



HAL
open science

History of CRISPR-Cas from Encounter with a Mysterious Repeated Sequence to Genome Editing Technology

Yoshizumi Ishino, Mart Krupovic, Patrick Forterre

► **To cite this version:**

Yoshizumi Ishino, Mart Krupovic, Patrick Forterre. History of CRISPR-Cas from Encounter with a Mysterious Repeated Sequence to Genome Editing Technology. *Journal of Bacteriology*, 2018, 200, pp.e00580-17. 10.1128/JB.00580-17. pasteur-01977345

HAL Id: pasteur-01977345

<https://pasteur.hal.science/pasteur-01977345>

Submitted on 10 Jan 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution - NonCommercial - ShareAlike 4.0 International License

1

2

3

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

4

5

6

Yoshizumi Ishino,^{1,2,*} Mart Krupovic,¹ Patrick Forterre^{1,3}

7

8

¹Unité de Biologie Moléculaire du Gène chez les Extrêmophiles, Département de

9

Microbiologie, Institut Pasteur, F-75015, Paris, France, ²Department of

10

Bioscience and Biotechnology, Faculty of Agriculture, Kyushu University,

11

Fukuoka 812-8581, Japan. ³Institute of Integrative Cellular Biology, Université

12

Paris Sud, 91405 Orsay, Cedex France

13

14

15

Running title: Discovery and development of CRISPR-Cas research

16

17

* Correspondence to

18

Prof. Yoshizumi Ishino

19

Department of Bioscience and Biotechnology,

20

Faculty of Agriculture, Kyushu University,

21

Fukuoka 812-8581, Japan

22

ishino@agr.kyushu-u.ac.jp

23 **ABSTRACT**

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

45

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

47

48 **INTRODUCTION**

49 CRISPR-Cas systems are currently in the spotlight of active research in biology.

50 The first clustered regularly interspaced short palindromic repeats (CRISPR)

51 were detected 30 years ago by one of the authors of this review (YI) in

52 *Escherichia coli* in the course of the analysis of the gene responsible for isozyme

53 conversion of alkaline phosphatase (1). The structural features of CRISPR are

54 shown in Figure 1. At the time, it was hardly possible to predict the biological

55 function of these unusual repeated sequences due to the lack of sufficient DNA

56 sequence data, especially for mobile genetic elements. The actual function of this

57 unique sequence remained enigmatic right up until the mid-2000s. In 1993,

58 CRISPRs were for the first time observed in Archaea, specifically in *Haloferax*

59 *mediterranei* (2), and subsequently detected in an increasing number of bacterial

60 and archaeal genomes, since life science moved into genomic era. Conservation

61 of these sequences in two of the three domains of life was critical for appreciating

62 their importance. In the early 2000s, the discovery of sequence similarity

63 between the spacer regions of CRISPR and sequences of bacteriophages,

64 archaeal viruses and plasmids finally shed light on the function of CRISPR as an

65 immune system. This dramatic discovery by Mojica and others was grossly

66 underappreciated at that time, and was published in 2005 by three research

67 groups independently (3-5). In parallel, several genes previously proposed to
68 encode for DNA repair proteins specific for hyperthermophilic archaea (6) were
69 identified to be strictly associated with CRISPR, and designated as *cas*
70 (CRISPR-associated genes) (7). Comparative genomic analyses thus suggested
71 that CRISPR and Cas proteins (the *cas* gene products) actually work together
72 and constitute an acquired immunity system to protect the prokaryotic cells
73 against invading viruses and plasmids, analogous to the eukaryotic RNA
74 interference (RNAi) system (8).

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

82

83 **A PUZZLING SEQUENCE FROM BACTERIA CHALLENGES THE EARLY** 84 **SEQUENCING METHODOLOGY**

85 In the mid-80s, when studying isozyme conversion of alkaline
86 phosphatase (AP), one of us (YI), in an attempt to identify the protein responsible
87 for the isozyme conversion of AP in the periplasm of the *E. coli* K12 cells,
88 sequenced a 1.7 kbp *E. coli* DNA fragment spanning the region containing the *iap*

89 gene (designated from isozyme of alkaline phosphatase) (1). The isozyme of AP
90 was previously detected by biochemical and genetic analyses (15). At that time,
91 for conventional M13 dideoxy sequencing, single-stranded template DNA had to
92 be produced by cloning the target DNA into an M13 vector, whereas the dideoxy
93 chain-termination reaction was performed by Klenow fragment of *E. coli* Pol I.
94 The reaction products were labeled by incorporation of [$\alpha^{32}\text{P}$]dATP, and the
95 sequence ladder images were obtained by autoradiography. For sequencing, the
96 cloned DNA fragment had to be subcloned into M13 mp18 and 19 vectors (for the
97 coding and noncoding strands) after digestion into short fragments. During the
98 sequencing of the DNA fragment containing *iap*, one of the authors realized that
99 the same sequence appeared many times in different clones. Furthermore, it was
100 difficult to read the repeated sequences precisely, using the Klenow fragment at
101 37°C, because of non-specific termination of the dideoxynucleotide
102 incorporation reactions for the template DNA, due to secondary structure
103 formation by the palindromic sequence. This is why it took several months to
104 read the sequence of the CRISPR region precisely in 1987 (1). A peculiar
105 repeated sequence was detected downstream of the translation termination
106 codon for the *iap* gene (Fig. 2). It is remarkable that the exact same region can be
107 sequenced in just one day using current technology, by amplification of the target
108 region by PCR directly from the genome, followed by a fluorescent-labeling and
109 cycle-sequencing at 72°C (Fig. 3). The feature of the repetitive sequence was so
110 mysterious and unexpected that it was mentioned in the Discussion section, even

111 though its function was not understood (1). Notably, the same sequence
112 containing a dyad symmetry of 14 bp was repeated five times with a variable
113 32-nucleotide sequence interspersed between the repeats (Fig. 2).
114 Well-conserved nucleotide sequences containing a dyad symmetry, named REP
115 (Repetitive extragenic palindromic) sequences (16), had been previously found
116 in *E. coli* and *Salmonella typhimurium* and suggested to stabilize mRNA (17).
117 However, no similarities were found between the REP and the repeated
118 sequences detected downstream of the *iap* gene. In fact, this sequence was, at
119 the time, unique in sequence databases. As it later turned out, this was the first
120 encounter with a CRISPR sequence. Soon after, similar sequences were
121 detected by southern blot hybridization analysis in other *E. coli* strains (C600
122 and Ymel) and in two other members of the *Enterobacteriaceae*, *Shigella*
123 *dysenteriae* and *Salmonella typhimurium* (phylum *Proteobacteria*) (18).
124 Subsequently, similar repeated sequences were also found in members of the
125 phylum Actinobacteria, such as *Mycobacterium tuberculosis* (19), but not in the
126 closely related strain *M. leprae*, prompting the use of these highly polymorphic
127 repeated sequences for strain typing (20).

128

129 **DISCOVERY OF CRISPR IN ARCHAEA**

130 A major advance was made when similar repeated sequences were
131 identified by Mojica and co-workers in the archaeon *Haloferax mediterranei*
132 during the research on regulatory mechanisms allowing extremely halophilic

133 archaea to adapt to high salt environments (2). Transcription of the genomic
134 regions containing the repeated sequences was demonstrated by Northern blot
135 analysis (2), but compelling evidence for the processing of the transcripts into
136 several different RNA products was shown only more recently (12). The authors
137 first suggested that these repeated sequences could be involved in the
138 regulation of gene expression, possibly facilitating the conversion of the
139 double-stranded DNA from B to Z-form for the specific binding of a regulator
140 protein. It was indeed often suggested at that time that the high GC content of
141 halophilic genomes could facilitate such B-to-Z transition for regulatory purposes
142 at the high intracellular salt concentration characteristic of haloarchaea.
143 However, such explanation could not be valid for bacteria. Soon after, the same
144 authors found a similar repeated sequence in *Haloferax volcanii*, and
145 hypothesized that these repeated sequences could be involved in replicon
146 partitioning (21).

147 In the meantime, invention of the automated sequencing machines and
148 development of efficient procedures for DNA sequencing during the 90s
149 provided scientists for the first time with access to complete genome sequences.
150 Starting with *Haemophilus influenzae* (22), followed by *Methanocaldococcus*
151 *jannaschii* (23) and *Saccharomyces cerevisiae* (24), all three domains of life
152 entered into the genomics era. Then, the unusual repeated sequences
153 interspersed with non-conserved sequences, first detected in *E. coli* and *H.*
154 *mediterranei*, were identified in an increasing number of bacterial and archaeal

155 genomes, and were described using different names by different authors, such
156 as SRSRs, (Short Regularly Spaced Repeats (2), SPIDR (spacers interspersed
157 direct repeats) or LCTR (large cluster of tandem repeats) (25). In the
158 hyperthermophilic archaea *Pyrococcus abyssi* and *P. horikoshi* two sets of
159 “LCTR” sequences were located symmetrically on each side of the replication
160 origin, again suggesting a possible role in chromosome partitioning. However,
161 they were more numerous and scrambled in the genome of *P. furiosus*, casting
162 doubt on this interpretation (26).

163 Mojica *et al.* were the first to realize that all these bacterial and archaeal
164 sequences were functionally related (27). The term CRISPR, for clustered
165 regularly interspaced short palindromic repeats, was proposed by Jansen *et al* in
166 2002 (7) and became generally accepted by the community working on these
167 sequences, which precluded further confusion caused by many different names
168 for the related repeat sequences. Comparative genomics studies illuminated the
169 common characteristics of the CRISPR, namely that i) they are located in
170 intergenic regions; ii) contain multiple short direct repeats with very little
171 sequence variation; iii) the repeats are interspersed with non-conserved
172 sequences; iv) a common leader sequence of several hundred base pairs is
173 located on one side of the repeat cluster.

174 The fact that these mysterious sequences were conserved in two different
175 domains of life pointed to a more general role of these sequences. CRISPR
176 sequences were found in nearly all archaeal genomes and in about half of

177 bacterial genomes, rendering them the most widely distributed family of
178 repeated sequences in prokaryotes. As of today, CRISPR sequences have not
179 been found in any eukaryotic genome.

180

181 IDENTIFICATION OF THE CAS GENES

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

196 At about the same time, Kira Makarova, Eugene Koonin and colleagues
197 independently and systematically analyzed the conserved gene contexts in all
198 prokaryotic genomes available at the time and found several clusters of genes

199 corresponding to *cas* genes (encoding putative DNA polymerase, helicase and
200 RecB-like nuclease) in the genomes of hyperthermophilic archaea and in the two
201 hyperthermophilic bacteria with available genome sequences, *Aquifex* and
202 *Thermotoga* (8). These conserved genes were not found at that time in
203 mesophilic and moderate thermophilic archaea and bacteria. Based on this
204 observation, it was predicted that these proteins could be part of a “mysterious”
205 uncharacterized DNA repair system specific to thermophilic organisms.

206

207 **THE DISCOVERY OF CRISPR FUNCTION**

208 In the beginning of the genomic era, most of the archaeal genome
209 sequences were those of thermophilic and hyperthermophilic organisms.
210 Furthermore, thermophilic archaea, in addition to the hyperthermophilic bacteria,
211 such as *A. aeolicus* and *T. maritima*, have more and larger CRISPRs than
212 mesophilic organisms (7). These observations first suggested that the function of
213 CRISPR may be related to adaptation of organisms to high temperatures.
214 However, with more and more sequences becoming available, it turned out that
215 this correlation was not robust and that many mesophilic organisms also
216 contained CRISPR sequences. The *Eureka!* moment came when Francisco
217 Mojica in Alicante and Christine Pourcel in Orsay noticed independently that the
218 spacer regions between the repeat sequences are homologous to sequences of
219 bacteriophages, prophages and plasmids (3, 4). Importantly, based on the
220 literature review, they pointed out that the phages and plasmids do not infect host

221 strains harboring the homologous spacer sequences in the CRISPR. From these
222 observations, they independently proposed that CRISPR sequences function in
223 the framework of a biological defense system similar to the eukaryotic RNAi
224 system to protect the cells from the entry of these foreign mobile genetic
225 elements. The two groups also suggested that the CRISPRs can somehow
226 trigger the capture of pieces of foreign invading DNA to constitute a memory of
227 past genetic aggressions (3, 4). In a third influential paper of the same year,
228 Bolotin and colleagues confirmed these observations, further noticing a
229 correlation between the number of spacers of phage origin and the degree of
230 resistance to phage infection and suggested that CRISPR could be used to
231 produce antisense RNA (5) (for a brief historical account, see Morange, 2015)
232 (9).

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

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

260 The function of the CRISPR-Cas system as a prokaryotic acquired
261 immune system was finally experimentally proven in 2007, using the lactic acid
262 bacterium, *S. thermophilus* in 2007 (28). Insertion of the phage sequence into the
263 spacer region of the CRISPR of *S. thermophilus* made this strain resistant to the
264 corresponding phage. On the other hand, this bacterial resistance to the phage

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

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

302

303 **DIVERSITY AND CLASSIFICATION OF CRISPR-CAS**

304 It is striking that closely related strains can vary considerably in their
305 CRISPR content and distribution. For example, in *Mycobacterium* genus,
306 CRISPR exists in *M. tuberculosis*, but not in *M. leprae*. On the other hand,
307 phylogenetically distant *E. coli* and *M. avium* as well as *Methanothermobacter*
308 *thermautotrophicus* and *Archaeoglobus fulgidus* carry nearly identical CRISPR

309 repeat sequences (7). The number of CRISPR arrays in one genome varies from
310 1 to 18, and the number of repeat units in one CRISPR array varies from 2 to 374
311 (43). Based on the CRISPR database (<http://crispr.u-psud.fr/crispr/>), as of May
312 2017, CRISPRs were identified in the whole genome sequences of 202 (87%) out
313 of 232 analyzed archaeal species and 3059 (45%) of 6782 bacterial species.
314 Interestingly, a survey of 1,724 draft genomes suggested that CRISPR-Cas
315 systems are much less prevalent in environmental microbial communities (10.4%
316 in bacteria and 10.1% in archaea). This large difference between the prevalence
317 estimated from complete genomes of cultivated microbes compared to that of the
318 uncultivated ones was attributed to the lack of CRISPR-Cas systems across
319 major bacterial lineages that have no cultivated representatives (44).

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

328 Each class currently includes three types, namely, types I, III, and IV in
329 class 1, and types II, V, and VI in class 2. Types I, II, and III are readily
330 distinguishable by virtue of the presence of unique signature proteins: Cas3 for

331 type I, Cas9 for type II and Cas10 for type III. The multimeric effector complexes
332 of type I and type III systems, known as the CRISPR-associated complex for
333 antiviral defense (Cascade) and the Csm/Cmr complexes, respectively, are
334 architecturally similar and evolutionarily related (47-52). Unlike all other known
335 CRISPR-Cas systems, the functionally uncharacterized Type IV systems do not
336 contain the adaptation module consisting of nucleases Cas1 and Cas2 (47, 53).
337 Notably, the effector modules of subtype III-B systems are known to utilize
338 spacers produced by Type I systems, testifying to the modularity of the
339 CRISPR-Cas systems (54). Although many of the genomes encoding Type IV
340 systems do not carry identifiable CRISPR loci, it is not excluded that Type IV
341 systems, similar to subtype III-B systems, use crRNAs from different CRISPR
342 arrays once these become available (53).

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

347

348 **CLASS 2 SYSTEMS ARE SUITABLE FOR GENOME EDITING** 349 **TECHNOLOGY**

350 The simple architecture of the effector complexes has made class 2
351 CRISPR–Cas systems an attractive choice for developing a new generation of
352 genome-editing technologies (Fig. 6). Several distinct class 2 effectors have

353 been reported, including Cas9 in type II, Cas12a (formerly Cpf1), Cas12b (C2c1)
354 in Type V, and Cas13a (C2c2) and Cas13b (C2c3) in Type VI (45, 51). The most
355 common and best studied multidomain effector protein is Cas9, a
356 crRNA-dependent endonuclease, consisting of two unrelated nuclease domains,
357 RuvC and HNH, which are responsible for cleavage of the displaced (non-target)
358 and target DNA strands, respectively, in the crRNA–target DNA complex. Type II
359 CRISPR–cas loci also encode a *trans*-activating crRNA (tracrRNA) which might
360 have evolved from the corresponding CRISPR. The tracrRNA molecule is also
361 essential for pre-crRNA processing and target recognition in the type II systems.
362 The molecular mechanism of the target DNA cleavage by Cas9-crRNA complex,
363 schematically shown in Fig. 7, has been elucidated at the atomic level by the
364 crystal structure analysis of the DNA-Cas9-crRNA complex (55).

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

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

390

391 **ORIGINS OF CRISPR-CAS**

392 Analysis of clusters of poorly characterized, narrowly spread fast-evolving
393 genes in archaeal genomes, denoted as 'dark matter islands' (62), revealed
394 several islands encoding Cas1 proteins not associated with CRISPR loci
395 (Cas1-solo) (63). Comprehensive interrogation of the dark matter islands
396 revealed that *cas1*-solo genes are always located in vicinity of genes encoding

397 family B DNA polymerases and several other conserved genes (64).
398 Furthermore, these gene ensembles were found to be surrounded by long
399 inverted repeats and further flanked by shorter direct repeats, which respectively
400 resembled terminal inverted repeats (TIR) and target site duplications (TSD)
401 characteristic of various transposable elements. However, none of the identified
402 Cas1-solo-encoding genomic loci carried genes for known transposases or
403 integrases. Thus, it was hypothesized that Cas1 is the principal enzyme
404 responsible for the mobility of these novel genetic elements, which were
405 accordingly named 'Casposons' (64). Casposons were found to be widespread
406 in the genomes of methanogenic archaea as well as in thaumarchaea, but also
407 present in different groups of bacteria. Strong evidence of recent casposon
408 mobility was obtained by comparative genomic analysis of more than 60 strains
409 of the archaeon *Methanosarcina mazei*, in which casposons are variably
410 inserted in several distinct sites indicative of multiple, recent gains, and losses
411 (65). Based on the gene content, taxonomic distribution and phylogeny of the
412 Cas1 proteins, casposons are currently classified into 4 families (66).

413 Biochemical characterization of the casposon Cas1 ('casposase')
414 encoded in the genome of a thermophilic archaeon *Aciduliprofundum boonei*
415 has confirmed the predicted integrase activity (67, 68). Integration showed
416 strong target site preference and resulted in the duplication of the target site
417 regenerating the TSD observed in the *A. boonei* genome (68). The latter feature
418 resembles the duplication of the leader sequence-proximal CRISPR unit upon

419 integration of a protospacer catalyzed by the Cas1-Cas2 adaptation machinery
420 of CRISPR-Cas (69, 70). Remarkably, the sequence features of the casposon
421 target site are functionally similar to those required for directional insertion of
422 new protospacers into CRISPR arrays. In both systems, the functional target site
423 consists of two components: (i) a sequence which gets duplicated upon
424 integration of the incoming DNA duplex (i.e. the TSD segment in the case of
425 casposon and a CRISPR unit during protospacer integration) and (ii) the
426 upstream region which further determines the exact location of the integration
427 (i.e. the leader sequence located upstream of the CRISPR array and the
428 TSD-proximal segment in *A. boonei* genome) (68).

429 Collectively, the comparative genomics and experimental results
430 reinforced the mechanistic similarities and evolutionary connection between the
431 casposons and the adaptation module of the prokaryotic adaptive immunity
432 system, culminating in an evolutionary scenario for the origin of the
433 CRISPR-Cas systems. It has been proposed that casposon insertion near a
434 'solo-effector' innate immunity locus, followed by the immobilization of the
435 ancestral casposon via inactivation of the TIRs, gave rise to the adaptation and
436 effector modules, respectively, whereas the CRISPR repeats and the leader
437 sequence evolved directly from the preexisting casposon target site (71, 72). An
438 outstanding question in the above scenario is the switch in substrate specificity
439 of the ancestral casposase from integration of defined casposon TIRs to
440 insertion of essentially random, short (compared to casposon length)

441 protospacer sequences. It has been suggested that coupling between Cas1 and
442 Cas2 has been critical for this evolutionary transition (72).

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

448

449 **APPLICATION OF CRISPR-CAS TOOLS TO BACTERIA AND ARCHAEA**

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

463 (73-77). For instance, CRISPR-Cas-mediated genome editing technique
464 coupled with “heterologous recombineering” using linear single-stranded (SSDR
465 for single-stranded DNA recombineering) or double-stranded DNA (DSDR for
466 double-stranded DNA recombineering) templates, have been developed and
467 successfully applied in *E. coli* (78). In Archaea, gene silencing has been
468 established in *Sulfolobus solfataricus*, *S. islandicus* and *Haloferax volcanii* using
469 the endogenous CRISPR-Cas systems (reviewed in 77, 79). More recently,
470 Nayak and Metcalf have harnessed a bacterial Cas9 protein for genome editing
471 in the mesophilic archaeon *Methanosarcina acetovorans* (80). Hopfully a
472 thermophilic counterpart of the CRISPR-Cas9 system (or other class 2 systems)
473 will finally be established to perform genome editing in hyperthermophilic
474 species, which are difficult to manipulate genetically. From that perspective, the
475 diversity of CRISPR-Cas systems and mobile genetic elements, which remain to
476 be fully explored, is a treasure trove for future exploitation.

477

478 **APPLICATION OF CRISPR-CAS9 FOR PURPOSES OTHER THAN GENOME** 479 **EDITING**

480 The CRISPR loci are encoded by many bacterial and archaeal organisms
481 and are remarkably diverse, and thus they have been used as genetic markers
482 for species identification and typing, even before the elucidation of the actual
483 function of the CRISPR-Cas, as described above. For example, typing of
484 *Mycobacterium tuberculosis* is useful for diagnostic and epidemiological

485 purposes (20, 81). Typing by using CRISPR has been applied to *Yersinia pestis*
486 (4, 82), *Salmonella* (83, 84), and *Corynebacterium diphtheriae* (85).
487 CRISPR-Cas9 can be used as an antimicrobial agent by cleaving the genomes of
488 pathogenic bacteria, as an antibiotic agent with a novel mechanism of action. It is
489 expected to be a valuable remedy for the control of antibiotic-resistant bacteria.
490 For example, antibiotic-resistant bacteria, such as *Staphylococcus*, infecting the
491 skin of mice were selectively killed using CRISPR-Cas9 (86). CRISPR-Cas9 also
492 reportedly prevented intestinal infection by pathogenic *E. coli* (87). Although
493 there are technical challenges, such as delivery methods, which must be
494 overcome before CRISPR-Cas can be used as a safe therapeutic agent, active
495 research in this direction is ongoing and is expected to yield solutions in the near
496 future. Furthermore, imparting phage resistance in specific strains by the
497 CRISPR-Cas system is extremely useful for protecting various beneficial bacteria
498 in the fermented food industry from phage infection during the production
499 process.

500 Since the HNH nuclease domain and the RuvC nuclease domain are
501 responsible for the DNA cleavage activity of Cas9, Cas9 mutants devoid of
502 cleavage activity (dCas9) were obtained by replacing the amino acids within each
503 active center. The dCas9 protein is a useful tool for molecular biology
504 experiments to regulate gene expression. CRISPR-dCas9 binds to the target
505 DNA sequence, but cannot cleave it. This activity of CRISPR-dCas9 is applicable
506 to the labeling of a specific position, by fusing green fluorescent protein (GFP) to

507 dCas9, which binds to the target sequence depending on the sgRNA sequence
508 (88). In addition to this live intracellular site-specific labeling, gene expression
509 can be artificially controlled by linking dCas9 to either the promoter region or the
510 open reading frame of a gene (89-91). dCas9 can also be fused with a
511 transcription activator or the ω subunit of bacterial RNA polymerase. However, it
512 seems not to be as easy as compared with suppression, although ingenious
513 attempts have been made to promote transcription by designing a guide
514 sequence that ensures binding of dCas9 to a specific promoter.

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

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

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

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

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

546

547 **CONCLUDING REMARKS**

548 Only 30 years have passed since one of the authors of this review
549 discovered unique repeated sequence in the *E. coli* genome at the onset of his

550 post-doc career. It was impossible to predict the possible function of this
551 enigmatic sequence at the time; however, genomic revolution in the mid-90's,
552 coupled with development of powerful bioinformatics tools eventually enabled
553 elucidation of the CRISPR functions. CRISPR arrays and Cas proteins, broadly
554 distributed in the genomes of prokaryotes, especially in Archaea, are now known
555 to constitute the highly efficient acquired immunity system. Although discovery of
556 the CRISPR-Cas by itself was a great feat of fundamental biology, it also led to
557 the development of next-generation tools for genetic engineering. The
558 development of the genome editing technology by CRISPR-Cas9 reminds of the
559 times when the PCR was born.

560 When *in vitro* genetic engineering techniques using restriction
561 endonucleases and nucleic acid modifying enzymes were established, it was still
562 often a complex task to clone a single gene (as in the case of the *iap* gene).
563 However, this difficulty was alleviated by the invention of PCR using a
564 thermostable DNA polymerase that profoundly boosted the application of genetic
565 engineering techniques in all biological laboratories worldwide. The discovery of
566 a thermostable DNA polymerase was critical for the “PCR revolution” because it
567 enabled the design of a PCR apparatus for practical use. Similarly, in the case of
568 genome editing, the CRISPR revolution was made possible by identifying the
569 right enzymatic system (Cas9) that could simplify the methodology to exploit the
570 potential of the CRISPR-Cas system. The curiosity of a mysterious repetitive

571 sequence and a sustained inquiry mind for elucidating its function brought grand
572 discoveries.

573

574 **ACKNOWLEDGEMENTS**

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

582

583 REFERENCES

- 584 1. Ishino Y, Shinagawa H, Makino K, Amemura M, Nakata A. 1987.
585 Nucleotide sequence of the *iap* gene, responsible for alkaline phosphatase
586 isozyme conversion in *Escherichia coli*, and identification of the gene
587 product, J Bacteriol 169: 5429–5433.
- 588 2. Mojica MJ, Juez G, Rodríguez-Valera F. 1993. Transcription at different
589 salinities of *Haloferax mediterranei* sequences adjacent to partially modified
590 PstI sites. Mol Microbiol 9: 613–621.
- 591 3. Mojica FJM, Díez-Villaseñor C, García-Martínez J, Soria E. 2005.
592 Intervening sequences of regularly spaced prokaryotic repeats derive from
593 foreign genetic elements. J Mol Evol 60: 174–182.
- 594 4. Pourcel C, Salvignol G, Vergnaud G. 2005. CRISPR elements in *Yersinia*
595 *pestis* acquire new repeats by preferential uptake of bacteriophage DNA,
596 and provide additional tools for evolutionary studies. Microbiology 151: 653–
597 663.
- 598 5. Bolotin A, Quinquis B, Sorokin A, Ehrlich SD. 2005. Clustered regularly
599 interspaced short palindrome repeats (CRISPRs) have spacers of
600 extrachromosomal origin. Microbiology 151: 2551–2561.
- 601 6. Makarova KS, Aravind L, Grishin NV, Rogozin IB, Koonin EV, A (2002)
602 DNA repair system specific for thermophilic Archaea and bacteria predicted
603 by genomic context analysis. Nucleic Acids Res 30: 482–496.

- 604 7. Jansen R, Embden JD, Gaastra W, Schouls LW. 2002. Identification of
605 genes that are associated with DNA repeats in prokaryotes. *Mol Microbiol*
606 43: 1565–1575.
- 607 8. Makarova KS, Grishin NV, Shabalina SA, Wolf YI, Koonin EV. 2006. A
608 putative RNA-interference-based immune system in prokaryotes:
609 computational analysis of the predicted enzymatic machinery, functional
610 analogies with eukaryotic RNAi, and hypothetical mechanisms of action.
611 *Biol Direct* 1: 7.
- 612 9. Morange M. 2015. What history tells us XXXVII. CRISPR-Cas: The
613 discovery of an immune system in prokaryotes. *J Biosci* 40: 221-223
- 614 10. Morange M. 2015. What history tells us XXXIX. CRISPR-Cas: From a
615 prokaryotic immune system to a universal genome editing tool. *J Biosci* 40:
616 829-832.
- 617 11. Lander ES. 2016. The Heroes of CRISPR. *Cell* 164: 18–28.
- 618 12. Mojica FJ, Rodriguez-Valera F. 2016. The discovery of CRISPR in archaea
619 and bacteria, *FEBS J* 283: 3162–3169.
- 620 13. Barrangou R, Horvath PA. 2017. A decade of discovery: CRISPR functions
621 and applications. *Nat Microbiol* 2: 17092.
- 622 14. Han W, She Q. 2017. CRISPR history: discovery, characterization, and
623 prosperity. *Prog Mol Biol Transl Sci* 152:1-21.
- 624 15. Nakata A, Shinagawa H, Amemura M. 1982. Cloning of alkaline
625 phosphatase isozyme gene (*iap*) of *Escherichia coli*. *Gene* 19: 313–319.

- 626 16. Stern MJ, Ames GF-L, Smith NH, Robinson EC, Higgins CF. 1984.
627 Repetitive extragenic palindromic sequences: a major component of the
628 bacterial genome, *Cell* 37: 1015–1026.
- 629 17. Newburg SF, Smith NH, Robinson EC, Hiles IE, Higgins CF. 1987.
630 Stabilization of translationally active mRNA by prokaryotic REP sequences,
631 *Cell* 48: 297–310.
- 632 18. Nakata A, Amemura M, Makino K. 1989. Unusual nucleotide arrangement
633 with repeated sequences in the *Escherichia coli* K-12 chromosome, *J*
634 *Bacteriol* 171: 3553–3556.
- 635 19. Hermans PW, van Soolingen D, Bik EM, de Haas PE, Dale JW, van Embden
636 JD. 1991. Insertion element IS987 from *Mycobacterium bovis* BCG is
637 located in a hot-spot integration region for insertion elements in
638 *Mycobacterium tuberculosis* complex strains. *Infect Immun* 59: 2695–2705.
- 639 20. Groenen PM, Bunschoten AE, van Soolingen D, van Embden JD. 1993.
640 Nature of DNA polymorphism in the direct repeat cluster of *Mycobacterium*
641 *tuberculosis*; application for strain differentiation by a novel typing method.
642 *Mol Microbiol* 10: 1057–1065.
- 643 21. Mojica FJ, Ferrer C, Juez G, Rodríguez-Valera F. 1995. Long stretches of
644 short tandem repeats are present in the largest replicons of the Archaea
645 *Haloferax mediterranei* and *Haloferax volcanii* and could be involved in
646 replicon partitioning, *Mol Microbiol* 17: 85–93.

- 647 22. Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage,
648 AR Bult CJ, Tomb JF, Dougherty BA, Merrick JM, McKenney K, Sutton G,
649 FitzHugh W, Fields C, Gocayne JD, Scott J, Shirley R, Liu LI, Glodek A,
650 Kelley JM, Weidman JF, Phillips CA, Spriggs T, Hedblom E, Cotton MD,
651 Utterback TR, Hanna MC, Nguyen DT, Saudek DM, Brandon RC, Fine LD,
652 Fritchman JL, Fuhrmann JL, Geoghagen NSM, Gnehm CL, McDonald LA,
653 Small KV, Fraser CM, Smith HO, Venter JC. 1995. Whole-genome random
654 sequencing and assembly of *Haemophilus influenzae* Rd, Science 269 :
655 496–512.
- 656 23. Bult CJ, White O, Olsen GJ, Zhou L, Fleischmann RD, Sutton GG, Blake JA,
657 FitzGerald LM, Clayton RA, Gocayne JD, Kerlavage AR, Dougherty BA,
658 Tomb JF, Adams MD, Reich CI, Overbeek R, Kirkness EF, Weinstock KG,
659 Merrick JM, Glodek A, Scott JL, Geoghagen NS, Venter JC. 1996. Complete
660 genome sequence of the methanogenic archaeon, *Methanococcus*
661 *jannaschii*. Science 273: 1058–1073.
- 662 24. Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert
663 F, Hoheisel JD, Jacq C, Johnston M, Louis EJ, Mewes HW, Murakami Y,
664 Philippsen P, Tettelin H, Oliver SG. 1996. Life with 6000 genes. Science
665 274: 546–567.
- 666 25. She Q, Singh RK, Confalonieri F, Zivanovic Y, Allard G, Awayez MJ,
667 Chan-Weiher CC, Clausen IG, Curtis BA, De Moors A, Erauso G, Fletcher C,
668 Gordon PM, Heikamp-de Jong I, Jeffries AC, Kozera CJ, Medina N, Peng X,

- 669 Thi-Ngoc HP, Redder P, Schenk ME, Theriault C, Tolstrup N, Charlebois RL,
670 Doolittle WF, Duguet M, Gaasterland T, Garrett RA, Ragan MA, Sensen CW,
671 Van der Oost J. 2001. The complete genome of the crenarchaeon
672 *Sulfolobus solfataricus* P2. *Proc Natl Acad Sci USA*. 98: 7835-7840.
- 673 26. Zivanovic Y, Lopez P, Philippe H, Forterre P. 2002. *Pyrococcus* genome
674 comparison evidences chromosome shuffling-driven evolution. *Nucleic
675 Acids Res* 30: 1902-1910.
- 676 27. Mojica FJ, Díez-Villaseñor C, Soria E, Juez G. 2000. Biological significance
677 of a family of regularly spaced repeats in the genomes of Archaea, Bacteria
678 and mitochondria. *Mol Microbiol* 36: 244–246.
- 679 28. Barrangou R, Fremaux C, Deveau H, Richards M, Boyaval P, Moineau S,
680 Romero DA, Horvath P. 2007. CRISPR provides acquired resistance against
681 viruses in prokaryotes. *Science* 315: 1709–1712.
- 682 29. Marraffini LA, Sontheimer EJ. 2008. CRISPR interference limits horizontal
683 gene transfer in *Staphylococci* by targeting DNA. *Science* 322: 1843–1845.
- 684 30. Brouns SJ, Jore MM, Lundgren M, Westra ER, Slijkhuis RJ, Snijders AP,
685 Dickman MJ, Makarova KS, Koonin EV, van der Oost J. Small CRISPR
686 RNAs guide antiviral defense in prokaryotes. 2008. *Science* 321: 960-964.
- 687 31. Andersson AF, Banfield JF. Virus population dynamics and acquired virus
688 resistance in natural microbial communities. 2008. *Science*. 320:1047-1050.

- 689 32. Saprunauskas R, Gasiunas G, Fremaux C, Barrangou R, Horvath P,
690 Siksnyš V. 2011. The *Streptococcus thermophilus* CRISPR/Cas system
691 provides immunity in *Escherichia coli*, *Nucleic Acids Res* 39: 9275–9282.
- 692 33. Gasiunas G, Barrangou R, Horvath P, Siksnyš V. 2012. Cas9-crRNA
693 ribonucleoprotein complex mediates specific DNA cleavage for adaptive
694 immunity in bacteria, *Proc Natl Acad Sci USA* 109: E2579–2586.
- 695 34. Jinek M, Chylinski K, Fonfara I, Hauer M, Doudna JA, Charpentier EA. 2012.
696 A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial
697 immunity. *Science* 337: 816–821.
- 698 35. Cong L, Ran FA, Cox D, Lin S, Barretto R, Habib N, Hsu PD, Wu X, Jiang W,
699 Marraffini LA, Zhang F. 2013. Multiplex genome engineering using
700 CRISPR/Cas systems. *Science* 339: 819–823.
- 701 36. Mali P, Yang L, Esvelt KM, Aach J, Guell M, DiCarlo JE, Norville JE, Church
702 GM. 2013. RNA-guided human genome engineering via Cas9. *Science* 339:
703 823–826.
- 704 37. Horvath P, Barrangou R. (2010) CRISPR/Cas, the immune system of
705 bacteria and archaea. *Science* 327: 167–170.
- 706 38. Wiedenheft B, Sternberg SH, Doudna JA. 2012. RNA-guided genetic
707 silencing systems in bacteria and archaea. *Nature* 482: 331–338.
- 708 39. Jackson SA, McKenzie RE, Fagerlund RD, Kieper SN, Fineran PC, Brouns
709 SJ. CRISPR-Cas: Adapting to change. *Science* 356 eaal5056.

- 710 40. Tamulaitis G, Venclovas C, Siksnys V. 2017. Type III CRISPR-Cas
711 Immunity: Major differences brushed aside, *Trends Microbiol* 25: 49-61.
- 712 41. Mohanraju P, Makarova KS, Zetsche B, Zhang F, Koonin EV, van der Oost J.
713 2016. Diverse evolutionary roots and mechanistic variations of the
714 CRISPR-Cas systems. *Science* 353: aad5147.
- 715 42. Charpentier E, Richter H, van der Oost J, White MF. 2015. Biogenesis
716 pathways of RNA guides in archaeal and bacterial CRISPR-Cas adaptive
717 immunity. *FEMS Microbiol Rev* 39: 428-441.
- 718 43. Marraffini LA, Sontheimer EJ. 2010. CRISPR interference: RNA-directed
719 adaptive immunity in bacteria and archaea. *Nat Rev Genet.* 11:181-190.
- 720 44. Burstein D, Sun CL, Brown CT, Sharon I, Anantharaman K, Probst AJ,
721 Thomas BC, Banfield JF. 2016. Major bacterial lineages are essentially
722 devoid of CRISPR-Cas viral defence systems. *Nat Commun.* 7:10613.
- 723 45. Shmakov S, Smargon A, Scott D, Cox D, Pyzocha N, Yan W, Abudayyeh
724 OO, Gootenberg JS, Makarova KS, Wolf YI, Severinov K, Zhang F, Koonin
725 EV. 2017. Diversity and evolution of class 2 CRISPR-Cas systems. *Nat Rev*
726 *Microbiol* 15: 169-182.
- 727 46. Burstein D, Harrington LB, Strutt SC, Probst AJ, Anantharaman K, Thomas
728 BC, Doudna JA, Banfield JF, 2017. New CRISPR-Cas systems from
729 uncultivated microbes. *Nature* 542: 237-241.
- 730 47. Rouillon C, Zhou M, Zhang J, Politis A, Beilsten-Edmands V, Cannone G,
731 Graham S, Robinson CV, Spagnolo L, White MF. 2013. Structure of the

- 732 CRISPR interference complex CSM reveals key similarities with cascade.
733 Mol Cell 52: 124-134.
- 734 48. Jackson RN, Wiedenheft B. 2015. A conserved structural chassis for
735 mounting versatile CRISPR RNA-guided immune responses, Mol Cell 58:
736 722-728.
- 737 49. Makarova KS, Haft DH, Barrangou R, Brouns SJ, Charpentier E, Horvath P,
738 Moineau S, Mojica FJ, Terns RM, Terns MP, White MF, Yakunin AF, Garrett
739 RA, van der Oost J, Backofen R, Koonin EV. 2011. Evolution and
740 classification of the CRISPR-Cas systems. Nat Rev Microbiol 9: 467–477.
- 741 50. Koonin EV, Makarova KS. 2013. CRISPR-Cas: evolution of an RNA-based
742 adaptive immunity system in prokaryotes, RNA Biol 10: 679-686.
- 743 51. Koonin EV, Makarova KS, Zhang F. 2017. Diversity, classification and
744 evolution of CRISPR-Cas systems. Curr Opin Microbiol 37: 67-78.
- 745 52. Venclovas C. 2016. Structure of Csm2 elucidates the relationship between
746 small subunits of CRISPR-Cas effector complexes. FEBS Lett 590:
747 1521-1529.
- 748 53. Makarova KS, Wolf YI, Alkhnbashi OS, Costa F, Shah SA, Saunders SJ,
749 Barrangou R, Brouns SJ, Charpentier E, Haft DH, Horvath P, Moineau S,
750 Mojica FJ, Terns RM, Terns MP, White MF, Yakunin AF, Garrett RA, van der
751 Oost, Backofen R, Koonin EV. 2015. An updated evolutionary classification
752 of CRISPR-Cas systems. Nat Rev Microbiol 13: 722–736.

- 753 54. Garrett RA, Vestergaard G, Shah SA. 2011. Archaeal CRISPR-based
754 immune systems: exchangeable functional modules. *Trends Microbiol.*
755 19:549-556.
- 756 55. Nishimasu H, Ran FA, Hsu PD, Konermann S, Shehata SI, Dohmae N,
757 Ishitani R, Zhang F, Nureki O. 2014. Crystal structure of Cas9 in complex
758 with guide RNA and target DNA. *Cell* 156: 935–949.
- 759 56. Yamano T, Nishimasu H, Zetsche B, Hirano H, Slaymaker IM, Li Y,
760 Fedorova I, Nakane T, Makarova KS, Koonin EV, Ishitani R, Zhang F, Nureki
761 O. 2016. Crystal structure of Cpf1 in complex with guide RNA and target
762 DNA. *Cell* 165: 949–962.
- 763 57. Zetsche B, Gootenberg JS, Abudayyeh OO, Slaymaker IM, Makarova KS,
764 Essletzbichler P, Volz SE, Joung J, van der Oost J, Regev A, Koonin EV,
765 Zhang F. 2015. Cpf1 is a single RNA-guided endonuclease of a class 2
766 CRISPR-Cas system. *Cell* 163: 759–771.
- 767 58. Shmakov S, Abudayyeh OO, Makarova KS, Wolf YI, Gootenberg JS,
768 Semenova E, Minakhin L, Joung J, Konermann S, Severinov K, Zhang Z,
769 Koonin EV. 2015. Discovery and functional characterization of diverse class
770 2 CRISPR- Cas systems. *Mol Cell* 60: 385–397.
- 771 59. Jiang W, Samai P, Marraffini LA, 2016. Degradation of phage transcripts by
772 CRISPR-associated RNases enables type III CRISPR-Cas immunity. *Cell*
773 164: 710–721.

- 774 60. Abudayyeh OO, Gootenberg JS, Konermann S, Joung J, Slaymaker IM,
775 Cox DB, Shmakov S, Makarova KS, Semenova E, Minakhin L, Severinov K,
776 Regev A, Lander ES, Koonin EV, Zhang F, 2016. C2c2 is a
777 single-component programmable RNA-guided RNA-targeting CRISPR
778 effector, *Science* 353: 6299.
- 779 61. Murugan K, Babu K, Sundaresan R, Rajan R, Sashital DG. 2017. The
780 Revolution Continues: Newly Discovered Systems Expand the CRISPR-Cas
781 Toolkit. *Mol Cell* 68: 15-25.
- 782 62. Makarova KS, Wolf YI, Forterre P, Prangishvili D, Krupovic M, Koonin EV.
783 2014. Dark matter in archaeal genomes: a rich source of novel mobile
784 elements, defense systems and secretory complexes, *Extremophiles* 18:
785 877-893.
- 786 63. Makarova KS, Wolf YI, Koonin EV, 2013. The basic building blocks and
787 evolution of CRISPR-cas systems, *Biochem Soc Trans* 41: 1392–1400.
- 788 64. Krupovic M, Makarova KS, Forterre P, Prangishvili D, Koonin E.V. 2014.
789 Casposons: a new superfamily of self-synthesizing DNA transposons at the
790 origin of prokaryotic CRISPR-Cas immunity, *BMC Biol* 12: 36.
- 791 65. Krupovic M, Shmakov S, Makarova KS, Forterre P, Koonin EV. 2016.
792 Recent mobility of casposons, self-synthesizing transposons at the origin of
793 the CRISPR-Cas immunity. *Genome Biol Evol* 8: 375-386.

- 794 66. Krupovic M, Koonin EV. 2016. Self-synthesizing transposons: unexpected
795 key players in the evolution of viruses and defense systems. *Curr Opin*
796 *Microbiol* 31: 25-33.
- 797 67. Hickman AB, Dyda F. 2015. The casposon-encoded Cas1 protein from
798 *Aciduliprofundum boonei* is a DNA integrase that generates target site
799 duplications. *Nucleic Acids Res* 43: 10576-10587.
- 800 68. Béguin P, Charpin N, Koonin EV, Forterre P, Krupovic M. 2016. Casposon
801 integration shows strong target site preference and recapitulates
802 protospacer integration by CRISPR-Cas systems. *Nucleic Acids Res* 44:
803 10367-10376.
- 804 69. Yosef I, Goren MG, Qimron U. 2012. Proteins and DNA elements essential
805 for the CRISPR adaptation process in *Escherichia coli*. *Nucleic Acids Res.*
806 40: 5569-5576.
- 807 70. Rollie C, Schneider S, Brinkmann AS, Bolt EL, White MF. 2015. Intrinsic
808 sequence specificity of the Cas1 integrase directs new spacer acquisition,
809 *Elife* 4: e08716.
- 810 71. Koonin EV, Krupovic M. 2015. Evolution of adaptive immunity from
811 transposable elements combined with innate immune systems. *Nat Rev*
812 *Genet* 16: 184–192.
- 813 72. Krupovic M, Béguin P, Koonin EV. 2017. Casposons: mobile genetic
814 elements that gave rise to the CRISPR-Cas adaptation machinery. *Curr*
815 *Opin Microbiol* 38: 36-43.

- 816 73. Jiang W, Bikard D, Cox D, Zhang F, Marraffini LA. 2013. RNA-guided
817 editing of bacterial genomes using CRISPR-Cas systems. *Nat Biotechnol*
818 31: 233–239.
- 819 74. Jiang Y, Chen B, Duan C, Sun B, Yang J, Yang S. 2015. Multigene editing
820 in the *Escherichia coli* genome via the CRISPR-Cas9 system. *Appl Environ*
821 *Microbiol* 81: 2506–2514.
- 822 75. Oh JH, van Pijkeren JP. 2014 CRISPR-Cas9-assisted recombineering in
823 *Lactobacillus reuteri*. *Nucleic Acids Res* 42: e131.
- 824 76. Charpentier E, Marraffini LA. 2014. Harnessing CRISPR-Cas9 immunity for
825 genetic engineering. *Curr Opin Microbiol* 19: 114-119.
- 826 77. Gophna U, Allers T, Marchfelder A. 2017. Finally, archaea get their
827 CRISPR-Cas toolbox. *Trends Microbiol* 25: 430-432.
- 828 78. Mougiakos I, Bosma EF, de Vos WM, van Kranenburg R, van der Oost J.
829 2016. Next generation prokaryotic engineering: the CRISPR-Cas toolkit.
830 *Trends Biotechnol* 34: 575-587.
- 831 79. Peng N, Han W, Li Y, Liang Y, She Q. 2017. Genetic technologies for
832 extremely thermophilic microorganisms of *Sulfolobus*, the only genetically
833 tractable genus of crenarchaea. *Sci China Life Sci* 60: 370-385.
- 834 80. Nayak DD, Metcalf WW. 2017. Cas9-mediated genome editing in the
835 methanogenic archaeon *Methanosarcina acetivorans*. *Proc Natl Acad Sci*
836 *USA* 114: 2976-2981.

- 837 81. Kamerbeek J, Schouls L, Kolk A, van Agterveld M, van Soolingen D, Kuijper
838 S, Bunschoten A, Molhuizen H, Shaw R, Goyal M, van Embden J. 1997.
839 Simultaneous detection and strain differentiation of *Mycobacterium*
840 *tuberculosis* for diagnosis and epidemiology. J Clin Microbiol 35: 907-914.
- 841 82. Cui Y, Li Y, Gorgé O, Platonov ME, Yan Y, Guo Z, Pourcel C, Dentovskaya
842 SV, Balakhonov SV, Wang X, Song Y, Anisimov AP, Vergnaud G, Yang R.
843 2008. Insight into microevolution of *Yersinia pestis* by clustered regularly
844 interspaced short palindromic repeats. PLoS One 2008 3: e2652.
- 845 83. Liu F, Barrangou R, Gerner-Smidt P, Ribot EM, Knabel SJ, Dudley EG.
846 2011. Novel virulence gene and clustered regularly interspaced short
847 palindromic repeat (CRISPR) multilocus sequence typing scheme for
848 subtyping of the major serovars of *Salmonella enterica* subsp. *enterica*. Appl
849 Environ Microbiol 77,1946-1956.
- 850 84. Liu F, Kariyawasam S, Jayarao BM, Barrangou R, Gerner-Smidt P, Ribot
851 EM, Knabel SJ, Dudley EG. 2011. Subtyping *Salmonella enterica* serovar
852 *enteritidis* isolates from different sources by using sequence typing based on
853 virulence genes and clustered regularly interspaced short palindromic
854 repeats (CRISPRs). Appl Environ Microbiol 77, 4520-4526.
- 855 85. Mokrousov I, Narvskaya O, Limeschenko E, Vyazovaya A. 2005. Efficient
856 discrimination within a *Corynebacterium diphtheriae* epidemic clonal group
857 by a novel macroarray-based method. J Clin Microbiol 43:1662-1668.

- 858 86. Bikard D, Euler CW, Jiang W, Nussenzweig PM, Goldberg GW, Duportet X,
859 Fischetti VA, Marraffini LA. 2014. Exploiting CRISPR-Cas nucleases to
860 produce sequence-specific antimicrobials. *Nat Biotechnol* 32:1146-1150.
- 861 87. Citorik RJ, Mimee M, Lu TK. 2014. Sequence-specific antimicrobials using
862 efficiently delivered RNA-guided nucleases. *Nat Biotechnol* 32: 1141-1145.
- 863 88. Chen B, Gilbert LA, Cimini BA, Schnitzbauer J, Zhang W, Li GW, Park J,
864 Blackburn EH, Weissman JS, Qi LS, Huang B. 2013. Dynamic imaging of
865 genomic loci in living human cells by an optimized CRISPR/Cas system. *Cell*
866 155:1479-1491.
- 867 89. Bikard D, Jiang W, Samai P, Hochschild A, Zhang F, Marraffini LA. 2013.
868 Programmable repression and activation of bacterial gene expression using
869 an engineered CRISPR-Cas system. *Nucleic Acids Res* 41: 7429-7437.
- 870 90. Qi LS, Larson MH, Gilbert LA, Doudna JA, Weissman JS, Arkin AP, Lim WA.
871 2013. Repurposing CRISPR as an RNA-guided platform for
872 sequence-specific control of gene expression. *Cell* 152:1173-1183.
- 873 91. Gilbert LA, Larson MH, Morsut L, Liu Z, Brar GA, Torres SE, Stern-Ginossar
874 N, Brandman O, Whitehead EH, Doudna JA, Lim WA, Weissman JS, Qi LS.
875 2013. CRISPR-mediated modular RNA-guided regulation of transcription in
876 eukaryotes. *Cell* 154: 442-451.
- 877 92. Guillinger JP, Thompson DB, Liu DR. 2014. Fusion of catalytically inactive
878 Cas9 to FokI nuclease improves the specificity of genome modification. *Nat*
879 *Biotechnol* 32: 577-582.

- 880 93. Tsai SQ, Wyvekens N, Khayter C, Foden JA, Thapar V, Reyon D, Goodwin
881 MJ, Aryee MJ, Joung JK. 2014. Dimeric CRISPR RNA-guided FokI
882 nucleases for highly specific genome editing. *Nat Biotechnol* 32: 569-576.
- 883 94. Havlicek S, Shen Y, Alpagu Y, Bruntraeger MB, Zufir NB, Phuah ZY, Fu Z,
884 Dunn NR, Stanton LW. 2017. Re-engineered RNA-Guided FokI-Nucleases
885 for Improved Genome Editing in Human Cells. *Mol Ther* 25: 342-355.
- 886 95. Ran FA, Hsu PD, Lin CY, Gootenberg JS, Konermann S, Trevino AE, Scott
887 DA, Inoue A, Matoba S, Zhang Y, Zhang F. 2013. Double nicking by
888 RNA-guided CRISPR Cas9 for enhanced genome editing specificity. *Cell*
889 154:1380-1389.
- 890 96. Nishida K, Arazoe T, Yachie N, Banno S, Kakimoto M, Tabata M, Mochizuki
891 M, Miyabe A, Araki M, Hara KY, Shimatani Z, Kondo A. 2016. Targeted
892 nucleotide editing using hybrid prokaryotic and vertebrate adaptive immune
893 systems *Science*. 353: 6305.
- 894 97. Zetsche B, Volz SE, Zhang F. 2015. A split-Cas9 architecture for inducible
895 genome editing and transcription modulation. *Nat Biotechnol* 33:139-142.
- 896 98. Wright AV, Sternberg SH, Taylor DW, Staahl BT, Bardales JA, Kornfeld JE,
897 Doudna JA. 2015. Rational design of a split-Cas9 enzyme complex. *Proc*
898 *Natl Acad Sci USA*. 112: 2984-2989.
- 899 99. Nihongaki Y, Kawano F, Nakajima T, Sato M. 2015. Photoactivatable
900 CRISPR-Cas9 for optogenetic genome editing. *Nat Biotechnol* 33, 755-760.

901 **Figure legends**

902

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

908

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

915

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

920

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

928

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

939

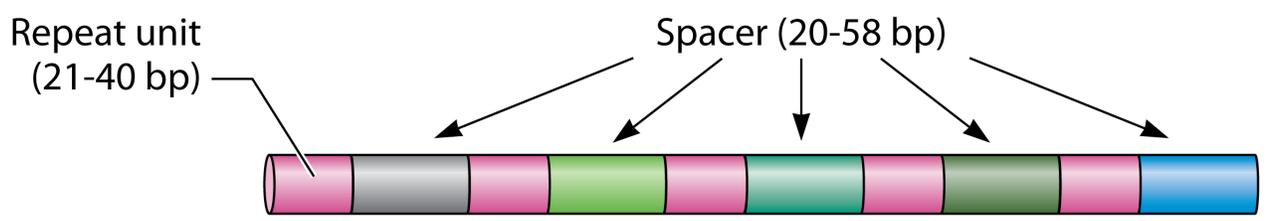
940 **FIG 6** Most recent classification of CRISPR-Cas immune systems. A. Based on
941 the detailed sequence analyses and gene organization of the Cas proteins,
942 CRISPR-Cas was classified into two major classes depending on whether the
943 effector is a complex composed of multiple Cas proteins or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

951

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

953 The Cas9-crRNA-tracrRNA complex binds to foreign DNA containing PAM,
954 where Cas9 binds and starts to unwind the double-strand of the foreign DNA to
955 induce duplex formation of crRNA and foreign DNA. Cas9 consists of two regions,
956 called REC (recognition) lobe and NUC (nuclease) lobe. REC lobe is responsible
957 for the nucleic acid recognition. NUC lobe contains the HNH and RuvC nuclease
958 domains, and a C-terminal region containing PAM-interacting (PI) domain. The
959 HNH domain and the RuvC domain cleave the DNA strand forming duplex with

960 crRNA and the other DNA strand, respectively, so that double-strand break
961 occurs in the target DNA.



E. coli (1987)

CGGTTTATCCCCGCTGCGCGGGGAACTC

H. mediterranei (1993)

GTTACAGACGAACCCTAGTTGGGTGAAGC

