences for phylogenetic studies.

In the future, the analysis of huge amounts of sequencing data from isolated microorganisms or from complex microbiomes will constitute a challenge. New bioinformatics tools will be necessary to identify and classify CRISPR elements and alleles. Efforts to maintain up-to-date databases will be needed in order to provide the community with high quality information.

Acknowledgments We thank Gilles Vergnaud for his helpful comments.

References

Andersson AF, Banfield JF (2008) Virus population dynamics and acquired virus resistance in natural microbial communities. Science 320:1047–1050

Bailey TL, Elkan C (1994) Fitting a mixture model by expectation maximization to dicover motifs in biopolymers. In: Proceedings of the second international conference on intelligent systems for molecular biology. pp 28–36

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

Beres SB, Sylva GL, Barbian KD, Lei B, Hoff JS, Mammarella ND, Liu MY, Smoot JC, Porcella SF, Parkins LD, Campbell DS, Smith TM, McCormick JK, Leung DY, Schlievert PM, Musser JM (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 USA 99:10078–10083

Berg Miller ME, Yeoman CJ, Chia N, Tringe SG, Angly FE, Edwards RA, Flint HJ, Lamed R, Bayer EA,and White BA (2011) Phage-bacteria relationships and CRISPR elements revealed by a metagenomic survey of the rumen microbiome. Environ Microbiol 14:207–227

Bland C, Ramsey TL, Sabree F, Lowe M, Brown K, Kyrpides NC, Hugenholtz P (2007) CRISPR recognition tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. BMC Bioinform 8:209

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

Bolt F, Cassiday P, Tondella ML, Dezoysa A, Efstratiou A, Sing A, Zasada A, Bernard K, Guiso N, Badell E, Rosso ML, Baldwin A, Dowson C (2010) Multilocus sequence typing identifies evidence for recombination and two distinct lineages of *Corynebacterium diphtheriae*. J Clin Microbiol 48:4177–4185

Bos KI, Schuenemann VJ, Golding GB, Burbano HA, Waglechner N, Coombes BK, McPhee JB, DeWitte SN, Meyer M, Schmedes S, Wood J, Earn DJ, Herring DA, Bauer P, Poinar HN, Krause J (2011) A draft genome of *Yersinia pestis* from victims of the Black Death. Nature 478:506–510

Brouns SJ, Jore MM, Lundgren M, Westra ER, Slijkhuis RJ, Snijders AP, Dickman MJ, Makarova KS, Koonin EV, van der Oost J (2008) Small CRISPR RNAs guide antiviral defense in prokaryotes. Science 321:960–964

Brudey K, Driscoll JR, Rigouts L, Prodinger WM, Gori A, Al-Hajoj SA, Allix C, Aristimuno L, Arora J, Baumanis V, Binder L, Cafrune P, Cataldi A, Cheong S, Diel R, Ellermeier C, Evans JT, Fauville-Dufaux M, Ferdinand S, Garcia de Viedma D, Garzelli C, Gazzola L, Gomes HM, Guttierez MC, Hawkey PM, van Helden PD, Kadival GV, Kreiswirth BN, Kremer K, Kubin M, Kulkarni SP, Liens B, Lillebaek T, Ho ML, Martin C, Martin C, Mokrousov I, Narvskaia O, Ngeow YF, Naumann L, Niemann S, Parwati I, Rahim Z, Rasolofo-Razanamparany V, Rasolonavalona T, Rossetti ML, Rusch-Gerdes S, Sajduda A, Samper S, Shemyakin IG, Singh UB, Somoskovi A, Skuce RA, van Soolingen D, Streicher EM, Suffys PN, Tortoli E, Tracevska T, Vincent V, Victor TC, Warren RM, Yap SF, Zaman K, Portaels F, Rastogi N, Sola C (2006) Mycobacterium tuberculosis complex genetic diversity: mining the fourth international spoligotyping database (SpolDB4) for classification, population genetics and epidemiology. BMC Microbiol 6:23

Cady KC, White AS, Hammond JH, Abendroth MD, Karthikeyan RS, Lalitha P, Zegans ME, O'Toole GA (2011) Prevalence, conservation and functional analysis of Yersinia and Escherichia CRISPR regions in clinical *Pseudomonas aeruginosa* isolates. Microbiology 157:430–437

Carte J, Wang R, Li H, Terns RM, Terns MP (2008) *Cas*6 is an endoribonuclease that generates guide RNAs for invader defense in prokaryotes. Genes Dev 22:3489–3496

Chakraborty S, Snijders AP, Chakravorty R, Ahmed M, Tarek AM, Hossain MA (2010) Comparative network clustering of direct repeats (DRs) and cas genes confirms the possibility of the horizontal transfer of CRISPR locus among bacteria. Mol Phylogenet Evol 56:878–887

Comas I, Homolka S, Niemann S, Gagneux S (2009) Genotyping of genetically monomorphic bacteria: DNA sequencing in *Mycobacterium tuberculosis* highlights the limitations of current methodologies. PLoS ONE 4:e7815

Cui Y, Li Y, Gorge O, Platonov ME, Yan Y, Guo Z, Pourcel C, Dentovskaya SV, Balakhonov SV, Wang X, Song Y, Anisimov AP, Vergnaud G, Yang R (2008) Insight into microevolution of *Yersinia pestis* by clustered regularly interspaced short palindromic repeats. PLoS ONE 3: e2652

Delaney NF, Balenger S, Bonneaud C, Marx CJ, Hill GE, Ferguson-Noel N, Tsai P, Rodrigo A, Edwards SV (2012) Ultrafast evolution and loss of CRISPRs following a host shift in a novel wildlife pathogen. *Mycoplasma gallisepticum* PLoS Genet 8:e1002511

Deltcheva E, Chylinski K, Sharma CM, Gonzales K, Chao Y, Pirzada ZA, Eckert MR, Vogel J, Charpentier E (2011) CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. Nature 471:602–607

Deveau H, Barrangou R, Garneau JE, Labonte J, Fremaux C, Boyaval P, Romero DA, Horvath P, Moineau S (2008) Phage response to CRISPR-encoded resistance in *Streptococcus thermophilus*. J Bacteriol 190:1390–1400

Diez-Villasenor C, Almendros C, Garcia-Martinez J, Mojica FJ (2010) Diversity of CRISPR loci in *Escherichia coli*. Microbiology 156:1351–1361

Dingle KE, Colles FM, Wareing DR, Ure R, Fox AJ, Bolton FE, Bootsma HJ, Willems RJ, Urwin R, Maiden MC (2001) Multilocus sequence typing system for *Campylobacter jejuni*. J Clin Microbiol 39:14–23

Durand P, Mahe F, Valin AS, Nicolas J (2006) Browsing repeats in genomes: Pygram and an application to non-coding region analysis. BMC Bioinform 7:477

Edgar RC (2007) PILER-CR: fast and accurate identification of CRISPR repeats. BMC Bioinform 8:18

Edgar RC (2010) Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26:2460–2461

Edgar R, Qimron U (2010) The *Escherichia coli* CRISPR system protects from lambda lysogenization, lysogens, and prophage induction. J Bacteriol 192:6291–6294

Fabre M, Koeck JL, Le Fleche P, Simon F, Herve 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

Fabre M, Hauck Y, Soler C, Koeck JL, van Ingen J, van Soolingen D, Vergnaud G, Pourcel C (2010) Molecular characteristics of "*Mycobacterium canettii*" the smooth *Mycobacterium tuberculosis* bacilli. Infect Genet Evol 10:1165–1173

Fabre L, Zhang J, Guigon G, Le Hello S, Guibert V, Accou-Demartin M, de Romans S, Lim C, Roux C, Passet V, Diancourt L, Guibourdenche M, Issenhuth-Jeanjean S, Achtman M, Brisse S, Sola C, Weill FX (2012) CRISPR typing and subtyping for improved laboratory surveillance of Salmonella infections. PLoS ONE 7:e36995

Filliol I, Driscoll JR, van Soolingen D, Kreiswirth BN, Kremer K, Valetudie G, Dang DA, Barlow R, Banerjee D, Bifani PJ, Brudey K, Cataldi A, Cooksey RC, Cousins DV, Dale JW, Dellagostin OA, Drobniewski F, Engelmann G, Ferdinand S, Gascoyne-Binzi D, Gordon M, Gutierrez MC, Haas WH, Heersma H, Kassa-Kelembho E, Ho ML, Makristathis A, Mammina C, Martin G, Mostrom P, Mokrousov I, Narbonne V, Narvskaya O, Nastasi A, Niobe-Eyangoh SN, Pape JW, Rasolofo-Razanamparany V, Ridell M, Rossetti ML, Stauffer F, Suffys PN, Takiff H, Texier-Maugein J, Vincent V, de Waard JH, Sola C, Rastogi N (2003) Snapshot of moving and expanding clones of *Mycobacterium tuberculosis* and their global distribution assessed by spoligotyping in an international study. J Clin Microbiol 41:1963–1970

Fricke WF, Mammel MK, McDermott PF, Tartera C, White DG, Leclerc JE, Ravel J, Cebula TA (2011) Comparative genomics of 28 *Salmonella enterica* isolates: evidence for CRISPR-mediated adaptive sublineage evolution. J Bacteriol 193:3556–3568

Gill SR, Fouts DE, Archer GL, Mongodin EF, Deboy RT, Ravel J, Paulsen IT, Kolonay JF, Brinkac L, Beanan M, Dodson RJ, Daugherty SC, Madupu R, Angiuoli SV, Durkin AS, Haft DH, Vamathevan J, Khouri H, Utterback T, Lee C, Dimitrov G, Jiang L, Qin H, Weidman J, Tran K, Kang K, Hance IR, Nelson KE, Fraser CM (2005) Insights on evolution of virulence and resistance from the complete genome analysis of an early methicillin-resistant *Staphylococcus aureus* strain and a biofilm-producing methicillin-resistant *Staphylococcus epidermidis* strain. J Bacteriol 187:2426–2438

Ginevra C, Jacotin N, Diancourt L, Guigon G, Arquilliere R, Meugnier H, Descours G, Vandenesch F, Etienne J, Lina G, Caro V, Jarraud S (2012) *Legionella pneumophila* sequence type 1/Paris pulsotype subtyping by spoligotyping. J Clin Microbiol 50:696–701

Glansdorff N, Xu Y, Labedan B (2008) The last universal common ancestor: emergence, constitution and genetic legacy of an elusive forerunner. Biol Direct 3:29

Godde JS, Bickerton A (2006) The repetitive DNA elements called CRISPRs and their associated genes: evidence of horizontal transfer among prokaryotes. J Mol Evol 62:718–729

Golding GR, Bryden L, Levett PN, McDonald RR, Wong A, Wylie J, Graham MR, Tyler S, Van Domselaar G, Simor AE, Gravel D, Mulvey MR (2010) Livestock-associated methicillin-resistant *Staphylococcus aureus* sequence type 398 in humans. Canada Emerg Infect Dis 16:587–594

Grissa I, Vergnaud G, Pourcel C (2007a) The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats. BMC Bioinform 8:172

Grissa I, Vergnaud G, Pourcel C (2007b) CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. Nucleic Acids Res 35:W52–W57

Grissa I, Bouchon P, Pourcel C, Vergnaud G (2008a) On-line resources for bacterial micro-evolution studies using MLVA or CRISPR typing. Biochimie 90:660–668

Grissa I, Vergnaud G, Pourcel C (2008b) CRISPRcompar: a website to compare clustered regularly interspaced short palindromic repeats. Nucleic Acids Res 36:W145–W148

Grissa I, Vergnaud G, Pourcel C (2009) Clustered regularly interspaced short palindromic repeats (CRISPRs) for the genotyping of bacterial pathogens. Methods Mol Biol 551:105–116

Groenen PM, Bunschoten AE, van Soolingen D, van Embden JD (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

Haft DH, Selengut J, Mongodin EF, Nelson KE (2005) A Guild of 45 CRISPR-Associated (Cas) Protein Families and Multiple CRISPR/Cas Subtypes Exist in Prokaryotic Genomes. PLoS Comput Biol 1:e60

Hauck Y, Soler C, Jault P, Merens A, Gerome P, Nab CM, Trueba F, Bargues L, Thien HV, Vergnaud G , Pourcel C. (2012) Diversity of acinetobacter baumannii in four french military hospitals, as assessed by multiple locus variable number of tandem repeats analysis. PLoS One 7:e44597

Heidelberg JF, Nelson WC, 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

Hoe N, Nakashima K, Grigsby D, Pan X, Dou SJ, Naidich S, Garcia M, Kahn E, Bergmire-Sweat D, Musser JM (1999) Rapid molecular genetic subtyping of serotype M1 group A *Streptococcus* strains. Emerg Infect Dis 5:254–263

Holt DC, Holden MT, Tong SY, Castillo-Ramirez S, Clarke L, Quail MA, Currie BJ, Parkhill J, Bentley SD, Feil EJ, Giffard PM (2011) A Very Early-Branching *Staphylococcus aureus* Lineage Lacking the Carotenoid Pigment Staphyloxanthin. Genome Biol Evol 3:881–895

Horvath P, Barrangou R (2010) CRISPR/Cas, the immune system of bacteria and archaea. Science 327:167–170

Horvath P, Romero DA, Coute-Monvoisin AC, 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

Horvath P, Coute-Monvoisin AC, Romero DA, Boyaval P, Fremaux C, Barrangou R (2009) Comparative analysis of CRISPR loci in lactic acid bacteria genomes. Int J Food Microbiol 131:62–70

Jansen R, Embden JD, Gaastra W, Schouls LM (2002) Identification of genes that are associated with DNA repeats in prokaryotes. Mol Microbiol 43:1565–1575

Jeffreys AJ, MacLeod A, Tamaki K, Neil DL, Monckton DG (1991) Minisatellite repeat coding as a digital approach to DNA typing. Nature 354:204–209

Kamerbeek J, Schouls L, Kolk A, van Agterveld M, van Soolingen D, Kuijper S, Bunschoten A, Molhuizen H, Shaw R, Goyal M, van Embden J (1997) Simultaneous detection and strain differentiation of *Mycobacterium tuberculosis* for diagnosis and epidemiology. J Clin Microbiol 35:907–914

Koonin EV, Wolf YI (2009) Is evolution Darwinian or/and Lamarckian? Biol Direct 4:42

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

Kuno S, Yoshida T, Kaneko T, Sako Y (2012) Intricate interactions between the bloom-forming cyanobacterium *Microcystis aeruginosa* and foreign genetic elements, revealed by diversified clustered regularly interspaced short palindromic repeat (CRISPR) signatures. Appl Environ Microbiol 78:5353–5360

Kurtz S, Choudhuri JV, Ohlebusch E, Schleiermacher C, Stoye J, Giegerich R (2001) REPuter: the manifold applications of repeat analysis on a genomic scale. Nucleic Acids Res 29: 4633–4642

Laukkanen-Ninios R, Didelot X, Jolley KA, Morelli G, Sangal V, Kristo P, Brehony C, Imori PF, Fukushima H, Siitonen A, Tseneva G, Voskressenskaya E, Falcao JP, Korkeala H, Maiden MC, Mazzoni C, Carniel E, Skurnik M, Achtman M (2011) Population structure of the *Yersinia pseudotuberculosis* complex according to multilocus sequence typing. Environ Microbiol 13:3114–3127

Lazzi C, Bove CG, Sgarbi E, Gatti M, La Gioia F, Torriani S, Neviani E (2009) Application of AFLP fingerprint analysis for studying the biodiversity of *Streptococcus thermophilus*. J Microbiol Methods 79:48–54

Lillestol RK, Redder P, Garrett RA, Brugger K (2006) A putative viral defence mechanism in archaeal cells. Archaea 2:59–72

Liu F, Barrangou R, Gerner-Smidt P, Ribot EM, Knabel SJ, Dudley EG (2011a) Novel virulence gene and clustered regularly interspaced short palindromic repeat (CRISPR) multilocus sequence typing scheme for subtyping of the major serovars of *Salmonella enterica* subsp. *enterica*. Appl Environ Microbiol 77:1946–1956

Liu F, Kariyawasam S, Jayarao BM, Barrangou R, Gerner-Smidt P, Ribot EM, Knabel SJ, Dudley EG (2011b) Subtyping *Salmonella enterica* serovar *enteritidis* isolates from different sources by using sequence typing based on virulence genes and clustered regularly interspaced short palindromic repeats (CRISPRs). Appl Environ Microbiol 77:4520–4526

Lopez-Sanchez MJ, Sauvage E, Da Cunha V, Clermont D, Ratsima Hariniaina E, Gonzalez-Zorn B, Poyart C, Rosinski-Chupin I,and Glaser P (2012) The highly dynamic CRISPR1 system of *Streptococcus agalactiae* controls the diversity of its mobilome. Mol Microbiol 85:1057–1071

Makarova KS, Sorokin AV, Novichkov PS, Wolf YI, Koonin EV (2007) Clusters of orthologous genes for 41 archaeal genomes and implications for evolutionary genomics of archaea. Biol Direct 2:33

Makarova KS, Haft DH, Barrangou R, Brouns SJ, Charpentier E, Horvath P, Moineau S, Mojica FJ, Wolf YI, Yakunin AF, van der Oost J, Koonin EV (2011a) Evolution and classification of the CRISPR-Cas systems. Nat Rev Microbiol 9:467–477

Makarova KS, Aravind L, Wolf YI, Koonin EV (2011b) Unification of Cas protein families and a simple scenario for the origin and evolution of CRISPR-Cas systems. Biol Direct 6:38

Marraffini LA, Sontheimer EJ (2010) Self versus non-self discrimination during CRISPR RNA-directed immunity. Nature 463:568–571

McGhee GC, Sundin GW (2012) *Erwinia amylovora* CRISPR Elements Provide New Tools for Evaluating Strain Diversity and for Microbial Source Tracking. PLoS ONE 7:e41706

Minot S, Sinha R, Chen J, Li H, Keilbaugh SA, Wu GD, Lewis JD, Bushman FD (2011) The human gut virome: inter-individual variation and dynamic response to diet. Genome Res 21:1616–1625

Mojica FJ, 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

Mojica FJ, Diez-Villasenor C, Garcia-Martinez J, Soria E (2005) Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. J Mol Evol 60:174–182

Mojica FJ, Diez-Villasenor C, Garcia-Martinez J, Almendros C (2009) Short motif sequences determine the targets of the prokaryotic CRISPR defence system. Microbiology 155:733–740

Mokrousov I (2009) *Corynebacterium diphtheriae*: genome diversity, population structure and genotyping perspectives. Infect Genet Evol 9:1–15

Mokrousov I, Narvskaya O, Limeschenko E, Vyazovaya A (2005) Efficient discrimination within a *Corynebacterium diphtheriae* epidemic clonal group by a novel macroarray-based method. J Clin Microbiol 43:1662–1668

Mokrousov I, Limeschenko E, Vyazovaya A, Narvskaya O (2007) *Corynebacterium diphtheriae* spoligotyping based on combined use of two CRISPR loci. Biotechnol J 2:901–906

Mokrousov I, Vyazovaya A, Kolodkina V, Limeschenko E, Titov L, Narvskaya O (2009) Novel macroarray-based method of Corynebacterium diphtheriae genotyping: evaluation in a field study in Belarus. Eur J Clin Microbiol Infect Dis 28:701–703

Morelli G, Song Y, Mazzoni CJ, Eppinger M, Roumagnac P, Wagner DM, Feldkamp M, Kusecek B, Vogler AJ, Li Y, Cui Y, Thomson NR, Jombart T, Leblois R, Lichtner P, Rahalison L, Petersen JM, Balloux F, Keim P, Wirth T, Ravel J, Yang R, Carniel E, Achtman M (2010) *Yersinia pestis* genome sequencing identifies patterns of global phylogenetic diversity. Nat Genet 42:1140–1143

Nozawa T, Furukawa N, Aikawa C, Watanabe T, Haobam B, Kurokawa K, Maruyama F, Nakagawa I (2011) CRISPR inhibition of prophage acquisition in *Streptococcus pyogenes*. PLoS ONE 6:e19543

Pourcel C, Andre-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

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

Price EP, Smith H, Huygens F, Giffard PM (2007) High-resolution DNA melt curve analysis of the clustered, regularly interspaced short-palindromic-repeat locus of *Campylobacter jejuni*. Appl Environ Microbiol 73:3431–3436

Rezzonico F, Smits TH, Duffy B (2011) Diversity, evolution, and functionality of clustered regularly interspaced short palindromic repeat (CRISPR) regions in the fire blight pathogen Erwinia amylovora. Appl Environ Microbiol 77:3819–3829

Riehm JM, Vergnaud G, Kiefer D, Damdindorj T, Dashdavaa O, Khurelsukh T, Zoller L, Wolfel R, Le Fleche P, Scholz HC (2012) *Yersinia pestis* lineages in Mongolia. PLoS ONE 7:e30624

Romling U, Kader A, Sriramulu DD, Simm R, Kronvall G (2005) Worldwide distribution of *Pseudomonas aeruginosa* clone C strains in the aquatic environment and cystic fibrosis patients. Environ Microbiol 7:1029–1038

Rousseau C, Gonnet M, Le Romancer M, Nicolas J (2009) CRISPI: a CRISPR interactive database. Bioinformatics 25:3317–3318

Schouls LM, Reulen S, Duim B, Wagenaar JA, Willems RJ, Dingle KE, Colles FM, Van Embden JD (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

Shah SA, Garrett RA (2011) CRISPR/Cas and Cmr modules, mobility and evolution of adaptive immune systems. Res Microbiol 162:27–38

Sonnhammer EL, Durbin R (1995) A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. Gene 167:GC1–GC10

Sorokin VA, Gelfand MS, Artamonova II (2010) Evolutionary dynamics of clustered irregularly interspaced short palindromic repeat systems in the ocean metagenome. Appl Environ Microbiol 76:2136–2144

Stern A, Keren L, Wurtzel O, Amitai G, Sorek R (2010) Self-targeting by CRISPR: gene regulation or autoimmunity? Trends Genet 26:335–340

Tamura K, Dudley J, Nei M, Kumar S (2007) MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Mol Biol Evol 24:1596–1599

Touchon M, Charpentier S, Clermont O, Rocha EP, Denamur E, Branger C (2011) CRISPR distribution within the *Escherichia coli* species is not suggestive of immunity-associated diversifying selection. J Bacteriol 193:2460–2467

Tyson GW, Banfield JF (2008) Rapidly evolving CRISPRs implicated in acquired resistance of microorganisms to viruses. Environ Microbiol 10:200–207

van der Oost J, Jore MM, Westra ER, Lundgren M, Brouns SJ (2009) CRISPR-based adaptive and heritable immunity in prokaryotes. Trends Biochem Sci 34:401–407

van der Ploeg JR (2009) Analysis of CRISPR in *Streptococcus mutans* suggests frequent occurrence of acquired immunity against infection by M102-like bacteriophages. Microbiology 155:1966–1976

van der Zanden AG, Kremer K, Schouls LM, Caimi K, Cataldi A, Hulleman A, Nagelkerke NJ, van Soolingen D (2002) Improvement of differentiation and interpretability of spoligotyping for *Mycobacterium tuberculosis* complex isolates by introduction of new spacer oligonucleotides. J Clin Microbiol 40:4628–4639

van Embden JD, van Gorkom T, Kremer K, Jansen R, van Der Zeijst BA, Schouls LM (2000) Genetic variation and evolutionary origin of the direct repeat locus of *Mycobacterium tuberculosis* complex bacteria. J Bacteriol 182:2393–2401

Warren RM, Streicher EM, Sampson SL, van der Spuy GD, Richardson M, Nguyen D, Behr MA, Victor TC, van Helden PD (2002) Microevolution of the direct repeat region of *Mycobacterium tuberculosis*: implications for interpretation of spoligotyping data. J Clin Microbiol 40:4457–4465

Zegans ME, Wagner JC, Cady KC, Murphy DM, Hammond JH, O'Toole GA (2009) Interaction between bacteriophage DMS3 and host CRISPR region inhibits group behaviors of *Pseudomonas aeruginosa*. J Bacteriol 191:210–219

Zhang J, Abadia E, Refregier G, Tafaj S, Boschiroli ML, Guillard B, Andremont A, Ruimy R, Sola C (2010) *Mycobacterium tuberculosis* complex CRISPR genotyping: improving efficiency, throughput and discriminative power of 'spoligotyping' with new spacers and a microbead-based hybridization assay. J Med Microbiol 59:285–294