CRISPR – a widespread system that provides acquired resistance against phages in bacteria and archaea

Rotem Sorek*, Victor Kunin and Philip Hugenholtz

Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, CA 94598

* Corresponding author: rsorek@lbl.gov

Abstract

Arrays of clustered, regularly spaced short palindromic repeats (CRISPR) are widespread in the genomes of many bacteria and almost all archaea. These arrays are composed of direct repeats sized 24-47 bp separated by similarly sized non-repetitive sequences (spacers). It was recently experimentally shown that CRISPR arrays, along with a group of associated proteins, confer resistance to phage. Following exposure to phage, bacteria integrate new spacer sequences that are derived from the phage genome. Acquisition of these spacers enables the bacterial cell to shutdown the phage attack, presumably by an RNA-interference-like mechanism. This Progress discusses the structure and function of CRISPRs and the implications of this new antiviral mechanism in bacteria.

Bacteriophages constitute the most populous life-forms on Earth[1]. In sea water, an environment in which phage abundance has been extensively studied, it has been estimated that there are 5-10 phage for every bacterial cell[2]. Despite being outnumbered by phage, bacteria proliferate and avoid extinction by using a battery of innate phage-resistance mechanisms such as restriction enzymes and abortive infection[3]. In this Progress article we describe the CRISPR system, a recently discovered defence mechanism, which is remarkable because it confers acquired phage resistance in Bacteria and Archaea. A hallmark of this system are arrays of short direct repeats interspersed by non-repetitive spacer sequences, the so-called Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR). Additional components of the system include CRISPR-associated (CAS) genes and a leader sequence (Fig. 1A).

Brief history of CRISPR research

The first report that described a CRISPR array, in 1987, was from Ishino *et al.* who found 14 repeats of 29bp interspersed by 32-33bp non-repeating spacer sequences[4, 5], adjacent to the isozyme converting alkaline phosphatase (*iap*) gene in *Escherichia coli*. In subsequent years similar CRISPR arrays were found in *Mycobacterium tuberculosis*[6], *Haloferax mediterranei*[7], *Methanocaldococcus jannaschii*[8], *Thermotoga maritima*[9] and other bacteria and archaea. The accumulation of sequenced microbial genomes allowed genome-wide computational searches for CRISPRs (the first such analysis was carried out by Mojica *et al.* in 2000[10]), and the most recent computational analyses revealed that CRISPRs are found in ~40% of bacterial and ~90% of archaeal genomes sequenced to date[11, 12] (Box 1).

In parallel to the initial appreciation of the abundance of CRISPRs[13], Jansen *et al*. identified four CRISPR-associated (CAS) genes that were almost always found adjacent to the repeat arrays[14]. Subsequent studies initiated by Koonin and colleagues[15, 16] and Haft *et al.*[17] uncovered 25-45 additional CAS genes appearing in close proximity to the arrays. The same set of genes is absent from genomes that lack CRISPRs.

Several hypotheses for the function of CRISPRs have been proposed. Early in 1995 Mojica *et al.* suggested that the repeats were involved in replicon partitioning, based on their observations that an increased copy number of the repeats in *Haloferax volcanii* resulted in altered replicon segregation[7]. This effect, however, was not reproduced in similar experiments carried out in *M. tuberculosis*[14]. Based on the presence of several CRISPR loci in some genomes Jansen *et al*. suggested that CRISPRs are mobile elements[14], while Makarova *et al*. suggested that the CRISPR system was involved in DNA repair, as many CRISPR-associated genes contained DNA-manipulating domains[15]. In 2005, three groups reported that the spacer sequences often contained plasmid- or phage-derived DNA, and hypothesized that CRISPRs mediate immunity against infection by extrachromosomal agents[18-20]. Bolotin *et al.* also reported on a negative correlation between the sensitivity of bacteria to phage infection and the number of CRISPR spacers in their genome[20]. Recently, Barrangou and colleagues have confirmed this hypothesis experimentally by showing that new spacers acquired following phage challenge confer resistance against the phage[21,49-50]. Their discovery is discussed in more detail below.

Structural features of CRISPR systems

CRISPR arrays and CAS genes (together forming the "CRISPR system") vary greatly among microbial species. The direct repeat sequences frequently diverge between species[14, 22], and an extreme sequence divergence is also observed in the CAS genes[16]. The size of the repeat can vary between 24bp and 47bp, with spacer sizes of 26-72bp[12]. The number of repeats per array can vary from two to 249 (in *Verminephrobacter eiseniae*[12]) and while many genomes contain a single CRISPR locus, the number of loci in *Methanocaldococcus jannaschii* reaches 18[8]. Finally, while in some CRISPR systems only six or fewer CAS genes were identified, others involve more than twenty[17]. Despite this great diversity most CRISPR systems have some conserved characteristics (Figure 1A):

Repeats: In a single array, repeats are almost always identical, with respect to size and sequence[14]. Despite being divergent between species, repeats can be clustered based on sequence similarity into at least twelve major groups[11]. Some of the larger groups contain a short (5bp-7bp) palindrome, and hence the word "palindromic" in the CRISPR acronym[14]. These palindromes were inferred as contributing to an RNA stem-loop secondary structure of the repeat[11], a hypothesis supported both by compensatory mutations existing in the repeats to maintain the stem structure, and by observations that the repeat-spacer array is transcribed into RNA[11, 23-25]. For other repeat groups evidence for RNA secondary structures is lacking. Apart from the structural feature, many repeats have a conserved 3' terminus of GAAA(C/G). Both the structural features and the conserved 3' motif were suggested to act as binding sites for one or more of the CRISPR-associated proteins[11].

Spacers: In any CRISPR system spacers are generally unique, with a few exceptions thought to result from segmental duplications[12]. Similarity searches of various CRISPRs consistently showed that many spacers frequently match (with high sequence identity) to phages and other extrachromosomal elements[16, 18-20, 25]. Mojica and coworkers have studied 4500 spacers from 67 microbial strains; 88 (2%) of them had similarity with known sequences, with more than 50% of these similar to a sequence found within a known phage and 10% within a plasmid[18]. Comparable numbers were reported in a separate study where 2156 spacers were examined[20]. The observation that only 2% of all spacers match any known sequence presumably reflect the general under-sampling of phage sequence space, and is in agreement with recent estimates of huge untapped phage environmental diversity[26]. Indeed, in lactic acid bacteria such as *S. thermophilus*, for which more than a dozen phage genomes have been isolated and sequenced, ~40% of the spacers had a homologue, matching either phage (75%) or plasmid (20%) sequences[20].

Spacers seem to be evenly distributed across the phage genomes and derive both from the sense (coding) and antisense (non-coding) orientations[18, 19, 21, 25], although one report suggested a preference towards spacers derived from one strand of the phage[20]. Two recent studies have reported on a short motif present in phage genomes 1-2 nucleotides

downstream to spacer-matching sequences[49-50]. This motif was hypothesized to be important for recognition, or cleavage, of phage sequences by the CRISPR system. The recognition motif can vary between CRISPR systems, being AGAA and GGNG for spacers found in CRISPR1 and CRISPR3 loci of *S. thermophilus*, respectively.

Leader: A sequence of up to 550bp is located 5' to most CRISPR loci, directly adjoining the first repeat[14, 25]. This common sequence was denoted the "leader" and is usually AT-rich[14]. Similar to the repeats, leaders lack an open reading frame and are generally not conserved between species; however, when several CRISPR loci are found in the same chromosome their leaders can be conserved[8, 27, 28]. When a new repeat-spacer unit is added to the CRISPR array, it almost always occurs between the leader and the previous unit, suggesting that the leader might function as a recognition sequence for the addition of new spacers[19, 21]. The leader was also suggested to act as the promoter of the transcribed CRISPR array, as it is found directly upstream of the first repeat[23, 24].

CAS genes: Two recent studies have characterized the large set of gene families that are associated with CRISPR arrays[16, 17] so in this review only the general features of these genes are discussed. CRISPR systems have been divided into 7 or 8 subtypes: each subtype contains 2-6 different subtype-specific CAS (CRISPR-associated) genes. In addition, six core CAS genes (*cas1-6*) are found associated with multiple subtypes, although the identity of *cas5* and *cas6* was not agreed upon[16, 17]. The *cas1* gene (COG1518; TIGR00287) is especially noteworthy as it serves as a universal marker of the CRISPR system (found linked to all CRISPR systems except for that of *Pyrococcus abyssii*[16]). Additional genes that are more loosely associated with CRISPRs, such as members of the Repeat Associated Mysterious Protein (RAMP)[15, 17] superfamily that occur only in genomes that contain CRISPR systems but not necessarily nearby the CRISPR, were also characterized. Specific functional domains identified in Cas proteins include endonuclease and exonuclease domains, helicases, RNA- and DNA- binding domains, and domains involved in transcription regulation[14, 16, 17, 29].

CRISPR is an anti-phage defence system

Very recently Barrangou and coworkers demonstrated experimentally that in response to phage infection bacteria integrate new spacers that are derived from phage genomic sequences, resulting in CRISPR-mediated phage resistance[21] (Fig. 1). These authors infected the lactic acid bacterium *Streptococcus thermophilus* with two different phages and recovered nine phage-resistant mutants. By sequencing the CRISPR1 locus they showed that each of the phage-resistant mutants had independently acquired between 1 and 4 new repeat-spacer units at the leader-proximal end of the array, and that in all cases the spacers were derived from the genome of the challenging phage. When a spacer matched the phage sequence exactly [100% identity], the mutant was phage resistant; but when one or more nucleotide changes were detected between the spacer and the phage sequence, bacteria were phage-sensitive. Barrangou and colleagues then inserted these resistance conferring spacers into the CRISPR array of a phage-sensitive *S. thermophilus* strain, causing it to become phage-resistant; finally, deletion of the acquired spacers led the strain to become sensitive again.

Together, these results showed that inclusion of phage-derived spacers in CRISPR arrays confers resistance to phage. Interestingly, in the course of their experiments Barrangou and coworkers noted that a small population of phage retained the ability to infect the resistant mutants. Further sequencing of the phage genomes revealed that the phage had mutated so that their sequence was no longer identical to the spacers[21]. Resistant phage having sequences identical to spacers were also isolated, but the AGAA downstream recognition motif was mutated in their genome, further strengthening the hypothesis that this motif is important for CRISPR function[49]. The selective pressure imposed by CRISPR on the phage therefore leads to rapid changes in their genome, and provides a glimpse into how CRISPR might be involved in driving the extremely high evolutionary rates observed in phage.

To begin to study the protein machinery behind the CRISPR function, Barrangou and colleagues also inactivated two subtype-specific CAS genes in a phage-resistant strain of *S. thermophilus*. Inactivation of *csn1* (according to the nomenclature of Haft *et al.*; denoted cas5 in ref. 21), which contains an endonuclease motif, resulted in loss of resistance even in the presence of phage-derived spacers. Mutants that had a different *cas* gene inactivated (named cas7 in ref. 21; might correspond to *cas2* or *csn2* by the nomenclature of Haft *et al.*) retained phage resistance when their CRISPR contained a phage-matching spacer, but were impaired in developing resistance to new phages, perhaps pointing to a role for this gene in acquiring new spacers[21].

A model for CRISPR activity

The exact mechanism by which CRISPR systems silence extrachromosomal DNA is not known, but a key observation towards mechanism elucidation was made by Tang *et al.*[23, 24] who found, in *Archaeoglobus* and *Sulfolobus*, that the repeat-spacer array is transcribed into a single transcript, which is further processed into small RNA units, each having a repeat + spacer size. The cleavage position seems to reside in the middle of the repeat, suggesting that the processed small-RNA unit corresponds to a full spacer flanked by two half repeats (Fig. 1C). The existence of palindromic motifs within many repeats might indicate that the two half repeats attach to each other, with the spacer forming a loop.

The observation that CRISPRs are processed into small RNAs as well as the assemblage of DNA- and RNA-manipulating protein domains within CAS genes has led Makarova *et al*. to suggest that CRISPR functions via an RNA-silencing (RNAi)-like mechanism[16]. This mechanism has been well-characterized in functioning as a defence against RNA viruses and transposable elements in eukaryotes[30]. In eukaryotic RNAi systems, long double stranded RNAs (dsRNA) of viruses are processed by a protein called dicer into small interfering RNAs (siRNAs) sized 21-22bp. These siRNAs are converted into single strands by the RNA-induced silencing protein complex (RISC), and the RISC-siRNA complex identifies viral mRNAs by base pairing, leading to their degradation by another nuclease denoted slicer[31]. According to the RNAi hypothesis the processed CRISPR spacers function as the microbial analogs of siRNAs. They bind to a RISC-like complex formed by Cas proteins and recognize the mRNA expressed from the foreign element by

base-pairing, resulting in subsequent degradation of the mRNA by other Cas proteins. Makarova *et al*. further proposed that *cas3*, a protein containing a helicase domain fused to a HD-nuclease domain, functions as the analog of dicer and processes the transcribed repeat-spacer array into siRNAs. *cas4*, which is a RecB-like nuclease, was suggested to be the analog of slicer[16]. A complication to this hypothesis stems from the observation that spacers can originate both from the sense and antisense strands of phage open reading frames[21]; a possible solution is that the spacers might first be converted into dsRNA so that both strands participate in silencing[16]. Indeed, Lillestøl *et al.* detected RNA transcripts corresponding to the both strands of the CRISPR repeats in *S. acidocaldarius*[25].

Evolution of CRISPR systems

CRISPR arrays can rapidly evolve, with CRISPR regions often being hypervariable between otherwise closely related strains[19]. A recent study revealed that within a nearly clonal population of *Leptospirillum* type II bacteria identified by metagenomics in an acidophilic microbial biofilm, evolution of the spacer collection in CRISPR regions is fast enough to promote cell individuality[32]. As new spacers are almost always inserted at the 5' end of the cluster next to the leader, the "older" spacers (farthest from the leader) are frequently common between isolates, while newer spacers are unique[19]. Deletion of repeat-spacer units is also frequently observed; this is necessary in order to prevent over-inflation of the CRISPR locus[12, 19, 49, 50]; however it is not clear whether such deletions occur actively or due to passive homologous recombination. Rare duplications of repeat-spacer units were also observed[12].

On a higher evolutionary scale CRISPR systems also greatly diversify. As indicated above, the repeats tend to vary between distantly-related species, but exceptions are often noted; for example, the arrays in *Escherichia coli* and *Mycobacterium avium* contain very similar repeats, although these two organisms are classified in different phyla[14]. This has been explained by horizontal gene transfer of CRISPR systems between organisms, a hypothesis supported by phylogenetic trees of core CAS genes[16, 17, 22]. Gene transfer was suggested to be mediated by megaplasmids, based on the identification of 10 such plasmids carrying CRISPR arrays[22, 33]. Interestingly, a CRISPR array was also found within a *Clostridium difficile* prophage, and it was suggested that the phage uses the CRISPR to limit dispersal of competing phages[34].

Current and future applications

Strain typing: More than a decade before the discovery that CRISPRs confer resistance to phage, Groenen *et al*. had spotted that these loci are among the most rapidly evolving structures in the genome of *Mycobacterium tuberculosis*, with strains varying in the number of repeats and in the presence and absence of specific spacers[35]. Based on this observation Kamerbeek and colleagues developed the spacer-oligotyping (spoligotyping) method for strain detection. In this method, probes for specific spacers are covalently bound to a membrane and hybridization patterns of labeled PCR products, primed from the CRISPR repeats, are measured[36](Fig. 2A). This has become the standard method for

genotyping of *M. tuberculosis* strains as part of ongoing efforts to control tuberculosis outbreaks[37, 38], and is also used for the typing of *Corynebacterium diphtheriae*[39]. Non-spoligotyping based methods for strain typing using CRISPR arrays are also used to study *Campylobacter jejuni*, *Thermotoga neapolitana* and other bacterial strains[40, 41], and Russell and colleagues recently filed a patent application on CRISPR-based methods to type *Lactobacillus* strains (US Patent application 20060199190).

Engineered defence against viruses: Many industries reliant on bacteria, such as the dairy and wine industries, are concerned about phage infection. Due to the high costs associated with phage-mediated culture losses, the dairy industry invests heavily in efforts to combat phage infection of dairy bacteria[3]. CRISPRs might offer a partial solution to this problem: by artificially adding spacers derived from conserved regions of known phages to the CRISPR array of the industrial bacteria, manufacturers could boost the immunity of their starter cultures against known phages (Fig. 2B). A recent patent application in this spirit was filed by Horvath et al. (US Patent Application 2007025097).

Selective silencing of endogenous genes: As noted above, it was proposed that the CRISPR system might be analogous to the eukaryotic RNAi system and that the spacers function as prokaryotic siRNAs by base-pairing with foreign mRNAs and promoting their degradation[16]. Should this hypothesis be confirmed, then manipulated CRISPR systems might revolutionize microbial physiology research, as they will allow selective gene knock-down without manipulation of the original microbial genome. Instead of knocking out the gene of interest, which is usually labour intensive, the same effect might be achieved by transforming a CRISPR-bearing plasmid into the organism of choice, with one of the spacers changed to match the studied gene (Fig. 2C). Moreover, the array nature of CRISPR could allow a simultaneous knockdown of multiple endogenous genes. Similar RNAi-based applications have revolutionized eukaryotic genetic studies; we envisage that CRISPRs would have a similar impact in the field of microbial genetics.

Outlook

Despite the recent advances in understanding the role of CRISPRs in microbial genomes, the mechanisms underlying CRISPR function are completely uncharacterized and hypotheses currently mainly rely on educated guesses based on bioinformatic analyses. Fundamental questions such as how new spacers are selected and inserted, how silencing of foreign DNA/RNA is achieved, and whether different CRISPR systems contain different functionalities are all expected to be addressed in the near future by the growing number of groups studying this system. Other questions that might be addressed in the future following extensive research on the system are detailed below.

The widespread occurrence of CRISPR systems in nearly half of all sequenced bacterial genomes points to their efficiency in providing protection against phage attacks (if this indeed is their predominant role). However, phages are the most populous biological entities on earth[1, 42] so it is plausible that phage have evolved various mechanisms to escape or inhibit CRISPRs. In fact, the high rates of evolution observed in CRISPR repeats and associated proteins indicate that an “arms race” between phage and CRISPR

systems might be occurring, in which mutations in CRISPR systems mediate escape from CRISPR shut-down mechanisms encoded by phages. If this hypothesis is correct, we would expect reports of phage-encoded anti-CRISPR systems. Hints that such a system exists can be found in the report by Peng *et al.* (2003) in which they described a *Sulfolobus* protein that specifically binds to the CRISPR repeat DNA, and induces an opening of the structure near the centre of the repeat[43]. We performed a homology search of this protein against all available microbial genomes, and found that its homologues are mainly found in bacterial prophages (Sorek R., unpublished data). We therefore propose that this protein might constitute part of an anti-CRISPR system encoded by phage; its exact role in this system is yet to be discovered.

The proposed analogy between the CRISPR system and the eukaryotic RNAi raises another possible important role for CRISPRs. In eukaryotes, RNAi functions both in silencing foreign elements through small interfering RNAs (siRNAs), as well as in endogenous gene regulation through genome-encoded micro-RNAs. Analogously, it is possible that CRISPR systems regulate endogenous functions in different bacteria. Indeed, 7%-35% of the spacers found in CRISPR arrays have homologues in the chromosomal DNA, which may indicate that CRISPR is being used to regulate expression of chromosomally-derived genes[18, 20, 50]. Moreover, the *devTRS* operon in *Myxococcus xanthus*, which encodes genes that are essential for spore differentiation inside the fruiting bodies of this species, is co-transcribed within a CRISPR operon, with DevS being a *bona fide* Cas5 protein[17, 44]. This might be another example of a CRISPR system regulating an endogenous mechanism.

Conclusions

Previously considered to be a simple family of repetitive elements, the CRISPR system has begun to take a centre stage in our understanding of acquired phage resistance in prokaryotes. The widespread presence of this system in many bacterial and archaeal phyla, as well as its extreme diversity, suggests that it may be one of the most ancient defence systems in the microbial world[16]. Future studies are expected to define how CRISPR functions and elucidate the role of this system in host-phage co-evolution.

Acknowledgements

We thank H. Garcia Martin, M.J. Blow, A. Visel and G. Tyson for helpful discussions. This work was performed under the auspices of the US Department of Energy's Office of Science, Biological and Environmental Research Program, and by the University of California, Lawrence Berkeley National Laboratory under contract No. DE-AC02-05CH11231, Lawrence Livermore National Laboratory under Contract No. DE-AC52-07NA27344, and Los Alamos National Laboratory under contract No. DE-AC02-06NA25396.

References

1. Breitbart, M. & Rohwer, F. Here a virus, there a virus, everywhere the same virus? Trends Microbiol 13, 278-84 (2005).
2. Wommack, K. E. & Colwell, R. R. Virioplankton: viruses in aquatic ecosystems. Microbiol Mol Biol Rev 64, 69-114 (2000).
3. Sturino, J. M. & Klaenhammer, T. R. Engineered bacteriophage-defence systems in bioprocessing. Nat Rev Microbiol 4, 395-404 (2006).
4. Ishino, Y., Shinagawa, H., Makino, K., Amemura, M. & Nakata, A. Nucleotide sequence of the iap gene, responsible for alkaline phosphatase isozyme conversion in *Escherichia coli*, and identification of the gene product. J Bacteriol 169, 5429-33 (1987).
5. Nakata, A., Amemura, M. & Makino, K. Unusual nucleotide arrangement with repeated sequences in the *Escherichia coli* K-12 chromosome. J Bacteriol 171, 3553-6 (1989).
6. Hermans, P. W. et al. Insertion element IS987 from *Mycobacterium bovis* BCG is located in a hot-spot integration region for insertion elements in *Mycobacterium tuberculosis* complex strains. Infect Immun 59, 2695-705 (1991).
7. Mojica, F. J., Ferrer, C., Juez, G. & Rodriguez-Valera, F. 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 (1995).
8. Bult, C. J. et al. Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. Science 273, 1058-73 (1996).
9. Nelson, K. E. et al. Evidence for lateral gene transfer between Archaea and bacteria from genome sequence of *Thermotoga maritima.* Nature 399, 323-9 (1999).
10. Mojica, F. J., Diez-Villasenor, C., Soria, E. & Juez, G. Biological significance of a family of regularly spaced repeats in the genomes of Archaea, Bacteria and mitochondria. Mol Microbiol 36, 244-6 (2000).
11. Kunin, V., Sorek, R. & Hugenholtz, P. Evolutionary conservation of sequence and secondary structures in CRISPR repeats. Genome Biol 8, R61 (2007).
12. Grissa, I., Vergnaud, G. & Pourcel, C. The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats. BMC Bioinformatics 8, 172 (2007).
13. Jansen, R., van Embden, J. D., Gaastra, W. & Schouls, L. M. Identification of a novel family of sequence repeats among prokaryotes. Omics 6, 23-33 (2002).
14. Jansen, R., Embden, J. D., Gaastra, W. & Schouls, L. M. Identification of genes that are associated with DNA repeats in prokaryotes. Mol Microbiol 43, 1565-75 (2002).
15. Makarova, K. S., Aravind, L., Grishin, N. V., Rogozin, I. B. & Koonin, E. V. A DNA repair system specific for thermophilic Archaea and bacteria predicted by genomic context analysis. Nucleic Acids Res 30, 482-96 (2002).
16. Makarova, K. S., Grishin, N. V., Shabalina, S. A., Wolf, Y. I. & Koonin, E. V. A putative RNA-interference-based immune system in prokaryotes: computational

analysis of the predicted enzymatic machinery, functional analogies with eukaryotic RNAi, and hypothetical mechanisms of action. Biol Direct 1, 7 (2006).
17. Haft, D. H., Selengut, J., Mongodin, E. F. & Nelson, K. E. A guild of 45 CRISPR-associated (Cas) protein families and multiple CRISPR/Cas subtypes exist in prokaryotic genomes. PLoS Comput Biol 1, e60 (2005).
18. Mojica, F. J., Diez-Villasenor, C., Garcia-Martinez, J. & Soria, E. Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. J Mol Evol 60, 174-82 (2005).
19. Pourcel, C., Salvignol, G. & Vergnaud, G. CRISPR elements in *Yersinia pestis* acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies. Microbiology 151, 653-63 (2005).
20. Bolotin, A., Quinquis, B., Sorokin, A. & Ehrlich, S. D. Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. Microbiology 151, 2551-61 (2005).
21. Barrangou, R. et al. CRISPR provides acquired resistance against viruses in prokaryotes. Science 315, 1709-12 (2007).
22. Godde, J. S. & Bickerton, A. The repetitive DNA elements called CRISPRs and their associated genes: evidence of horizontal transfer among prokaryotes. J Mol Evol 62, 718-29 (2006).
23. Tang, T. H. et al. Identification of 86 candidates for small non-messenger RNAs from the archaeon *Archaeoglobus fulgidus*. Proc Natl Acad Sci U S A 99, 7536-41 (2002).
24. Tang, T. H. et al. Identification of novel non-coding RNAs as potential antisense regulators in the archaeon *Sulfolobus solfataricus*. Mol Microbiol 55, 469-81 (2005).
25. Lillestøl, R. K., Redder, P., Garrett, R. A. & Brügger, K. A putative viral defence mechanism in archaeal cells. Archaea 2, 59-72 (2006).
26. Edwards, R. A. & Rohwer, F. Viral metagenomics. Nat Rev Microbiol 3, 504-10 (2005).
27. Klenk, H. P. et al. The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon *Archaeoglobus fulgidus*. Nature 390, 364-70 (1997).
28. Smith, D. R. et al. Complete genome sequence of *Methanobacterium thermoautotrophicum* deltaH: functional analysis and comparative genomics. J Bacteriol 179, 7135-55 (1997).
29. Ebihara, A. et al. Crystal structure of hypothetical protein TTHB192 from *Thermus thermophilus* HB8 reveals a new protein family with an RNA recognition motif-like domain. Protein Sci 15, 1494-9 (2006).
30. Hannon, G. J. RNA interference. Nature 418, 244-51 (2002).
31. Sontheimer, E. J. Assembly and function of RNA silencing complexes. Nat Rev Mol Cell Biol 6, 127-38 (2005).
32. Tyson, G. W. & Banfield, J. F. Rapidly evolving CRISPRs implicated in acquired resistance of microorganisms to viruses. Environ Microbiol (2007) [Epub ahead of print].
33. Greve, B., Jensen, S., Brügger, K., Zillig, W. & Garrett, R. A. Genomic comparison of archaeal conjugative plasmids from *Sulfolobus*. Archaea 1, 231-9 (2004).

34. Sebaihia, M. et al. The multidrug-resistant human pathogen *Clostridium difficile* has a highly mobile, mosaic genome. Nat Genet 38, 779-86 (2006).
35. Groenen, P. M., Bunschoten, A. E., van Soolingen, D. & van Embden, J. D. Nature of DNA polymorphism in the direct repeat cluster of *Mycobacterium tuberculosis*; application for strain differentiation by a novel typing method. Mol Microbiol 10, 1057-65 (1993).
36. Kamerbeek, J. et al. Simultaneous detection and strain differentiation of *Mycobacterium tuberculosis* for diagnosis and epidemiology. J Clin Microbiol 35, 907-14 (1997).
37. Crawford, J. T. Genotyping in contact investigations: a CDC perspective. Int J Tuberc Lung Dis 7, S453-7 (2003).
38. Brudey, K. et al. *Mycobacterium tuberculosis* complex genetic diversity: mining the fourth international spoligotyping database (SpolDB4) for classification, population genetics and epidemiology. BMC Microbiol 6, 23 (2006).
39. Mokrousov, I., Limeschenko, E., Vyazovaya, A. & Narvskaya, O. *Corynebacterium diphtheriae* spoligotyping based on combined use of two CRISPR loci. Biotechnol J 2, 901-6 (2007).
40. Schouls, L. M. et al. 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 (2003).
41. DeBoy, R. T., Mongodin, E. F., Emerson, J. B. & Nelson, K. E. Chromosome evolution in the Thermotogales: large-scale inversions and strain diversification of CRISPR sequences. J Bacteriol 188, 2364-74 (2006).
42. Suttle, C. A. Viruses in the sea. Nature 437, 356-61 (2005).
43. Peng, X. et al. Genus-specific protein binding to the large clusters of DNA repeats (short regularly spaced repeats) present in *Sulfolobus* genomes. J Bacteriol 185, 2410-7 (2003).
44. Viswanathan, P., Murphy, K., Julien, B., Garza, A. G. & Kroos, L. Regulation of dev, an operon that includes genes essential for *Myxococcus xanthus* development and CRISPR-associated genes and repeats. J Bacteriol 189, 3738-50 (2007).
45. Edgar, R. C. PILER-CR: fast and accurate identification of CRISPR repeats. BMC Bioinformatics 8, 18 (2007).
46. Bland, C. et al. CRISPR Recognition Tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. BMC Bioinformatics 8, 209 (2007).
47. Grissa, I., Vergnaud, G. & Pourcel, C. CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. Nucleic Acids Res 35, W52-7 (2007).
48. Durand, P., Mahé, F., Valin, A. S. & Nicolas, J. Browsing repeats in genomes: Pygram and an application to non-coding region analysis. BMC Bioinformatics 7, 477 (2006).
49. Deveau, H. et al. Phage response to CRISPR-encoded resistance in Streptococcus thermophilus. J Bacteriol., *in press*
50. Horvath, P. et al. Diversity, activity and evolution of CRISPR loci in Streptococcus thermophilus. J Bacteriol., *in press*

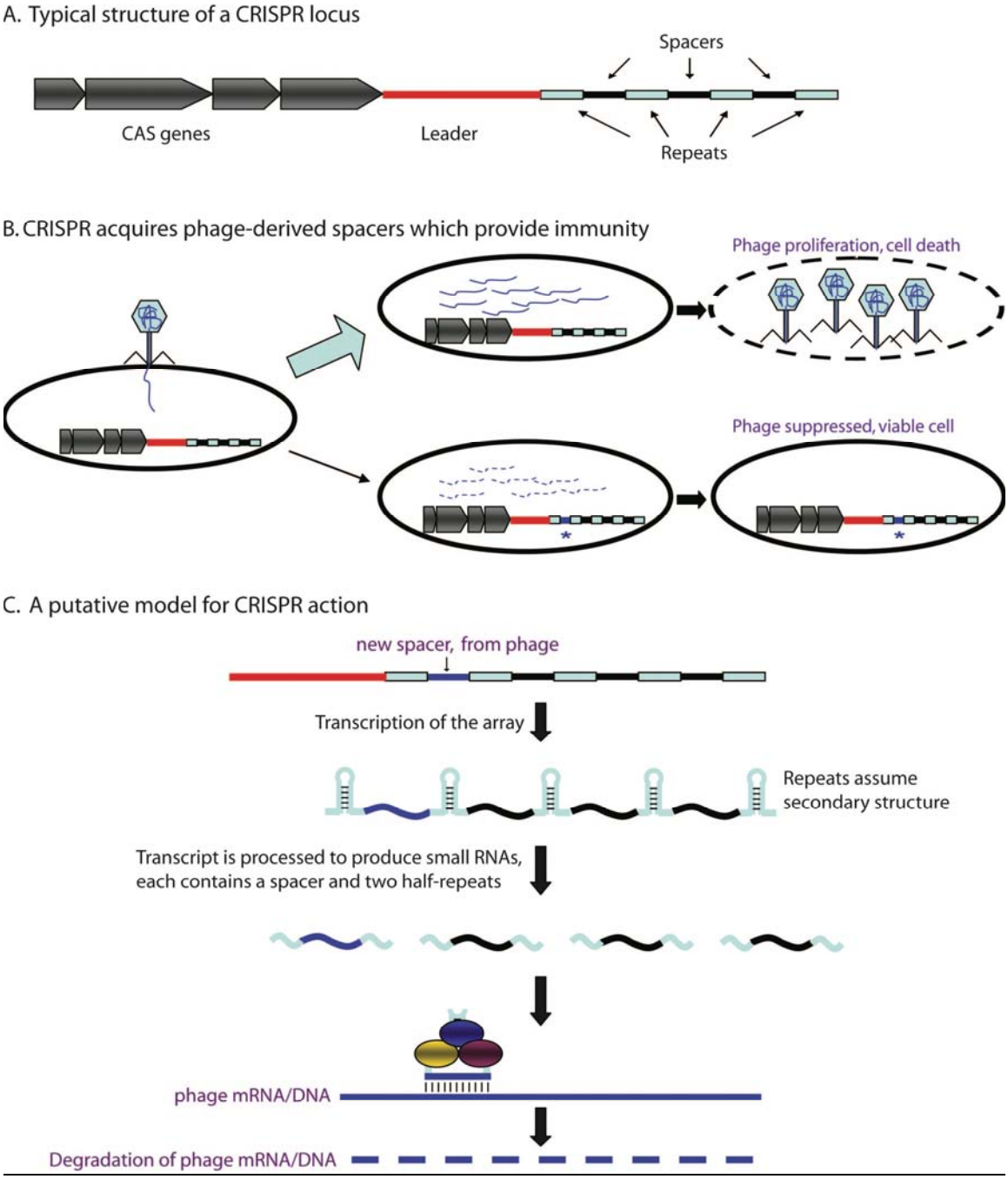

Figure 1: CRISPR structure and function. (A) Typical structure of a CRISPR locus. (B) CRISPR acquires phage-derived spacers that provide immunity. Following an attack by phage, phage nucleic acids proliferate in the cell and new particles are produced leading to death of the majority of sensitive bacteria. A small number of bacteria acquire phage derived spacers (blue spacer, marked by asterix) leading to survival, presumably via CRISPR-mediated degradation of phage mRNA or DNA. (C) Putative (simplified) model for CRISPR action. The repeat-spacer array is transcribed into a long RNA, and the repeats assume a secondary structure. Cas proteins recognize the sequence/structure of the repeats and process the RNA to produce small RNAs (sRNAs), each containing a spacer and two half repeats. The sRNAs, complexed with additional Cas proteins, base pair with phage nucleic acids leading to its degradation, putatively mediated by one or more of the Cas proteins.

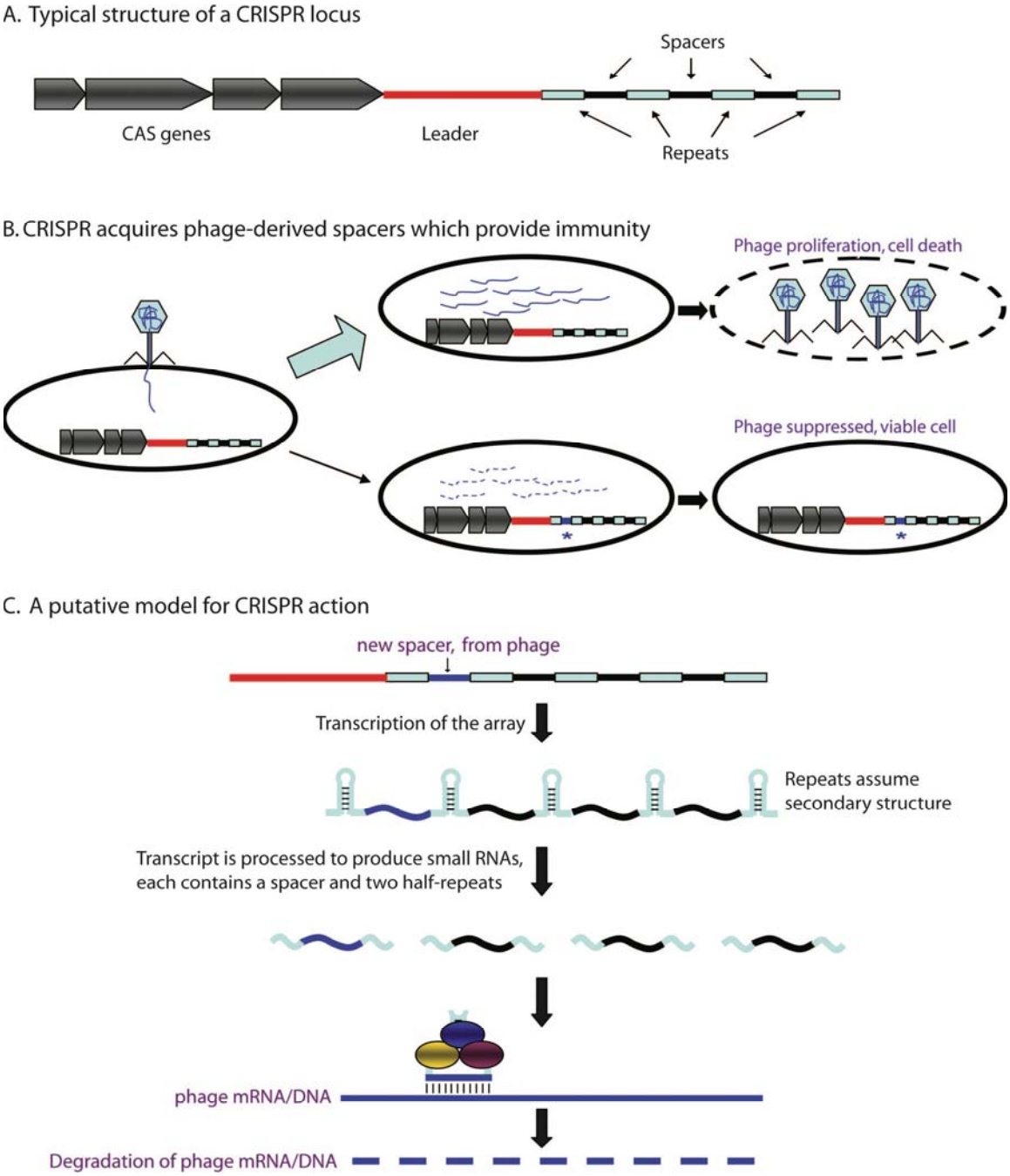

Figure 2: Applications of CRISPR. (A) Spoligotyping: Labeled primers (a,b) are designed from the repeat region to amplify the CRISPR array. Probes matching known spacers are printed on a membrane, and hybridization with the amplified products for each isolate occurs. Black boxes represent existence of a spacer; white boxes represent spacer absence. In the figure, isolates 1 and 3 belong to the same strain, as well as isolates 2, 6 and 7. Adapted from ref. [36] (B) Engineering of phage resistance into sensitive industrial bacteria. Sequences from known phages are inserted as spacers into a CRISPR array and the CRISPR system is then transformed into bacteria. (C) Silencing of endogenous genes as an alternative to knockout methods. Fragments from a chromosome-encoded gene (green) are engineered into a CRISPR array as spacers. If the CRISPR system indeed functions via silencing of RNA as suggested[16], this might lead to silencing of the endogenous gene.

Box 1: Tools for CRISPR detection and analysis

A growing interest in CRISPRs has led to the development of different computer software and web resources for analysis of CRISPR systems (see Table). These tools include software for CRISPR detection such as Piler-CR[45], CRISPR recognition tool[46] and CRISPRFinder[47]; online repositories of pre-analyzed CRISPRs such as CRISPRdb[12]; and tools for browsing CRISPRs in microbial genomes such as Pygram[48]. The Institute for Genomic Research (TIGR) also provides a web-page that displays the occurrence profile of all Cas proteins[17] for each available microbial genome. Among these tools CRISPRdb is especially notable as, apart from containing an automatically updated database of CRISPR arrays in published genomes (currently ~700 arrays in 232 genomes), it also provides various analysis tools allowing the extraction and alignment of specific repeats and spacers as well as the flanking leader sequences. Despite this recent proliferation of tools for CRISPR analysis there is still a need for tools that would allow the combined analysis of CAS and CRISPRs, because most tools either focus on the repeat arrays or on the related CAS genes. Reports showing the association between specific repeat types and specific CAS subsystems[11] highlight the need for such a combined web resource.

Resource	Web page	Description	Ref.
Piler-CR	http://www.drive5.com/pilercr/	A software tool for detection of CRISPRs in microbial genomic sequences; based on local alignments in the genome represented by mathematical graphs [a]	45
CRISPR recognition tool (CRT)	http://www.room220.com/crt/	A software tool for detection of CRISPRs in microbial genomic sequences; based on detection of exact k-mer matches separated by similar distances [a]	46
CRISPRFinder	http://crispr.u-psud.fr/crispr/	A software tool for detection of CRISPRs in microbial genomic sequences; based on enhanced suffix arrays [a]	47
CRISPRdb	http://crispr.u-psud.fr/crispr/	Automatically updated database of CRISPR arrays in published microbial genomes; also contains CRISPR analysis tools allowing alignment of repeats and spacers as well as BLASTing them against the public databases	12
Pygram	http://www.irisa.fr/symbiose/projets/Modulome/article.php3?id_article=18	Visualization application providing graphical browser for studying repeats	48
TIGR Comprehensive Microbial Resource (CMR)	http://rice.tigr.org/tigr-scripts/CMR2/genome_property.spl?subproperty=CRISPR%20region!&select_count=1	Provides a clickable table depicting, for each sequenced genome, the presence/absence of the 45 Cas protein families as defined in [17].	17

[a] All CRISPR detection software apply post-processing filters to separate real CRISPR arrays from false predictions.