J. Bacteriol. doi:10.1128/JB.00580-17

# History of CRISPR-Cas from encounter with a mysterious repeated sequence to genome editing technology

**Yoshizumi Ishino,[1, 2,]* Mart Krupovic,[1] Patrick Forterre[1, 3]**

[1]*Unité de Biologie Moléculaire du Gène chez les Extrêmophiles, Département de Microbiologie, Institut Pasteur, F-75015, Paris, France,* [2]*Department of Bioscience and Biotechnology, Faculty of Agriculture, Kyushu University, Fukuoka 812-8581, Japan.* [3]*Institute of Integrative Cellular Biology, Université Paris Sud, 91405 Orsay, Cedex France*

Running title: Discovery and development of CRISPR-Cas research

* Correspondence to

Prof. Yoshizumi Ishino

Department of Bioscience and Biotechnology,

Faculty of Agriculture, Kyushu University,

Fukuoka 812-8581, Japan

ishino@agr.kyushu-u.ac.jp

## ABSTRACT

CRISPR-Cas systems are well known acquired immunity systems that are widespread in Archaea and Bacteria. The RNA-guided nucleases from CRISPR-Cas systems are currently regarded as the most reliable tools for genome editing and engineering. The first hint of their existence came in 1987, when an unusual repetitive DNA sequence, which subsequently defined as a cluster of regularly interspersed short palindromic repeats (CRISPR), was discovered in the *Escherichia coli* genome during the analysis of genes involved in phosphate metabolism. Similar sequence patterns were then reported in a range of other bacteria as well as in halophilic archaea, suggesting an important role for such evolutionarily conserved clusters of repeated sequences. A critical step towards functional characterization of the CRISPR-Cas systems was the recognition of a link between CRISPRs and the associated Cas proteins, which were initially hypothesized to be involved in DNA repair in hyperthermophilic archaea. Comparative genomics, structural biology and advanced biochemistry could then work hand in hand, culminating not only in the explosion of genome editing tools based on CRISPR-Cas9 and other class II CRISPR-Cas systems, but also providing insights into the origin and evolution of this system from mobile genetic elements denoted casposons. To celebrate the 30th anniversary of the discovery of CRISPR, this minireview briefly discusses the fascinating history of CRISPR-Cas systems, from the original observation of an enigmatic sequence in *E. coli* to genome editing in humans.

**KEYWORDS** *Repeated sequence, RAMP, Casposon, Archaea, Genome editing*

## INTRODUCTION

CRISPR-Cas systems are currently in the spotlight of active research in biology. The first clustered regularly interspaced short palindromic repeats (CRISPR) were detected 30 years ago by one of the authors of this review (YI) in *Escherichia coli* in the course of the analysis of the gene responsible for isozyme conversion of alkaline phosphatase (1). The structural features of CRISPR are shown in Figure 1. At the time, it was hardly possible to predict the biological function of these unusual repeated sequences due to the lack of sufficient DNA sequence data, especially for mobile genetic elements. The actual function of this unique sequence remained enigmatic right up until the mid-2000s. In 1993, CRISPRs were for the first time observed in Archaea, specifically in *Haloferax mediterranei* (2), and subsequently detected in an increasing number of bacterial and archaeal genomes, since life science moved into genomic era. Conservation of these sequences in two of the three domains of life was critical for appreciating their importance. In the early 2000s, the discovery of sequence similarity between the spacer regions of CRISPR and sequences of bacteriophages, archaeal viruses and plasmids finally shed light on the function of CRISPR as an immune system. This dramatic discovery by Mojica and others was grossly underappreciated at that time, and was published in 2005 by three research groups independently (3-5). In parallel, several genes previously proposed to encode for DNA repair proteins specific for hyperthermophilic archaea (6) were identified to be strictly associated with CRISPR, and designated as *cas* (CRISPR-associated genes) (7). Comparative genomic analyses thus suggested that CRISPR and Cas proteins (the *cas* gene products) actually work together and constitute an acquired immunity system to protect the prokaryotic cells against invading viruses and plasmids, analogous to the eukaryotic RNA interference (RNAi) system (8).

This minireview focuses on the contribution of early fundamental microbiological research to the discovery of the CRISPR-Cas system and to our understanding of its function and mode of action (for other recent reviews on the history of the research on CRISPR-Cas system see refs 9-14). We also emphasize recent discoveries that shed light on the origins of the system and suggest that more tools remain to be discovered in the microbial world that could still improve our genome editing capacity.

## A PUZZLING SEQUENCE FROM BACTERIA CHALLENGES THE EARLY SEQUENCING METHODOLOGY

In the mid-80s, when studying isozyme conversion of alkaline phosphatase (AP), one of us (YI), in an attempt to identify the protein responsible for the isozyme conversion of AP in the periplasm of the *E. coli* K12 cells, sequenced a 1.7 kbp *E. coli* DNA fragment spanning the region containing the *iap* gene (designated from isozyme of alkaline phosphatase) (1). The isozyme of AP was previously detected by biochemical and genetic analyses (15). At that time, for conventional M13 dideoxy sequencing, single-stranded template DNA had to be produced by cloning the target DNA into an M13 vector, whereas the dideoxy chain-termination reaction was performed by Klenow fragment of *E. coli* Pol I. The reaction products were labeled by incorporation of $[\alpha^{32}P]$dATP, and the sequence ladder images were obtained by autoradiography. For sequencing, the cloned DNA fragment had to be subcloned into M13 mp18 and 19 vectors (for the coding and noncoding strands) after digestion into short fragments. During the sequencing of the DNA fragment containing *iap*, one of the authors realized that the same sequence appeared many times in different clones. Furthermore, it was difficult to read the repeated sequences precisely, using the Klenow fragment at 37°C, because of non-specific termination of the dideoxynucleotide incorporation reactions for the template DNA, due to secondary structure formation by the palindromic sequence. This is why it took several months to read the sequence of the CRISPR region precisely in 1987 (1). A peculiar repeated sequence was detected downstream of the translation termination codon for the *iap* gene (Fig. 2). It is remarkable that the exact same region can be sequenced in just one day using current technology, by amplification of the target region by PCR directly from the genome, followed by a fluorescent-labeling and cycle-sequencing at 72°C (Fig. 3). The feature of the repetitive sequence was so mysterious and unexpected that it was mentioned in the Discussion section, even though its function was not understood (1). Notably, the same sequence containing a dyad symmetry of 14 bp was repeated five times with a variable 32-nucleotide sequence interspersed between the repeats (Fig. 2). Well-conserved nucleotide sequences containing a dyad symmetry, named REP (Repetitive extragenic palindromic) sequences (16), had been previously found in *E. coli* and *Salmonella typhimurium* and suggested to stabilize mRNA (17). However, no similarities were found between the REP and the repeated sequences detected downstream of the *iap* gene. In fact, this sequence was, at the time, unique in sequence databases. As it later turned out, this was the first encounter with a CRISPR sequence. Soon after, similar sequences were detected by southern blot hybridization analysis in other *E. coli* strains (C600 and Ymel) and in two other members of the *Enterobacteriaceae*, *Shigella dysenteriae* and *Salmonella typhimurium* (phylum *Proteobacteria*) (18). Subsequently, similar repeated sequences were also found in members of the phylum Actinobacteria, such as *Mycobactrium tuberculosis* (19), but not in the closely related strain *M. leprae,* prompting the use of these highly polymorphic repeated sequences for strain typing (20).

## DISCOVERY OF CRISPR IN ARCHAEA

A major advance was made when similar repeated sequences were identified by Mojica and co-workers in the archaeon *Haloferax mediterranei* during the research on regulatory mechanisms allowing extremely halophilic archaea to adapt to high salt environments (2). Transcription of the genomic regions containing the repeated sequences was demonstrated by Northern blot analysis (2), but compelling evidence for the processing of the transcripts into several different RNA products was shown only more recently (12). The authors first suggested that these repeated sequences could be involved in the regulation of gene expression, possibly facilitating the conversion of the double-stranded DNA from B to Z-form for the specific binding of a regulator protein. It was indeed often suggested at that time that the high GC content of halophilic genomes could facilitate such B-to-Z transition for regulatory purposes at the high intracellular salt concentration characteristic of haloarchaea. However, such explanation could not be valid for bacteria. Soon after, the same authors found a similar repeated sequence in *Haloferax volcanii,* and hypothesized that these repeated sequences could be involved in replicon partitioning (21).

In the meantime, invention of the automated sequencing machines and development of efficient procedures for DNA sequencing during the 90s provided scientists for the first time with access to complete genome sequences. Starting with *Haemophilus influenzae* (22), followed by *Methanocaldococcus jannaschii* (23) and *Sacchamyces cerevisiae* (24), all three domains of life entered into the genomics era. Then, the unusual repeated sequences interspersed with non-conserved sequences, first detected in *E. coli* and *H. mediterranei*, were identified in an increasing number of bacterial and archaeal genomes, and were described using different names by different authors, such as SRSRs, (Short Regularly Spaced Repeats (2), SPIDR (spacers interspersed direct repeats) or LCTR (large cluster of tandem repeats) (25). In the hyperthermophilic archaea *Pyrococcus abyssi* and *P. horikoshi* two sets of "LCTR" sequences were located symmetrically on each side of the replication origin, again suggesting a possible role in chromosome partitioning. However, they were more numerous and scrambled in the genome of *P. furiosus*, casting doubt on this interpretation (26).

Mojica *et al.* were the first to realize that all these bacterial and archaeal sequences were functionally related (27). The term CRISPR, for clustered regularly interspaced short palindromic repeats, was proposed by Jansen *et al* in 2002 (7) and became generally accepted by the community working on these sequences, which precluded further confusion caused by many different names for the related repeat sequences. Comparative genomics studies illuminated the common characteristics of the CRISPR, namely that i) they are located in intergenic regions; ii) contain multiple short direct repeats with very little sequence variation; iii) the repeats are interspersed with non-conserved sequences; iv) a common leader sequence of several hundred base pairs is located on one side of the repeat cluster.

The fact that these mysterious sequences were conserved in two different domains of life pointed to a more general role of these sequences. CRISPR sequences were found in nearly all archaeal genomes and in about half of bacterial genomes, rendering them the most widely distributed family of repeated sequences in prokaryotes. As of today, CRISPR sequences have not been found in any eukaryotic genome.

## IDENTIFICATION OF THE *CAS* GENES

The accumulation of genomic sequences in the beginning of this century enabled scientists to compare the genomic context of CRISPR regions in many organisms, which led to the discovery of four conserved genes regularly present adjacent to the CRISPR regions. The genes were designated as CRISPR-associated genes 1 through 4 (*cas1*-*cas4*) (7). No similarity to functional domains of any known protein was identified for the Cas1 and Cas2. By contrast, Cas3 contained the seven motifs characteristic of the superfamily 2 helicases, whereas Cas4 was found to be related to RecB exonucleases, which work as part of the RecBCD complex for the terminal resection of the double-strand breaks to start homologous recombination. Therefore, Cas3 and Cas4 were predicted to be involved in DNA metabolism, including DNA repair and recombination, transcriptional regulation or chromosome segregation. Due to their association with CRISPR, it was suggested that Cas proteins are involved in the genesis of the CRISPR loci (7).

At about the same time, Kira Makarova, Eugene Koonin and colleagues independently and systematically analyzed the conserved gene contexts in all prokaryotic genomes available at the time and found several clusters of genes corresponding to *cas* genes (encoding putative DNA polymerase, helicase and RecB-like nuclease) in the genomes of hyperthermophilic archaea and in the two hyperthermophilic bacteria with available genome sequences, *Aquifex* and *Thermotog*a (8). These conserved genes were not found at that time in mesophilic and moderate thermophilic archaea and bacteria. Based on this observation, it was predicted that these proteins could be part of a "mysterious" uncharacterized DNA repair system specific to thermophilic organisms.

## THE DISCOVERY OF CRISPR FUNCTION

In the beginning of the genomic era, most of the archaeal genome sequences were those of thermophilic and hyperthermophilic organisms. Furthermore, thermophilic archaea, in addition to the hyperthermophilic bacteria, such as *A. aeolicus* and *T. maritima,* have more and larger CRISPRs than mesophilic organisms (7). These observations first suggested that the function of CRISPR may be related to adaptation of organisms to high temperatures. However, with more and more sequences becoming available, it turned out that this correlation was not robust and that many mesophilic organisms also contained CRISPR sequences. The *Eureka!* moment came when Francisco Mojica in Alicante and Christine Pourcel in Orsay noticed independently that the spacer regions between the repeat sequences are homologous to sequences of bacteriophages, prophages and plasmids (3, 4). Importantly, based on the literature review, they pointed out that the phages and plasmids do not infect host strains harboring the homologous spacer sequences in the CRISPR. From these observations, they independently proposed that CRISPR sequences function in the framework of a biological defense system similar to the eukaryotic RNAi system to protect the cells from the entry of these foreign mobile genetic elements. The two groups also suggested that the CRISPRs can somehow trigger the capture of pieces of foreign invading DNA to constitute a memory of past genetic aggressions (3, 4). In a third influential paper of the same year, Bolotin and colleagues confirmed these observations, further noticing a correlation between the number of spacers of phage origin and the degree of resistance to phage infection and suggested that CRISPR could be used to produce antisense RNA (5) (for a brief historical account, see Morange, 2015) (9).

As mentioned above, these seminal publications were grossly underappreciated at the time and published in specialized journals (12). Interestingly, Morange suggested that lack of adequate recognition of the 2005 papers at that time and in subsequent years in some publications and reviews might be due to both cultural and sociological reasons based partly on the predominance of experimental molecular biologists over microbiologists and evolutionists (9). In two of the three 2005 papers, the authors acknowledged the previous discovery of the *cas* genes, suggesting that proteins encoded by these genes should be involved in the functioning of this new putative prokaryotic immune system (4, 5).

The predicted role of Cas proteins as effectors of prokaryotic immunity was emphasized a year after in an exhaustive analytical paper published by the Koonin group (8). Building on their previous work, Makarova *et al.* performed a detailed analysis of the Cas protein sequences and attempted to predict their functions in a mechanism similar to the eukaryotic RNAi system (8). Notably, in many cases, these, often non-trivial, functional predictions, as in the case of Cas1 integrase, were fully confirmed experimentally several years later and continue to guide experimental research on the CRISPR-Cas systems. Importantly, they pinpointed that the CRISPR-Cas system, with its memory component, rather resembles the adaptive immune system of vertebrates, with the crucial difference that the animal immune system is not inheritable. Considering the diversity of the CRISPR-Cas systems, their erratic distribution suggesting high mobility, and their ubiquity in Archaea, Makarova *et al* suggested that the CRISPR-Cas system emerged in an ancient ancestor of archaea and spread to bacteria horizontally. They concluded on a practical note, suggesting that CRISPR-Cas systems could be exploited to silence genes in organisms encoding Cas proteins (8).

The function of the CRISPR-Cas system as a prokaryotic acquired immune system was finally experimentally proven in 2007, using the lactic acid bacterium, *S. thermophilus* in 2007 (28). Insertion of the phage sequence into the spacer region of the CRISPR of *S. thermophilus* made this strain resistant to the corresponding phage. On the other hand, this bacterial resistance to the phage infection disappeared when the corresponding protospacer sequence was deleted from the phage genome. In addition, it was experimentally demonstrated that CRISPR-Cas restricts transformation of plasmids carrying sequences matching the CRISPR spacers (29). Then, van der Oost's group reconstituted the immunity system using *E. coli* CRISPR, which was originally discovered in 1987. They demonstrated that the processed RNA molecules from the transcription of the CRISPR region function by cooperation with the Cas proteins produced from the genes located next to the CRISPR (30). Around the same time, metagenomic analysis of archaea by Banfield's group indicated dynamic changing of sequences at CRISPR loci on a time scale of months, and new spacer sequences corresponding to phages in the same communities appeared (31). Subsequently, the CRISPR-Cas system of *S. thermophilus* expressed in *E. coli* showed heterologous protection against plasmid transformation and phage infection by the reconstituted CRISPR-Cas9 system of *S. thermophilus* (32). This work also showed that *cas9* is, in that case, the sole *cas* gene necessary for CRISPR-encoded interference. Soon after, it has been proven that the purified Cas9-CRISPR RNA (crRNA) complex is capable of cleaving the target DNA *in vitro* (33, 34). The CRISPR-Cas system of *S. pyogenes* was then applied to perform genome editing in human nerve and mouse kidney cells (35, 36). Thus, CRISPR-Cas came to be widely known as the prokaryotic acquired immunity system (37, 38). The various steps underlying the functioning of this system are schematically shown in Fig. 4.

Numerous and highly diverse Cas proteins are involved in different stages of CRISPR immunity; they exhibit a variety of predicted nucleic acid-manipulating activities such as nucleases, helicases and polymerases, which have been described in detail in several excellent recent reviews (39-42). In a nutshell, Cas1 and Cas2 are conserved throughout most known types of CRISPR–Cas systems and form a complex that represents the adaptation module required for the insertion of new spacers into the CRISPR arrays. During the expression stage, the CRISPR locus is transcribed and the pre-crRNA transcript is processed by the type-specific Cas endonucleases into the mature crRNAs. During the interference stage, the crRNAs are bound by the effector Cas enonucleases and the corresponding complexes are recruited to and cleave the target DNA or RNA in a sequence-dependent manner (Fig. 4). Notably, unlike the adaptation module, Cas enzymes involved in the expression and interference stages vary from one CRISPR-Cas type to the other and the same enzymes may participate in both stages of immunity.

## DIVERSITY AND CLASSIFICATION OF CRISPR-CAS

It is striking that closely related strains can vary considerably in their CRISPR content and distribution. For example, in *Mycobacterium* genus, CRISPR exists in *M. tuberculosis*, but not in *M. leprae*. On the other hand, phylogenetically distant *E. coli* and *M. avium* as well as *Methanothermobacter thermautotrophicus* and *Archaeoglobus fulgidus* carry nearly identical CRISPR repeat sequences (7). The number of CRISPR arrays in one genome varies from 1 to 18, and the number of repeat units in one CRISPR array varies from 2 to 374 (43). Based on the CRISPR database (http://crispr.u-psud.fr/crispr/), as of May 2017, CRISPRs were identified in the whole genome sequences of 202 (87%) out of 232 analyzed archaeal species and 3059 (45%) of 6782 bacterial species. Interestingly, a survey of 1,724 draft genomes suggested that CRISPR-Cas systems are much less prevalent in environmental microbial communities (10.4% in bacteria and 10.1% in archaea). This large difference between the prevalence estimated from complete genomes of cultivated microbes compared to that of the uncultivated ones was attributed to the lack of CRISPR-Cas systems across major bacterial lineages that have no cultivated representatives (44).

As shown in Fig. 5, the latest classification of CRISPR–Cas systems includes two classes, class 1 and 2, based on the encoded effector proteins (45). Class 1 CRISPR–Cas systems work with multisubunit effector complexes consisting of 4–7 Cas proteins present in an uneven stoichiometry. This system is widespread in Bacteria and Archaea, including in all hyperthermophiles, comprising ~90% of all identified CRISPR–*cas* loci. The remaining ~10% belong to class 2, which use a single multidomain effector protein and are found almost exclusively in Bacteria (46).

Each class currently includes three types, namely, types I, III, and IV in class 1, and types II, V, and VI in class 2. Types I, II, and III are readily distinguishable by virtue of the presence of unique signature proteins: Cas3 for type I, Cas9 for type II and Cas10 for type III. The multimeric effector complexes of type I and type III systems, known as the CRISPR-associated complex for antiviral defense (Cascade) and the Csm/Cmr complexes, respectively, are architecturally similar and evolutionarily related (47-52). Unlike all other known CRISPR-Cas systems, the functionally uncharacterized Type IV systems do not contain the adaptation module consisting of nucleases Cas1 and Cas2 (47, 53). Notably, the effector modules of subtype III-B systems are known to utilize spacers produced by Type I systems, testifying to the modularity of the CRISPR-Cas systems (54). Although many of the genomes encoding Type IV systems do not carry identifiable CRISPR loci, it is not excluded that Type IV systems, similar to subtype III-B systems, use crRNAs from different CRISPR arrays once these become available (53).

Finally, each type is classified into multiple subtypes (I-A~F, and U; III-A~D in class 1; II-A~C; V-A~E and U; VI-A~C in class 2) based on additional signature genes and characteristic gene arrangements (45, 51). The figure 6B shows distribution of CRISPR-Cas systems in Archaea and Bacteria.

## CLASS 2 SYSTEMS ARE SUITABLE FOR GENOME EDITING TECHNOLOGY

The simple architecture of the effector complexes has made class 2 CRISPR–Cas systems an attractive choice for developing a new generation of genome-editing technologies (Fig. 6). Several distinct class 2 effectors have been reported, including Cas9 in type II, Cas12a (formerly Cpf1), Cas12b (C2c1) in Type V, and Cas13a (C2c2) and Cas13b (C2c3) in Type VI (45, 51). The most common and best studied multidomain effector protein is Cas9, a crRNA-dependent endonuclease, consisting of two unrelated nuclease domains, RuvC and HNH, which are responsible for cleavage of the displaced (non-target) and target DNA strands, respectively, in the crRNA–target DNA complex. Type II CRISPR–*cas* loci also encode a *trans*-activating crRNA (tracrRNA) which might have evolved from the corresponding CRISPR. The tracrRNA molecule is also essential for pre-crRNA processing and target recognition in the type II systems. The molecular mechanism of the target DNA cleavage by Cas9-crRNA complex, schematically shown in Fig. 7, has been elucidated at the atomic level by the crystal structure analysis of the DNA-Cas9-crRNA complex (55).

A gene originally denoted as *cpf1* is present in several bacterial and archaeal genomes, where it is adjacent to *cas1*, *cas2* and CRISPR array (45). Cas12a (Cpf1), the prototype of type V effectors, contains two RuvC-like nuclease domains, but lacks the HNH domain. However, recent structural analysis of Cas12a-crRNA-target DNA complex revealed a second nuclease domain with a unique fold that is functionally analogous to the HNH domain of Cas9 (56). Cas12a is a single-RNA-guided nuclease that does not require a tracrRNA, which is indispensable for Cas9 activity (57). The protein also differs from Cas9 in its cleavage pattern and in its PAM recognition, which determines the target strands.

The discovery of two distantly related class 2 effector proteins, Cas9 and Cas12a, suggested that other distinct variants of such systems could exist. Indeed, more recently, Cas12b (type V), Cas13a and Cas13b (type VI), which are distinct from Cas9 or Cas12a, have been discovered through directed bioinformatics search for class II effectors, and their activities were confirmed (58). Type V effectors, similar to Cas9, need a tracrRNA for the targeted activity. Most of the functionally characterized CRISPR-Cas systems, to date, have been reported to target DNA, and only the multi-component type III-A and III-B systems additionally target RNA (59). By contrast, type VI effectors, Cas13a and Cas13b, specifically target RNA, thereby mediating RNA interference. Unlike type II and type V effectors, Cas13a and Cas13b lack characteristic RuvC-like nuclease domains and instead contain a pair of HEPN (higher eukaryotes and prokaryotes nucleotide-binding) domains (60). The discovery of novel class 2 effectors will most likely provide new opportunities for the application of CRISPR systems to genome engineering technology (61).

## ORIGINS OF CRISPR-CAS

Analysis of clusters of poorly characterized, narrowly spread fast-evolving genes in archaeal genomes, denoted as ‘dark matter islands’ (62), revealed several islands encoding Cas1 proteins not associated with CRISPR loci (Cas1-solo) (63). Comprehensive interrogation of the dark matter islands revealed that *cas1*-solo genes are always located in vicinity of genes encoding family B DNA polymerases and several other conserved genes (64). Furthermore, these gene ensembles were found to be surrounded by long inverted repeats and further flanked by shorter direct repeats, which respectively resembled terminal inverted repeats (TIR) and target site duplications (TSD) characteristic of various transposable elements. However, none of the identified Cas1-solo-encoding genomic loci carried genes for known transposases or integrases. Thus, it was hypothesized that Cas1 is the principal enzyme responsible for the mobility of these novel genetic elements, which were accordingly named ‘Casposons’ (64). Casposons were found to be widespread in the genomes of methanogenic archaea as well as in thaumarchaea, but also present in different groups of bacteria. Strong evidence of recent casposon mobility was obtained by comparative genomic analysis of more than 60 strains of the archaeon *Methanosarcina mazei*, in which casposons are variably inserted in several distinct sites indicative of multiple, recent gains, and losses (65). Based on the gene content, taxonomic distribution and phylogeny of the Cas1 proteins, casposons are currently classified into 4 families (66).

Biochemical characterization of the casposon Cas1 (‘casposase’) encoded in the genome of a thermophilic archaeon *Aciduliprofundum boonei* has confirmed the predicted integrase activity (67, 68). Integration showed strong target site preference and resulted in the duplication of the target site regenerating the TSD observed in the *A. boonei* genome (68). The latter feature resembles the duplication of the leader sequence-proximal CRISPR unit upon integration of a protospacer catalyzed by the Cas1-Cas2 adaptation machinery of CRISPR-Cas (69, 70). Remarkably, the sequence features of the casposon target site are functionally similar to those required for directional insertion of new protospacers into CRISPR arrays. In both systems, the functional target site consists of two components: (i) a sequence which gets duplicated upon integration of the incoming DNA duplex (i.e. the TSD segment in the case of casposon and a CRISPR unit during protospacer integration) and (ii) the upstream region which further determines the exact location of the integration (i.e. the leader sequence located upstream of the CRISPR array and the TSD-proximal segment in *A. boonei* genome) (68).

Collectively, the comparative genomics and experimental results reinforced the mechanistic similarities and evolutionary connection between the casposons and the adaptation module of the prokaryotic adaptive immunity system, culminating in an evolutionary scenario for the origin of the CRISPR-Cas systems. It has been proposed that casposon insertion near a ‘solo-effector’ innate immunity locus, followed by the immobilization of the ancestral casposon via inactivation of the TIRs, gave rise to the adaptation and effector modules, respectively, whereas the CRISPR repeats and the leader sequence evolved directly from the preexisting casposon target site (71, 72). An outstanding question in the above scenario is the switch in substrate specificity of the ancestral casposase from integration of defined casposon TIRs to insertion of essentially random, short (compared to casposon length) protospacer sequences. It has been suggested that coupling between Cas1 and Cas2 has been critical for this evolutionary transition (72).

Remarkably, casposons are not the only mobile genetic elements that contributed to the origin and evolution of the CRISPR-Cas systems. It has been demonstrated that class 2 effector proteins of type II and type V have independently evolved from different groups of small transposons, which donated the corresponding RuvC-like nuclease domains (45, 58).

## APPLICATION OF CRISPR-CAS TOOLS TO BACTERIA AND ARCHAEA

Microbial engineering directly influences the development of the bioindustry. High-throughput genome editing tools are useful for breeding economically valuable strains. It is remarkable how quickly the practical application of the CRISPR-Cas system has been adapted to genome editing in eukaryotic cells. Such rapid success of this technology in eukaryotic cells was linked to the fact that eukaryotes employ the error-prone non-homologous end joining (NHEJ) to repair double-strand breaks introduced by the CRISPR-Cas in the target sequence. The use of the CRISPR-Cas technology was not as ‘revolutionary’ in bacteria, likely because other methods based on homologous recombination (HR) were already available for efficient manipulation of their genomes. Nevertheless, DNA Toolkits based on CRISPR-Cas technology for genome editing, gene silencing and genome-wide screening of essential genes in bacterial and archaeal genomes are gradually emerging and diversifying

(73-77). For instance, CRISPR-Cas-mediated genome editing technique coupled with “heterologous recombineering” using linear single-stranded (SSDR for single-stranded DNA recombineering) or double-stranded DNA (DSDR for double-stranded DNA recombineering) templates, have been developed and successfully applied in *E. coli* (78). In Archaea, gene silencing has been established in *Sulfolobus solfataricus*, *S. islandicus* and *Haloferax volcanii* using the endogenous CRISPR-Cas systems (reviewed in 77, 79). More recently, Nayak and Metcalf have harnessed a bacterial Cas9 protein for genome editing in the mesophilic archaeon *Methanosarcina acetovorans* (80). Hopfully a thermophilic counterpart of the CRISPR-Cas9 system (or other class 2 systems) will finally be established to perform genome editing in hyperthermophilic species, which are difficult to manipulate genetically. From that perspective, the diversity of CRISPR-Cas systems and mobile genetic elements, which remain to be fully explored, is a treasure trove for future exploitation.

## APPLICATION OF CRISPR-CAS9 FOR PURPOSES OTHER THAN GENOME EDITING

The CRISPR loci are encoded by many bacterial and archaeal organisms and are remarkably diverse, and thus they have been used as genetic markers for species identification and typing, even before the elucidation of the actual function of the CRISPR-Cas, as described above. For example, typing of *Mycobacterium tuberculosis* is useful for diagnostic and epidemiological purposes (20, 81). Typing by using CRISPR has been applied to *Yersinia pestis* (4, 82), *Salmonella* (83, 84), and *Corynebacterium diphtheriae* (85). CRISPR-Cas9 can be used as an antimicrobial agent by cleaving the genomes of pathogenic bacteria, as an antibiotic agent with a novel mechanism of action. It is expected to be a valuable remedy for the control of antibiotic-resistant bacteria. For example, antibiotic-resistant bacteria, such as *Staphylococcus*, infecting the skin of mice were selectively killed using CRISPR-Cas9 (86). CRISPR-Cas9 also reportedly prevented intestinal infection by pathogenic *E. coli* (87). Although there are technical challenges, such as delivery methods, which must be overcome before CRISPR-Cas can be used as a safe therapeutic agent, active research in this direction is ongoing and is expected to yield solutions in the near future. Furthermore, imparting phage resistance in specific strains by the CRISPR-Cas system is extremely useful for protecting various beneficial bacteria in the fermented food industry from phage infection during the production process.

Since the HNH nuclease domain and the RuvC nuclease domain are responsible for the DNA cleavage activity of Cas9, Cas9 mutants devoid of cleavage activity (dCas9) were obtained by replacing the amino acids within each active center. The dCas9 protein is a useful tool for molecular biology experiments to regulate gene expression. CRISPR-dCas9 binds to the target DNA sequence, but cannot cleave it. This activity of CRISPR-dCas9 is applicable to the labeling of a specific position, by fusing green fluorescent protein (GFP) to dCas9, which binds to the target sequence depending on the sgRNA sequence (88). In addition to this live intracellular site-specific labeling, gene expression can be artificially controlled by linking dCas9 to either the promoter region or the open reading frame of a gene (89-91). dCas9 can also be fused with a transcription activator or the ω subunit of bacterial RNA polymerase. However, it seems not to be as easy as compared with suppression, although ingenious attempts have been made to promote transcription by designing a guide sequence that ensures binding of dCas9 to a specific promoter.

The dCas9 protein is also useful for the techniques to reduce off-target cleavage in the genomes. An artificial CRISPR-Cas nuclease RFN (RNA-guided FokI nuclease), in which the nuclease domain of FokI is fused to dCas9 like ZFN or TALEN, was developed by designing the guide RNA so that the nuclease domain can form a dimer at the target site. Since it can be used for double-strand cleavage with different guide RNAs for top and bottom DNA strands, the probability of non-specific binding decreases (92-94). The reduction of off-target cleavage was also achieved by using Cas9 nickase (Cas9n). A mutant Cas9, in which the Asp10 active residue in the RuvC domain was substituted with alanine, showed a nickase activity that cleaved only one strand of the target site with an appropriate sgRNA (33,34). Therefore, nicking of both DNA strands by a pair of Cas9 nickases with different sgRNA leads to site-specific double-strand DNA breaks (DSBs). This paired nickase strategy can reduce off-target activity by 50-

to 1,500-fold in cell lines and to facilitate gene knockout in mouse zygotes without sacrificing on-target cleavage efficiency (95).

A method for site-specifc mutagenesis of genomic DNA by fusion of dCas9 with a cytidine deaminase has been developed (96). The sgRNA-induced cytidine deaminase causes base substitution at the target site without cutting DNA. This method significantly reduces cytotoxicity compared to artificial nucleases and Cas9 nuclease, and efficiently achieves intended modifications.

Another interesting solution was to split the Cas9 protein into two parts and reconstitute the Cas9 nuclease from the corresponding proteins (97, 98). The photoactivatable Cas9 (paCas9), which is activated by light irradiation, can be used for conditional genome editing. The activity of paCas9 is about 60% compared with the original Cas9, but it can be fully used for cutting the desired double-strand by light irradiation from the outside without changing the culture conditions (99).

Thus, as described above, the genome editing technique using the CRISPR-Cas immune system is not limited to the use of *S. pyogenes* CRISPR-Cas9, but further variants continue to be developed. These devices will certainly contribute to improvement of genome editing technologies.

## CONCLUDING REMARKS

Only 30 years have passed since one of the authors of this review discovered unique repeated sequence in the *E. coli* genome at the onset of his post-doc career. It was impossible to predict the possible function of this enigmatic sequence at the time; however, genomic revolution in the mid-90's, coupled with development of powerful bioinformatics tools eventually enabled elucidation of the CRISPR functions. CRISPR arrays and Cas proteins, broadly distributed in the genomes of prokaryotes, especially in Archaea, are now known to constitute the highly efficient acquired immunity system. Although discovery of the CRISPR-Cas by itself was a great feat of fundamental biology, it also led to the development of next-generation tools for genetic engineering. The development of the genome editing technology by CRISPR-Cas9 reminds of the times when the PCR was born.

When *in vitro* genetic engineering techniques using restriction endonucleases and nucleic acid modifying enzymes were established, it was still often a complex task to clone a single gene (as in the case of the *iap* gene). However, this difficulty was alleviated by the invention of PCR using a thermostable DNA polymerase that profoundly boosted the application of genetic engineering techniques in all biological laboratories worldwide. The discovery of a thermostable DNA polymerase was critical for the "PCR revolution" because it enabled the design of a PCR apparatus for practical use. Similarly, in the case of genome editing, the CRISPR revolution was made possible by identifying the right enzymatic system (Cas9) that could simplify the methodology to exploit the potential of the CRISPR-Cas system. The curiosity of a mysterious repetitive sequence and a sustained inquiry mind for elucidating its function brought grand discoveries.

## ACKNOWLEDGEMENTS

This perspective manuscript stemmed from discussions during Dr. Ishino's stay in Dr. Forterre's laboratory thanks to the generous support from the program of visiting researchers in Institut Pasteur and also the two-country exchange program of the Japan Society for the Promotion of Science (JSPS). PF is supported by an ERC grant from the European Union's Seventh Framework Program (FP/2007-2013)/Project. Finally, we apologize to the researchers whose work was not cited due to space limitations.

## REFERENCES

1. Ishino Y, Shinagawa H, Makino K, Amemura M, Nakata A. 1987. Nucleotide sequence of the *iap* gene, responsible for alkaline phosphatase isozyme conversion in *Escherichia coli*, and identification of the gene product, J Bacteriol 169: 5429–5433.
2. Mojica MJ, Juez G, Rodríguez-Valera F. 1993. Transcription at different salinities of *Haloferax mediterranei* sequences adjacent to partially modified PstI sites. Mol Microbiol 9: 613–621.
3. Mojica FJM, Díez-Villaseñor C, García-Martínez J, Soria E. 2005. Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. J Mol Evol 60: 174–182.
4. 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.
5. 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.
6. Makarova KS, Aravind L, Grishin NV, Rogozin IB, Koonin EV, A  (2002) DNA repair system specific for thermophilic Archaea and bacteria predicted by genomic context analysis. Nucleic Acids Res 30: 482–496.

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

8. Makarova KS, Grishin NV, Shabalina SA, Wolf YI, Koonin EV. 2006. A putative RNA-interference-based immune system in prokaryotes: computational analysis of the predicted enzymatic machinery, functional analogies with eukaryotic RNAi, and hypothetical mechanisms of action. Biol Direct 1: 7.

9. Morange M. 2015. What history tells us XXXVII. CRISPR-Cas: The discovery of an immune system in prokaryotes. J Biosci 40: 221-223

10. Morange M. 2015. What history tells us XXXIX. CRISPR-Cas: From a prokaryotic immune system to a universal genome editing tool. J Biosci 40: 829-832.

11. Lander ES. 2016. The Heroes of CRISPR. Cell 164: 18–28.

12. Mojica FJ, Rodriguez-Valera F. 2016. The discovery of CRISPR in archaea and bacteria, FEBS J 283: 3162–3169.

13. Barrangou R, Horvath PA. 2017. A decade of discovery: CRISPR functions and applications. Nat Microbiol 2: 17092.

14. Han W, She Q. 2017. CRISPR history: discovery, characterization, and prosperity. Prog Mol Biol Transl Sci 152:1-21.

15. Nakata A, Shinagawa H, Amemura M. 1982. Cloning of alkaline phosphatase isozyme gene (*iap*) of *Escherichia coli.* Gene 19: 313–319.

16. Stern MJ, Ames GF-L, Smith NH, Robinson EC, Higgins CF. 1984. Repetitive extragenic palindromic sequences: a major component of the bacterial genome, Cell 37: 1015–1026.

17. Newburg SF, Smith NH, Robinson EC, Hiles IE, Higgins CF. 1987. Stabilization of translationally active mRNA by prokaryotic REP sequences, Cell 48: 297–310.

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

19. Hermans PW, van Soolingen D, Bik EM, de Haas PE, Dale JW, van Embden JD. 1991. 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: 2695–2705.

20. 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.

21. Mojica FJ, Ferrer C, Juez G, Rodríguez-Valera F. 1995. Long stretches of short tandem repeats are present in the largest replicons of the Archaea Haloferax. mediterranei and *Haloferax volcanii* and could be involved in replicon partitioning, Mol Microbiol 17: 85–93.

22. Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage, AR Bult CJ, Tomb JF, Dougherty BA, Merrick JM, McKenney K, Sutton G, FitzHugh W, Fields C, Gocayne JD, Scott J, Shirley R, Liu Ll, Glodek A, Kelley JM, Weidman JF, Phillips CA, Spriggs T, Hedblom E, Cotton MD, Utterback TR, Hanna MC, Nguyen DT, Saudek DM, Brandon RC, Fine LD, Fritchman JL, Fuhrmann JL, Geoghagen NSM, Gnehm CL, McDonald LA, Small KV, Fraser CM, Smith HO, Venter JC. 1995. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd, Science 269 : 496–512.

23. Bult CJ, White O, Olsen GJ, Zhou L, Fleischmann RD, Sutton GG, Blake JA, FitzGerald LM, Clayton RA, Gocayne JD, Kerlavage AR, Dougherty BA, Tomb JF, Adams MD, Reich CI, Overbeek R, Kirkness EF, Weinstock KG, Merrick JM, Glodek A, Scott JL, Geoghagen NS, Venter JC. 1996. Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. Science 273: 1058–1073.

24. Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M, Louis EJ, Mewes HW, Murakami Y, Philippsen P, Tettelin H, Oliver SG. 1996. Life with 6000 genes. Science 274: 546–567.

25. She Q, Singh RK, Confalonieri F, Zivanovic Y, Allard G, Awayez MJ, Chan-Weiher CC, Clausen IG, Curtis BA, De Moors A, Erauso G, Fletcher C, Gordon PM, Heikamp-de Jong I, Jeffries AC, Kozera CJ, Medina N, Peng X,

Thi-Ngoc HP, Redder P, Schenk ME, Theriault C, Tolstrup N, Charlebois RL, Doolittle WF, Duguet M, Gaasterland T, Garrett RA, Ragan MA, Sensen CW, Van der Oost J. 2001. The complete genome of the crenarchaeon Sulfolobus solfataricus P2. Proc Natl Acad Sci USA. 98: 7835-7840.

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

27. Mojica FJ, Díez-Villaseñor C, Soria E, Juez G. 2000. Biological significance of a family of regularly spaced repeats in the genomes of Archaea, Bacteria and mitochondria. Mol Microbiol 36: 244–246.

28. 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.

29. Marraffini LA, Sontheimer EJ. 2008. CRISPR interference limits horizontal gene transfer in *Staphylococci* by targeting DNA. Science 322: 1843–1845.

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

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

32. Sapranauskas R, Gasiunas G, Fremaux C, Barrangou R, Horvath P, Siksnys V. 2011. The *Streptococcus thermophilus* CRISPR/Cas system provides immunity in *Escherichia coli,* Nucleic Acids Res 39: 9275–9282.

33. Gasiunas G, Barrangou R, Horvath P, Siksnys V. 2012. Cas9-crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria, Proc Natl Acad Sci USA 109: E2579–2586.

34. Jinek M, Chylinski K, Fonfara I, Hauer M, Doudna JA, Charpentier EA. 2012. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. Science 337: 816–821.

35. Cong L, Ran FA, Cox D, Lin S, Barretto R, Habib N, Hsu PD, Wu X, Jiang W, Marraffini LA, Zhang F. 2013. Multiplex genome engineering using CRISPR/Cas systems. Science 339: 819–823.

36. Mali P, Yang L, Esvelt KM, Aach J, Guell M, DiCarlo JE, Norville JE, Church GM. 2013. RNA-guided human genome engineering via Cas9. Science 339: 823–826.

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

38. Wiedenheft B, Sternberg SH, Doudna JA. 2012. RNA-guided genetic silencing systems in bacteria and archaea. Nature 482: 331–338.

39. Jackson SA, McKenzie RE, Fagerlund RD, Kieper SN, Fineran PC, Brouns SJ. CRISPR-Cas: Adapting to change. Science 356 eaal5056.

40. Tamulaitis G, Venclovas C, Siksnys V. 2017. Type III CRISPR-Cas Immunity: Major differences brushed aside, Trends Microbiol 25: 49-61.

41. Mohanraju P, Makarova KS, Zetsche B, Zhang F, Koonin EV, van der Oost J. 2016. Diverse evolutionary roots and mechanistic variations of the CRISPR-Cas systems. Science 353: aad5147.

42. Charpentier E, Richter H, van der Oost J, White MF. 2015. Biogenesis pathways of RNA guides in archaeal and bacterial CRISPR-Cas adaptive immunity. FEMS Microbiol Rev 39: 428-441.

43. Marraffini LA, Sontheimer EJ. 2010. CRISPR interference: RNA-directed adaptive immunity in bacteria and archaea. Nat Rev Genet. 11:181-190.

44. Burstein D, Sun CL, Brown CT, Sharon I, Anantharaman K, Probst AJ, Thomas BC, Banfield JF. 2016. Major bacterial lineages are essentially devoid of CRISPR-Cas viral defence systems. Nat Commun. 7:10613.

45. Shmakov S, Smargon A, Scott D, Cox D, Pyzocha N, Yan W, Abudayyeh OO, Gootenberg JS, Makarova KS, Wolf YI, Severinov K, Zhang F, Koonin EV. 2017. Diversity and evolution of class 2 CRISPR-Cas systems. Nat Rev Microbiol 15: 169-182.

46. Burstein D, Harrington LB, Strutt SC, Probst AJ, Anantharaman K, Thomas BC, Doudna JA, Banfield JF, 2017. New CRISPR-Cas systems from uncultivated microbes. Nature 542: 237-241.

47. Rouillon C, Zhou M, Zhang J, Politis A, Beilsten-Edmands V, Cannone G, Graham S, Robinson CV, Spagnolo L, White MF. 2013. Structure of the

CRISPR interference complex CSM reveals key similarities with cascade. Mol Cell 52: 124-134.

48. Jackson RN, Wiedenheft B. 2015. A conserved structural chassis for mounting versatile CRISPR RNA-guided immune responses, Mol Cell 58: 722-728.

49. Makarova KS, Haft DH, Barrangou R, Brouns SJ, Charpentier E, Horvath P, Moineau S, Mojica FJ, Terns RM, Terns MP, White MF, Yakunin AF, Garrett RA, van der Oost J, Backofen R, Koonin EV. 2011. Evolution and classification of the CRISPR-Cas systems. Nat Rev Microbiol 9: 467–477.

50. Koonin EV, Makarova KS. 2013. CRISPR-Cas: evolution of an RNA-based adaptive immunity system in prokaryotes, RNA Biol 10: 679-686.

51. Koonin EV, Makarova KS, Zhang F. 2017. Diversity, classification and evolution of CRISPR-Cas systems. Curr Opin Microbiol 37: 67-78.

52. Venclovas C. 2016. Structure of Csm2 elucidates the relationship between small subunits of CRISPR-Cas effector complexes. FEBS Lett 590: 1521-1529.

53. Makarova KS, Wolf YI, Alkhnbashi OS, Costa F, Shah SA, Saunders SJ, Barrangou R, Brouns SJ, Charpentier E, Haft DH, Horvath P, Moineau S, Mojica FJ, Terns RM, Terns MP, White MF, Yakunin AF, Garrett RA, van der Oost, Backofen R, Koonin EV. 2015. An updated evolutionary classification of CRISPR-Cas systems. Nat Rev Microbiol 13: 722–736.

54. Garrett RA, Vestergaard G, Shah SA. 2011. Archaeal CRISPR-based immune systems: exchangeable functional modules. Trends Microbiol. 19:549-556.

55. Nishimasu H, Ran FA, Hsu PD, Konermann S, Shehata SI, Dohmae N, Ishitani R, Zhang F, Nureki O. 2014. Crystal structure of Cas9 in complex with guide RNA and target DNA. Cell 156: 935–949.

56. Yamano T, Nishimasu H, Zetsche B, Hirano H, Slaymaker IM, Li Y, Fedorova I, Nakane T, Makarova KS, Koonin EV, Ishitani R, Zhang F, Nureki O. 2016. Crystal structure of Cpf1 in complex with guide RNA and target DNA. Cell 165: 949–962.

57. Zetsche B, Gootenberg JS, Abudayyeh OO, Slaymaker IM, Makarova KS, Essletzbichler P, Volz SE, Joung J, van der Oost J, Regev A, Koonin EV, Zhang F. 2015. Cpf1 is a single RNA-guided endonuclease of a class 2 CRISPR-Cas system. Cell 163: 759–771.

58. Shmakov S, Abudayyeh OO, Makarova KS, Wolf YI, Gootenberg JS, Semenova E, Minakhin L, Joung J, Konermann S, Severinov K, Zhang Z, Koonin EV. 2015. Discovery and functional characterization of diverse class 2 CRISPR- Cas systems. Mol Cell 60: 385–397.

59. Jiang W, Samai P, Marraffini LA, 2016. Degradation of phage transcripts by CRISPR-associated RNases enables type III CRISPR-Cas immunity. Cell 164: 710–721.

60. Abudayyeh OO, Gootenberg JS, Konermann S, Joung J, Slaymaker IM, Cox DB, Shmakov S, Makarova KS, Semenova E, Minakhin L, Severinov K, Regev A, Lander ES, Koonin EV, Zhang F, 2016. C2c2 is a single-component programmable RNA-guided RNA-targeting CRISPR effector, Science 353: 6299.

61. Murugan K, Babu K, Sundaresan R, Rajan R, Sashital DG. 2017. The Revolution Continues: Newly Discovered Systems Expand the CRISPR-Cas Toolkit. Mol Cell 68: 15-25.

62. Makarova KS, Wolf YI, Forterre P, Prangishvili D, Krupovic M, Koonin EV. 2014. Dark matter in archaeal genomes: a rich source of novel mobile elements, defense systems and secretory complexes, Extremophiles 18: 877-893.

63. Makarova KS, Wolf YI, Koonin EV, 2013. The basic building blocks and evolution of CRISPR-cas systems, Biochem Soc Trans 41: 1392–1400.

64. Krupovic M, Makarova KS, Forterre P, Prangishvili D, Koonin E.V. 2014. Casposons: a new superfamily of self-synthesizing DNA transposons at the origin of prokaryotic CRISPR-Cas immunity, BMC Biol 12: 36.

65. Krupovic M, Shmakov S, Makarova KS, Forterre P, Koonin EV. 2016. Recent mobility of casposons, self-synthesizing transposons at the origin of the CRISPR-Cas immunity. Genome Biol Evol 8: 375-386.

66. Krupovic M, Koonin EV. 2016. Self-synthesizing transposons: unexpected key players in the evolution of viruses and defense systems. Curr Opin Microbiol 31: 25-33.

67. Hickman AB, Dyda F. 2015. The casposon-encoded Cas1 protein from *Aciduliprofundum boonei* is a DNA integrase that generates target site duplications. Nucleic Acids Res 43: 10576-10587.

68. Béguin P, Charpin N, Koonin EV, Forterre P, Krupovic M. 2016. Casposon integration shows strong target site preference and recapitulates protospacer integration by CRISPR-Cas systems. Nucleic Acids Res 44: 10367-10376.

69. Yosef I, Goren MG, Qimron U. 2012. Proteins and DNA elements essential for the CRISPR adaptation process in *Escherichia coli.* Nucleic Acids Res. 40: 5569-5576.

70. Rollie C, Schneider S, Brinkmann AS, Bolt EL, White MF. 2015. Intrinsic sequence specificity of the Cas1 integrase directs new spacer acquisition, Elife 4: e08716.

71. Koonin EV, Krupovic M. 2015. Evolution of adaptive immunity from transposable elements combined with innate immune systems. Nat Rev Genet 16: 184–192.

72. Krupovic M, Béguin P, Koonin EV. 2017. Casposons: mobile genetic elements that gave rise to the CRISPR-Cas adaptation machinery. Curr Opin Microbiol 38: 36-43.

73. Jiang W, Bikard D, Cox D, Zhang F, Marraffini LA. 2013. RNA-guided editing of bacterial genomes using CRISPR-Cas systems. Nat Biotechnol 31: 233–239.

74. Jiang Y, Chen B, Duan C, Sun B, Yang J, Yang S. 2015. Multigene editing in the *Escherichia coli* genome via the CRISPR-Cas9 system. Appl Environ Microbiol 81: 2506–2514.

75. Oh JH, van Pijkeren JP. 2014 CRISPR-Cas9-assisted recombineering in *Lactobacillus reuteri.* Nucleic Acids Res 42: e131.

76. Charpentier E, Marraffini LA. 2014. Harnessing CRISPR-Cas9 immunity for genetic engineering. Curr Opin Microbiol 19: 114-119.

77. Gophna U, Allers T, Marchfelder A. 2017. Finally, archaea get their CRISPR-Cas toolbox. Trends Microbiol 25: 430-432.

78. Mougiakos I, Bosma EF, de Vos WM, van Kranenburg R, van der Oost J. 2016. Next generation prokaryotic engineering: the CRISPR-Cas toolkit. Trends Biotechnol 34: 575-587.

79. Peng N, Han W, Li Y, Liang Y, She Q. 2017. Genetic technologies for extremely thermophilic microorganisms of *Sulfolobus*, the only genetically tractable genus of crenarchaea. Sci China Life Sci 60: 370-385.

80. Nayak DD, Metcalf WW. 2017. Cas9-mediated genome editing in the methanogenic archaeon *Methanosarcina acetivorans.* Proc Natl Acad Sci USA 114: 2976-2981.

81. 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.

82. Cui Y, Li Y, Gorgé 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 2008 3: e2652.

83. Liu F, Barrangou R, Gerner-Smidt P, Ribot EM, Knabel SJ, Dudley EG. 2011. 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.

84. Liu F, Kariyawasam S, Jayarao BM, Barrangou R, Gerner-Smidt P, Ribot EM, Knabel SJ, Dudley EG. 2011. Subtyping *Salmonella enterica serovar enteritidi*s 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.

85. 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.

86. Bikard D, Euler CW, Jiang W, Nussenzweig PM, Goldberg GW, Duportet X, Fischetti VA, Marraffini LA. 2014. Exploiting CRISPR-Cas nucleases to produce sequence-specific antimicrobials. Nat Biotechnol 32:1146-1150.

87. Citorik RJ, Mimee M, Lu TK. 2014. Sequence-specific antimicrobials using efficiently delivered RNA-guided nucleases. Nat Biotechnol 32: 1141-1145.

88. Chen B, Gilbert LA, Cimini BA, Schnitzbauer J, Zhang W, Li GW, Park J, Blackburn EH, Weissman JS, Qi LS, Huang B. 2013. Dynamic imaging of genomic loci in living human cells by an optimized CRISPR/Cas system. Cell 155:1479-1491.

89. Bikard D, Jiang W, Samai P, Hochschild A, Zhang F, Marraffini LA. 2013. Programmable repression and activation of bacterial gene expression using an engineered CRISPR-Cas system. Nucleic Acids Res 41: 7429-7437.

90. Qi LS, Larson MH, Gilbert LA, Doudna JA, Weissman JS, Arkin AP, Lim WA. 2013. Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. Cell 152:1173-1183.

91. Gilbert LA, Larson MH, Morsut L, Liu Z, Brar GA, Torres SE, Stern-Ginossar N, Brandman O, Whitehead EH, Doudna JA, Lim WA, Weissman JS, Qi LS. 2013. CRISPR-mediated modular RNA-guided regulation of transcription in eukaryotes. Cell 154: 442-451.

92. Guilinger JP, Thompson DB, Liu DR. 2014. Fusion of catalytically inactive Cas9 to FokI nuclease improves the specificity of genome modification. Nat Biotechnol 32: 577-582.

93. Tsai SQ, Wyvekens N, Khayter C, Foden JA, Thapar V, Reyon D, Goodwin MJ, Aryee MJ, Joung JK. 2014. Dimeric CRISPR RNA-guided FokI nucleases for highly specific genome editing. Nat Biotechnol 32: 569-576.

94. Havlicek S, Shen Y, Alpagu Y, Bruntraeger MB, Zufir NB, Phuah ZY, Fu Z, Dunn NR, Stanton LW. 2017. Re-engineered RNA-Guided FokI-Nucleases for Improved Genome Editing in Human Cells. Mol Ther 25: 342-355.

95. Ran FA, Hsu PD, Lin CY, Gootenberg JS, Konermann S, Trevino AE, Scott DA, Inoue A, Matoba S, Zhang Y, Zhang F. 2013. Double nicking by RNA-guided CRISPR Cas9 for enhanced genome editing specificity. Cell 154:1380-1389.

96. Nishida K, Arazoe T, Yachie N, Banno S, Kakimoto M, Tabata M, Mochizuki M, Miyabe A, Araki M, Hara KY, Shimatani Z, Kondo A. 2016. Targeted nucleotide editing using hybrid prokaryotic and vertebrate adaptive immune systems Science. 353: 6305.

97. Zetsche B, Volz SE, Zhang F. 2015. A split-Cas9 architecture for inducible genome editing and transcription modulation. Nat Biotechnol 33:139-142.

98. Wright AV, Sternberg SH, Taylor DW, Staahl BT, Bardales JA, Kornfeld JE, Doudna JA. 2015. Rational design of a split-Cas9 enzyme complex. Proc Natl Acad Sci USA. 112: 2984-2989.

99. Nihongaki Y, Kawano F, Nakajima T, Sato M. 2015. Photoactivatable CRISPR-Cas9 for optogenetic genome editing. Nat Biotechnol 33, 755-760.

## Figure legends

**FIG 1** The structural features of CRISPR. The repeat sequences with constant length generally have dyad symmetry to form a palindromic structure (shown by arrows). Two examples are shown by the first identified CRISPR from *E. coli* (bacteria) and *H. mediterranei* (archaea), respectively. The spacer regions are also constant length, but no sequence homology.

**FIG 2** The first CRISPR found in *E. coli.* As a result of the *iap* gene analysis from *E. coli*, a very ordered repeating sequence was found downstream of the *iap* gene. The conserved sequence unit was repeated 5 times with constant length of spaces in 1987. It turns out that the repeat was 14 times in total by the subsequent genome analysis. The *cas* gene cluster was also identified at the downstream region.

**FIG 3** The first CRISPR sequence in *E. coli*. The exact same region, downstream of the *iap* gene, which was found in 1987 by a conventional dideoxy-sequencing was read by a cycle-sequencing with fluorescent labeling recently. The CRISPR repeat units are shown by pink shadow.

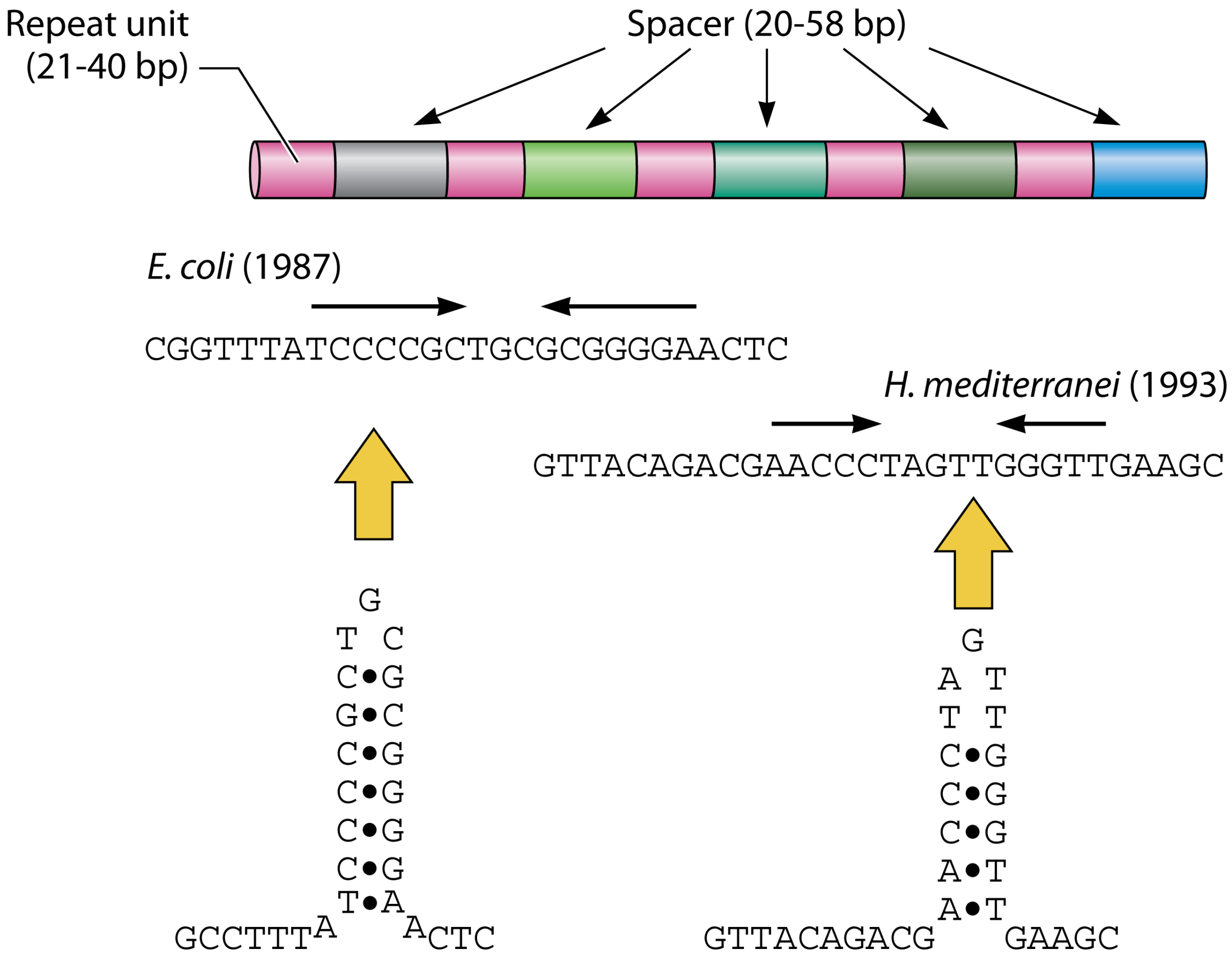
**FIG 4** Process of CRISPR-Cas acquired immune system. A. Adaptation: The invading DNA is recognized by Cas proteins, fragmented and incorporated into the spacer region of CRISPR and stored in the genome. B. Expression: Pre-crRNA is generated by transcription of the CRISPR region, and is processed into smaller units of RNA, named crRNA. Interference: By taking advantage of the homology of the spacer sequence present in crRNA, foreign DNA is captured and a complex with Cas protein having nuclease activity cleaves DNA.

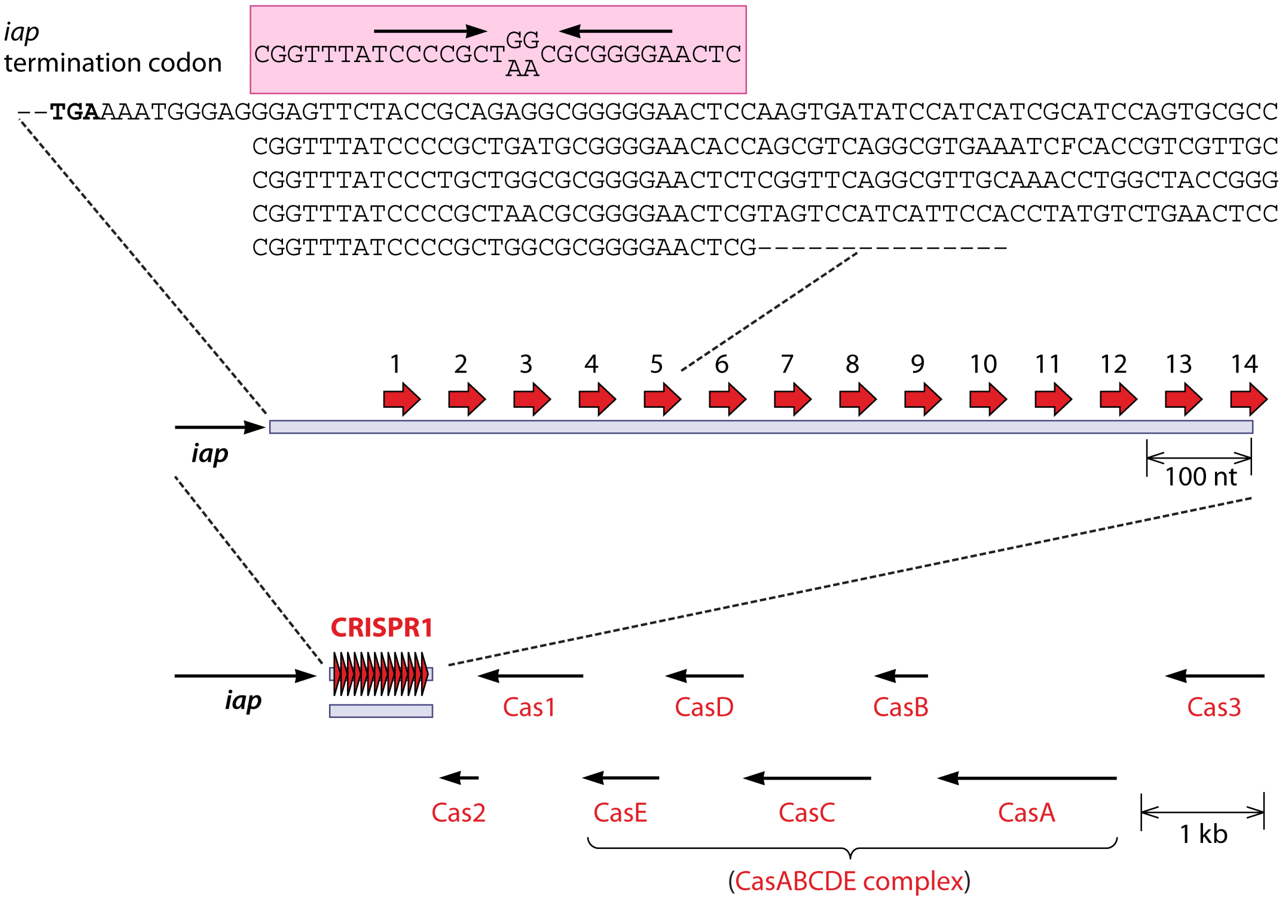
**FIG. 5** Genome editing by CRISPR-Cas9. The principle of genome editing is the cleavage of double-stranded DNA at a targeted position on the genome. The Type II is the simplest as a targeted nuclease among the CRISPR-Cas systems. The CRISPR RNA (crRNA), having a sequence homologous to the target site, and trans activating RNA (tracrRNA) are enough to bring the Cas9 nuclease to the target site. The artificial linkage of crRNA and tracrRNA into one RNA chain (single guide RNA; sgRNA) has no effect on function. Once the Cas9-gRNA complex cleaves the target gene, it is easy to disrupt the function of the gene by deletion or insertion mutation. This overwhelmingly simple method is now rapidly spreading as a practical genomic editing technique.

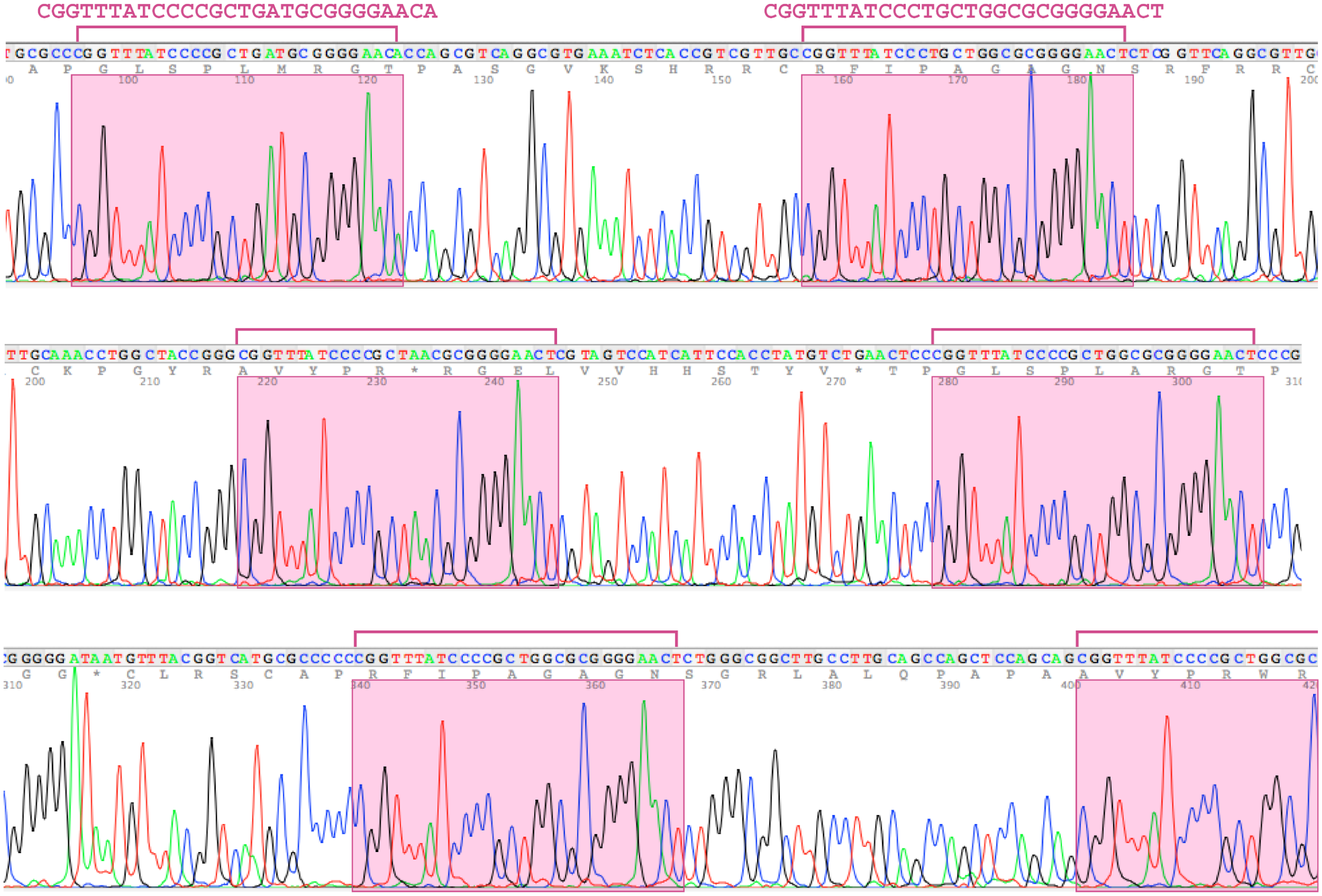
**FIG 6** Most recent classification of CRISPR-Cas immune systems. A. Based on the detailed sequence analyses and gene organization of the Cas proteins, CRISPR-Cas was classified into two major classes depending on whether the effector is a complex composed of multiple Cas proteins or a single effector. In addition to the conventional types I, II and III, the types IV and V were added to the classes 1 and 2, respectively. Types IV and V are those which do not have Cas1 and Cas2, necessary for adaptation process, in the same CRISPR loci. Type VI was added most recently in class 2. B. Chart showing the proportions of identified CRISPR-*cas* loci in the total genomes of bacteria and archaea referred from the literatures (51, 53). The proportions of loci that encode incomplete systems or that could not be classified unambiguously are not included.

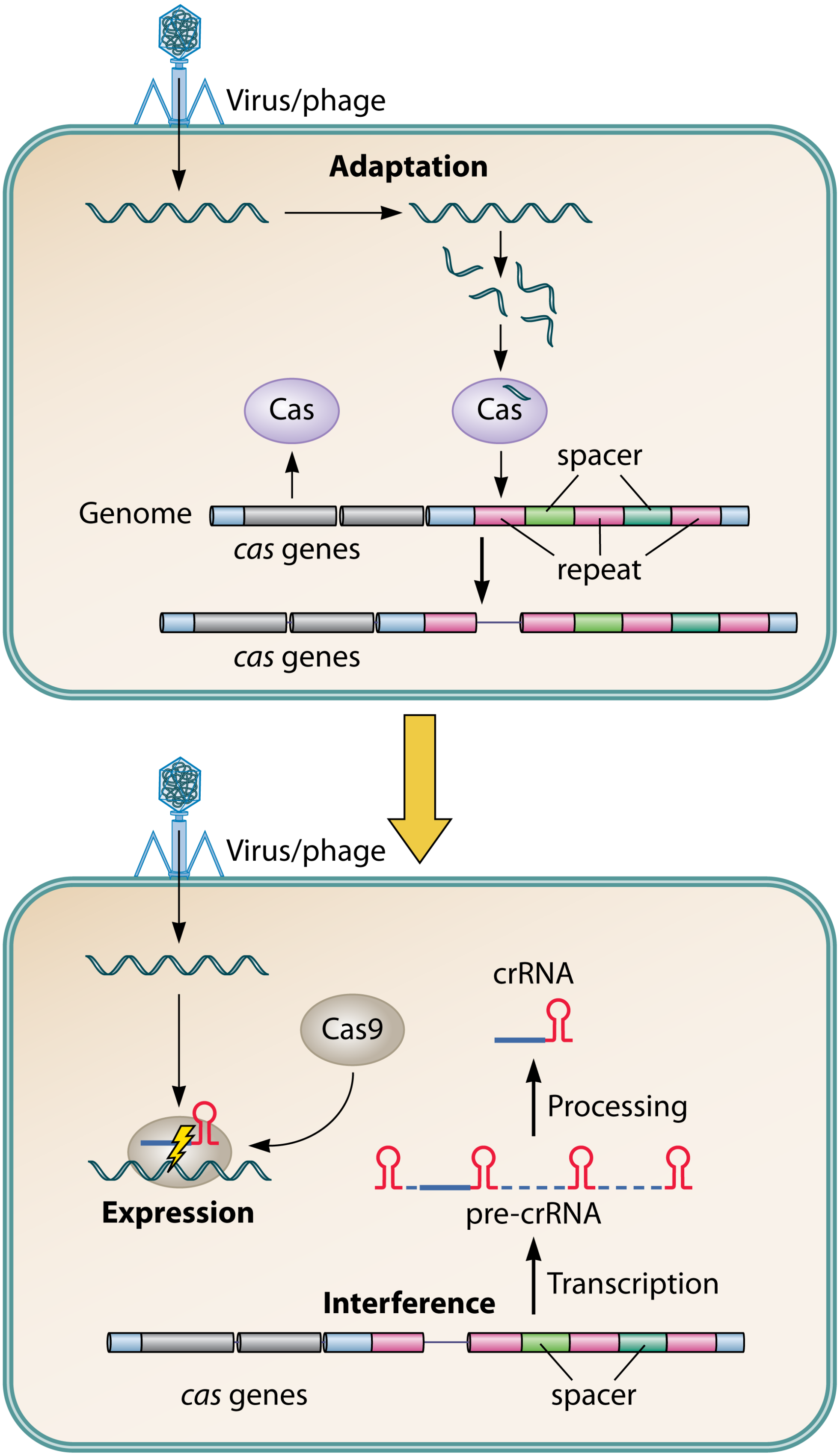
**FIG 7 Cleavage mechanism of target DNA by crRNA-tracrRNA-Cas9**

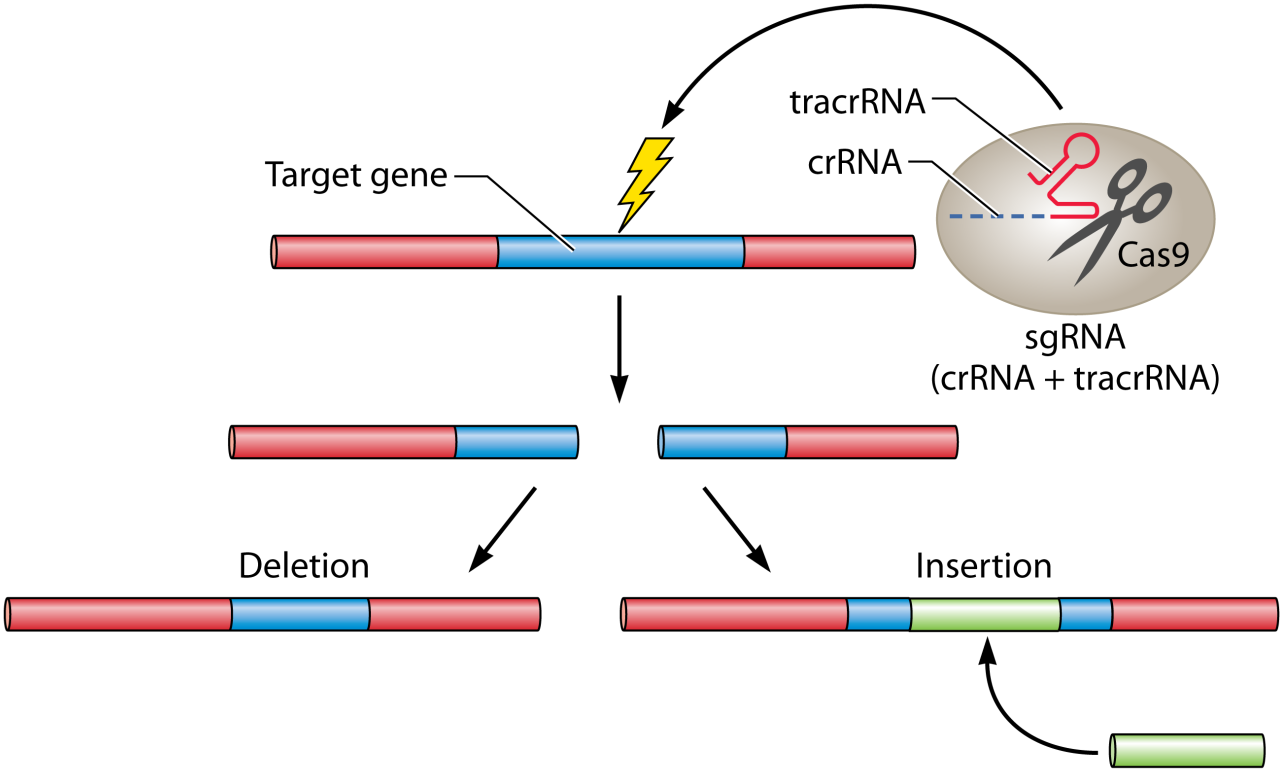
The Cas9-crRNA-tracrRNA complex binds to foreign DNA containing PAM, where Cas9 binds and starts to unwind the double-strand of the foreign DNA to induce duplex formation of crRNA and foreign DNA. Cas9 consists of two regions, called REC (recognition) lobe and NUC (nuclease) lobe. REC lobe is responsible for the nucleic acid recognition. NUC lobe contains the HNH and RuvC nuclease domains, and a C-terminal region containing PAM-interacting (PI) domain. The HNH domain and the RuvC domain cleave the DNA strand forming duplex with crRNA and the other DNA strand, respectively, so that double-strand break occurs in the target DNA.

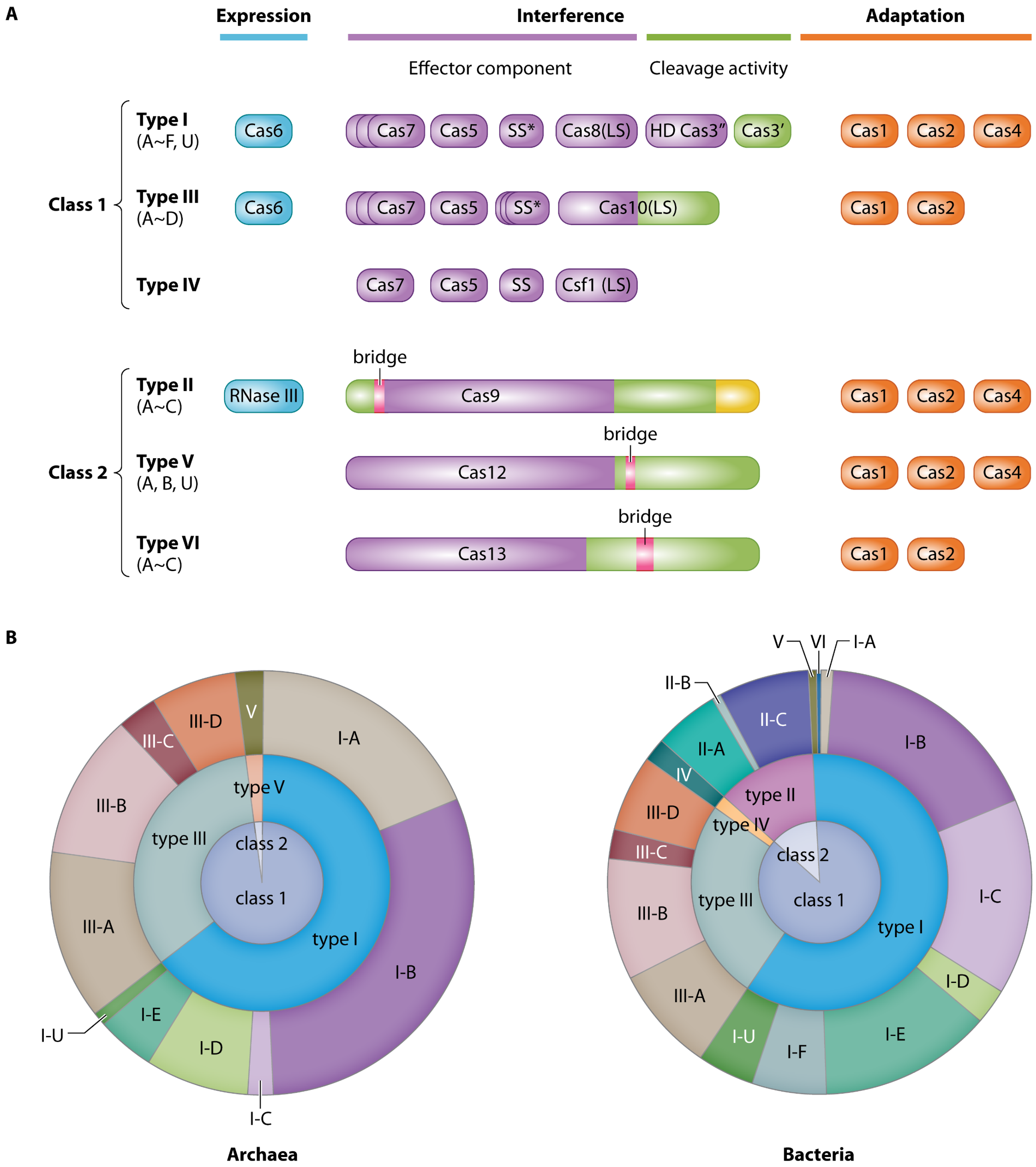

Repeat unit
(21-40 bp)
Spacer (20-58 bp)
*E. coli* (1987)
CGGTTTATCCCCGCTGCGCGGGGAACTC
*H. mediterranei* (1993)
GTTACAGACGAACCCTAGTTGGGTTGAAGC
G
T C
C•G
G•C
C•G
C•G
C•G
C•G
T•A
GCCTTTA ACTC
G
A T
T T
C•G
C•G
C•G
A•T
A•T
GTTACAGACG GAAGC

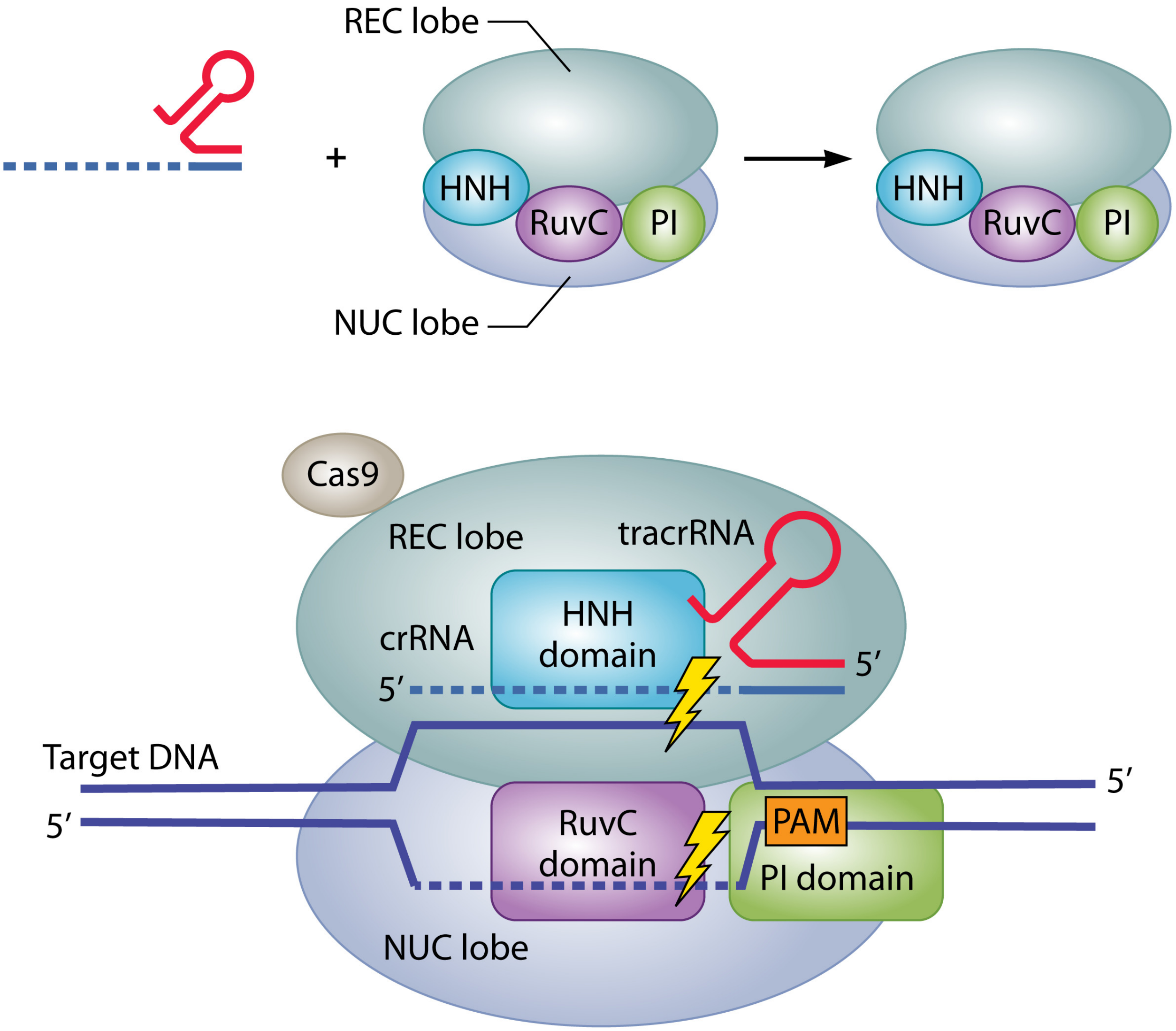

iap
termination codon
CGGTTTATCCCCGCT GG AA CGCGGGGAACTC
--TGAAAATGGGAGGGAGTTCTACCGCAGAGGCGGGGAACTCCAAGTGATATCCATCATCGCATCCAGTGCGCC
CGGTTTATCCCCGCTGATGCGGGGAACACCAGCGTCAGGCGTGAAATCFCACCGTCGTTGC
CGGTTTATCCCTGCTGGCGCGGGGAACTCTCGGTTCAGGCGTTGCAAACCTGGCTACCGGG
CGGTTTATCCCCGCTAACGCGGGGAACTCGTAGTCCATCATTCCACCTATGTCTGAACTCC
CGGTTTATCCCCGCTGGCGCGGGGAACTCG---------------
1
2
3
4
5
6
7
8
9
10
11
12
13
14
iap
100 nt
CRISPR1
iap
Cas1
CasD
CasB
Cas3
Cas2
CasE
CasC
CasA
1 kb
(CasABCDE complex)


CGGTTTATCCCCGCTGATGCGGGGAACA
CGGTTTATCCCTGCTGGCGCGGGGAACT
TGCGCCCGGTTTATCCCCGCTGATGCGGGGAACACCAGCGTCAGGCGTGAAATCTCACCGTCGTTGCCGGTTTATCCCTGCTGGCGCGGGGAACTCTCCGGTTCAGGCGTTG
A P G L S P L M R G T P A S G V K S H R R C R F I P A G A G N S R F R R C
100 110 120 130 140 150 160 170 180 190 200
TTGCAAACCTGGCTACCGGGCGGTTTATCCCCGCTAACGCGGGGAACTCGTAGTCCATCATTCCACCTATGTCTGAACTCCCGGTTTATCCCCGCTGGCGCGGGGAACTCCCG
C K P G Y R A V Y P R * R G E L V V H H S T Y V * T P G L S P L A R G T P
200 210 220 230 240 250 260 270 280 290 300 310
CGGGGGATAATGTTTACGGTCATGCGCCCCCCGGTTTATCCCCGCTGGCGCGGGAACTCTGGGCGGCTTGCCTTGCAGCCAGCTCCAGCAGCGGTTTATCCCCGCTGGCGC
G G * C L R S C A P R F I P A G A G N S G R L A L Q P A P A A V Y P R W R
310 320 330 340 350 360 370 380 390 400 410 420


Virus/phage
Adaptation
Cas
Cas
spacer
Genome
cas genes
repeat
cas genes
Virus/phage
crRNA
Cas9
Processing
Expression
pre-crRNA
Transcription
Interference
cas genes
spacer


tracrRNA
crRNA
Target gene
Cas9
sgRNA
(crRNA + tracrRNA)
Deletion
Insertion


A
Expression
Interference
Adaptation
Effector component
Cleavage activity
Class 1
Type I
(A~F, U)
Cas6
Cas7
Cas5
SS*
Cas8(LS)
HD Cas3"
Cas3'
Cas1
Cas2
Cas4
Type III
(A~D)
Cas6
Cas7
Cas5
SS*
Cas10(LS)
Cas1
Cas2
Type IV
Cas7
Cas5
SS
Csf1 (LS)
Class 2
Type II
(A~C)
RNase III
bridge
Cas9
Cas1
Cas2
Cas4
Type V
(A, B, U)
bridge
Cas12
Cas1
Cas2
Cas4
Type VI
(A~C)
bridge
Cas13
Cas1
Cas2
B
III-D
V
I-A
III-C
III-B
type V
type III
class 2
class 1
III-A
type I
I-B
I-U
I-E
I-D
I-C
Archaea
V
VI
I-A
II-B
II-C
I-B
II-A
IV
type II
III-D
type IV
III-C
class 2
I-C
III-B
type III
class 1
type I
I-D
III-A
I-U
I-F
I-E
Bacteria


REC lobe
+
HNH
RuvC
PI
NUC lobe
HNH
RuvC
PI
Cas9
REC lobe
tracrRNA
crRNA
HNH domain
5′
5′
Target DNA
5′
5′
RuvC domain
PAM
PI domain
NUC lobe