

CRISPR–Cas adaptation: insights into the mechanism of action

Gil Amitai and Rotem Sorek

Abstract | Since the first demonstration that CRISPR–Cas systems provide bacteria and archaea with adaptive immunity against phages and plasmids, numerous studies have yielded key insights into the molecular mechanisms governing how these systems attack and degrade foreign DNA. However, the molecular mechanisms underlying the adaptation stage, in which new immunological memory is formed, have until recently represented a major unresolved question. In this Progress article, we discuss recent discoveries that have shown both how foreign DNA is identified by the CRISPR–Cas adaptation machinery and the molecular basis for its integration into the chromosome to form an immunological memory. Furthermore, we describe the roles of each of the specific CRISPR–Cas components that are involved in memory formation, and consider current models for their evolutionary origin.

Organisms from all domains of life are engaged in a constant conflict with viruses and other infecting pathogens. As part of the perpetual arms race between pathogens and their hosts, two strategic approaches of immunity have evolved: innate immunity and adaptive immunity. Whereas innate immunity involves pre-existing, genetically encoded systems that recognize general features of pathogens, adaptive immunity involves systems that are capable of generating specific immune responses and immunological memory against previously unencountered invaders.

Until a decade ago, adaptive immunity was considered to be a feature found only in eukaryotes. However, the discovery of CRISPR and the CRISPR-associated (Cas) proteins^{1,2} led to the hypothesis, formulated in 2005 (REFS 3–5), and then the demonstration, in 2007 (REF. 6), that many bacteria and archaea also possess a functionally complex adaptive immune system. CRISPR–Cas systems are found in approximately 40% of all sequenced bacterial genomes and the vast majority of sequenced archaeal genomes⁷. In these systems, the CRISPR array stores the immunological memory in the form of ‘spacers’ (REFS 3–5) — short DNA sequences originating from invading pathogens — that

are interleaved with the CRISPR DNA repeats (FIG. 1a). The CRISPR array is preceded by a regulatory leader sequence^{8–10}. Finally, a set of Cas proteins is responsible for generating and executing the adaptive immune function of CRISPR–Cas systems¹¹.

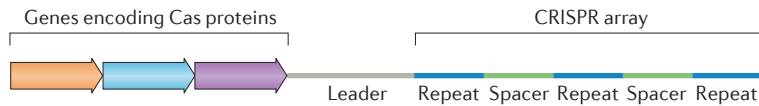
The mechanism of action of CRISPR–Cas systems can be divided into three stages: adaptation, expression and maturation, and interference (FIG. 1b–d). In the adaptation stage, Cas proteins identify the target DNA and acquire a new spacer sequence derived from this target. It is this spacer sequence that is integrated into the CRISPR array to form the immunological memory (FIG. 1b). During the expression and maturation stage, the CRISPR array is transcribed into a precursor RNA transcript that is further processed into smaller units of RNA known as CRISPR RNAs (crRNAs), each containing a single spacer flanked by a part of the repeat sequence. These crRNAs are combined with one or more Cas proteins to form the active Cas–crRNA complex (FIG. 1c). In the interference stage, the Cas–crRNA complex scans the cell for foreign nucleic acid targets that are recognized by base-pairing with complementary crRNA sequences. Successful recognition leads to the cleavage and degradation of the target nucleic acid (FIG. 1d).

Three major types of CRISPR–Cas system have been identified, and these can be further divided into several subtypes that encompass considerable structural and functional diversity (recently reviewed in REF. 11). In type I and type III systems (the two major types of what are now denoted class 1 CRISPR–Cas systems)¹¹, both the expression and maturation stage and the interference stage are executed by a multisubunit protein complex (the CRISPR-associated complex for antiviral defence (Cascade) complex (assisted by the Cas3 helicase)¹², the Csm complex¹³ or the Cmr complex¹⁴ for type I, type III-A and type III-B CRISPR–Cas systems, respectively), whereas in type II, type V and type VI systems (belonging to class 2 CRISPR–Cas systems)^{11,15}, these processes are carried out by a single large polypeptide, such as Cas9 in type II systems and Cpf1 or related proteins in type V systems^{15–18}. In addition, Cas9 requires association with a structural non-coding RNA known as transactivating CRISPR RNA (tracrRNA)¹⁹ for its activity.

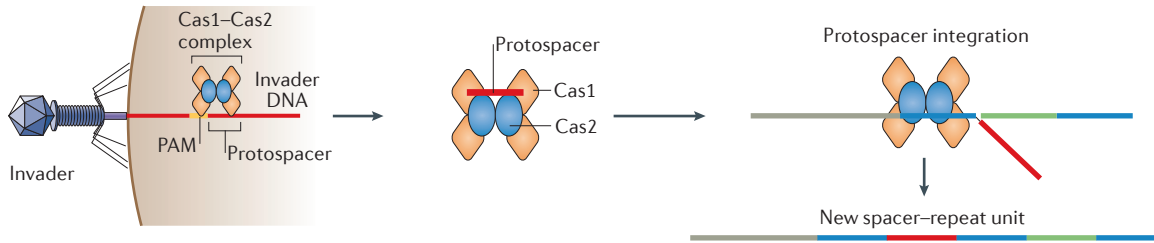
In the past few years, the expression and maturation stage and the interference stage of several CRISPR–Cas systems have been studied in depth, leading to a near comprehensive mechanistic understanding of these processes (reviewed in REFS 20–23). By contrast, many of the molecular details of the adaptation stage are still being investigated, although research in this field has recently benefited from the development and application of new experimental approaches (BOX 1). In this Progress article, we discuss the recent advances that have shed light on how new immunological memory is formed and stored during the adaptation stage. Specifically, we focus on the roles of the individual CRISPR–Cas components involved in the adaptation stage, the mechanistic and structural details of the spacer integration process, the recognition of foreign DNA as a substrate for new spacers and the evolution of the adaptation mechanism.

Studies into spacer acquisition in CRISPR–Cas systems have been carried out on many of the subtypes of these systems and in several organisms, including *Escherichia coli* (type I-E)^{10,24–33}, *Streptococcus thermophilus* (type II-A)^{6,9,34}, *Streptococcus pyogenes* (type II-A)³⁵,

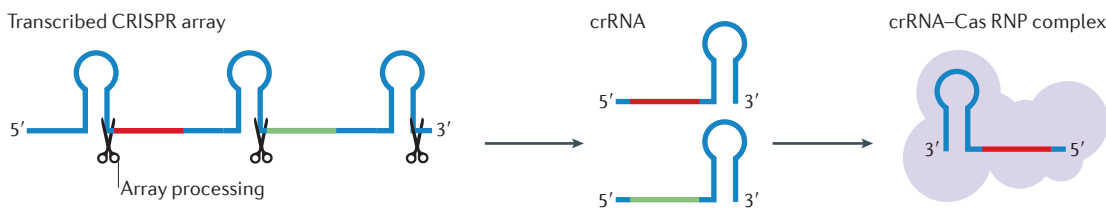
a Locus organization



b Adaptation



c Expression and maturation



d Interference

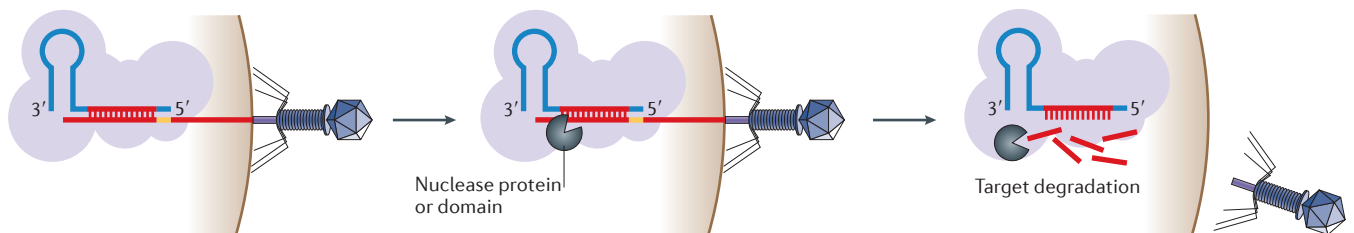


Figure 1 | The three stages of CRISPR–Cas immunity. **a** | Organization of a typical CRISPR–cas locus in a bacterial or archaeal genome. The numbers, order and identities of the *cas* genes are variable between CRISPR–Cas subtypes, and the number of spacer–repeat units varies between species. **b** | In the adaptation stage, the Cas1–Cas2 complex, which comprises two Cas1 dimers and a single Cas2 dimer, acquires a protospacer from the invader DNA and integrates it as a new spacer into the CRISPR array. Integration is coupled with a duplication of the first repeat. **c** | In the expression and maturation stage, the CRISPR array is transcribed and then processed into mature CRISPR RNAs (crRNAs), each containing a then transcribed spacer and part of the repeat sequence. These

crRNAs form ribonucleoprotein (RNP) complexes with Cas proteins. The Cas proteins in these complexes vary between subtypes and include CRISPR-associated complex for antiviral defence (Cascade) proteins (type I CRISPR–Cas systems), Cas9 (type II systems), Csm proteins (type III-A systems) and Cmr proteins (type III-B systems). **d** | In the interference stage, the crRNA–Cas RNP complex identifies the target DNA through complementary base-pairing in the presence of a protospacer-adjacent motif (PAM; in type I, type II and type V systems), and the target sequence is then degraded by nuclease proteins or domains. Both the position of the PAM and the identity of the nuclease that degrades the target vary between CRISPR–Cas subtypes.

Streptococcus agalactiae (type II-A)³⁶, *Pseudomonas aeruginosa* (type I-F)³⁷, *Haloarcula hispanica* (type I-B)³⁸, *Sulfolobus solfataricus* (type I-A and type III-B)^{33,39}, *Sulfolobus islandicus* (type III-B and type I-A)⁴⁰ and *Campylobacter jejuni* (type II-C)⁴¹. However, mainly for historical reasons, most of the detailed mechanistic insights were derived from studies of *E. coli* type I-E systems^{10,24–33} and *Streptococcus* spp. type II-A systems^{6,9,34–36}. Therefore, this Progress article focuses on these systems as the main models for spacer acquisition.

Composition of the adaptation machinery CRISPR–Cas adaptation is a complex, multistage process in which a protospacer needs to be extracted from an invading foreign DNA and subsequently stored within the CRISPR array as a spacer. First, the foreign DNA needs to be recognized as a target for spacer acquisition. Second, a sequence of a specific size (typically 30–40 bp, depending on the subtype of CRISPR–Cas system)⁴² needs to be acquired from the foreign DNA. Finally, the acquired sequence must be integrated

as a new spacer into the CRISPR array, and the adjacent repeat sequence needs to be duplicated (FIG. 1b). Although the components and prerequisites of the spacer acquisition machinery vary between organisms and subtypes of CRISPR–Cas system, several components seem to be universally conserved and are essential among all CRISPR–Cas subtypes. These components are the Cas proteins Cas1 and Cas2 and, within the CRISPR array locus, the leader sequence and the first CRISPR repeat.

Cas1 and Cas2 are essential for spacer acquisition in all studied CRISPR–Cas systems, but do not seem to have any role in the expression and maturation stage

or the interference stage^{11,43,44}. Cas1 and Cas2 are usually encoded in the same operon⁴⁴ and form a structurally stable protein complex (Cas1–Cas2)^{28,45,46} that

was shown to be involved in all stages of spacer acquisition^{24,28,29,33,45,46}. Cas1 is an endonuclease^{46–49}, and its endonuclease activity is essential for spacer acquisition²⁸.

Box 1 | Experimental systems for controlled studies of spacer acquisition

The adaptation process of CRISPR–Cas systems involves the acquisition of new spacers from foreign DNA and their addition to the CRISPR array in the bacterial chromosome. Major breakthroughs in understanding the principles of spacer acquisition were achieved following the development of experimental systems that enable direct and/or high-throughput investigation of the adaptation process. These systems generally use one of three approaches. The first approach takes advantage of the fact that new spacers are preferentially added juxtaposed to a regulatory ‘leader’ sequence that is found directly upstream of the CRISPR array. Each adaptation event entails the expansion of the array by one repeat and one spacer, together sized about 60 nucleotides; hence, PCR amplification with primers that anneal to the leader and a parental spacer can provide data on new spacer integration, either through the use of gel electrophoresis or by direct sequencing^{10,25–27,30–32,34,40,56,57,67,79} (see the figure, part a). This technique is compatible with various genetic backgrounds that include native^{25,30,31,67,74} or partial^{10,26,57} repertoires of cas genes, as well as manipulated leader and repeat sequences^{9,10}. The biochemical intermediates of spacer integration can also be examined using this approach, by Southern blotting of the PCR-amplified CRISPR array²⁴.

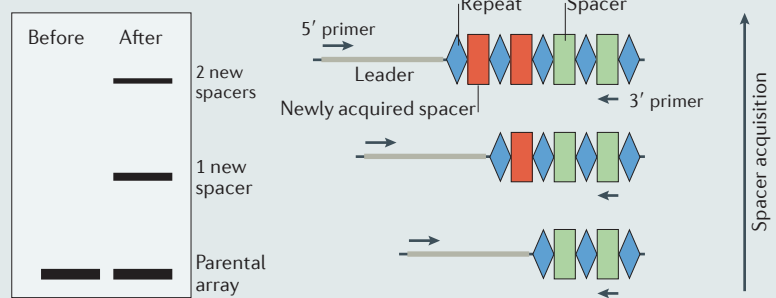
The second approach relies on a viability-based assay for clones that have incorporated new spacers into a CRISPR array that is genetically fused to an out-of-frame antibiotic-resistance gene²⁶ (*antibioticR*⁺; see the figure, part b). When a new spacer is integrated into the CRISPR array, the coding frame is repaired and the antibiotic-resistance gene is expressed, so that the survival of a clone following exposure to antibiotics denotes successful spacer integration. This system therefore provides antibiotic-based selection for spacer acquisition events and enables high-sensitivity detection of newly incorporated spacers without the need for the overexpression of Cas proteins²⁶.

The third approach is to assay spacer acquisition *in vitro*. In this approach, isotopically labelled protospacers are mixed *in vitro* with purified Cas1 and Cas2, which form a protein complex (Cas1–Cas2) that is essential for spacer acquisition, and a CRISPR array-containing plasmid that is an acceptor for spacer integration²⁹. The integration of spacers into the CRISPR array results in a relaxation of supercoiled plasmid DNA or the production of linear DNA, both of which can be observed through an electrophoretic mobility

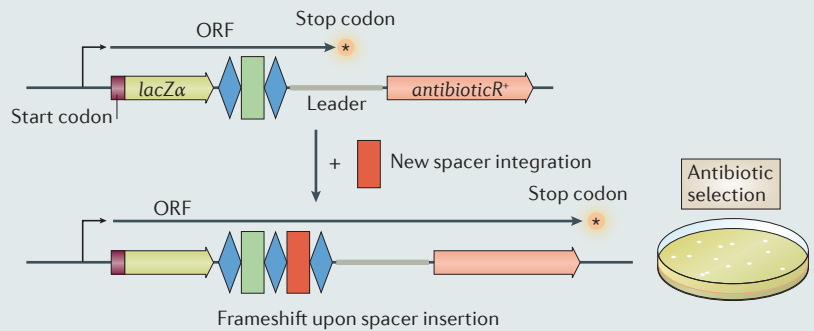
shift assay (EMSA) as mobility shifts on the gel (see the figure, part c). This assay demonstrated that the chemical steps of spacer integration are reminiscent of the activity of retroviral integrases²⁹. Alternatively, *in vitro* assays have been used to study the reverse reaction of spacer integration: disintegration. These assays

expose short labelled branched DNA intermediates of spacer integration to the acquisition machinery, thus producing reaction products that can be observed on a gel as a measure of activity. This assay was used to study the sequence specificity of the integration site in the CRISPR array³³.

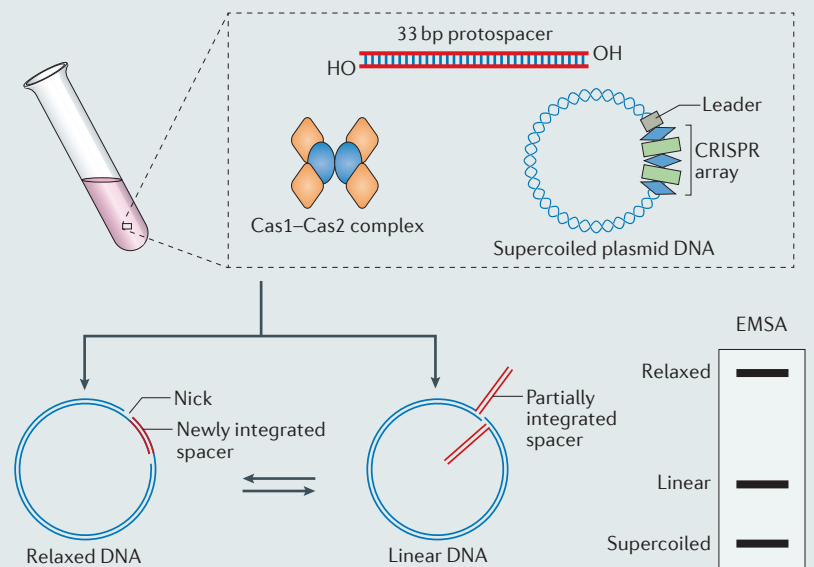
a Gel electrophoresis and sequencing



b



c



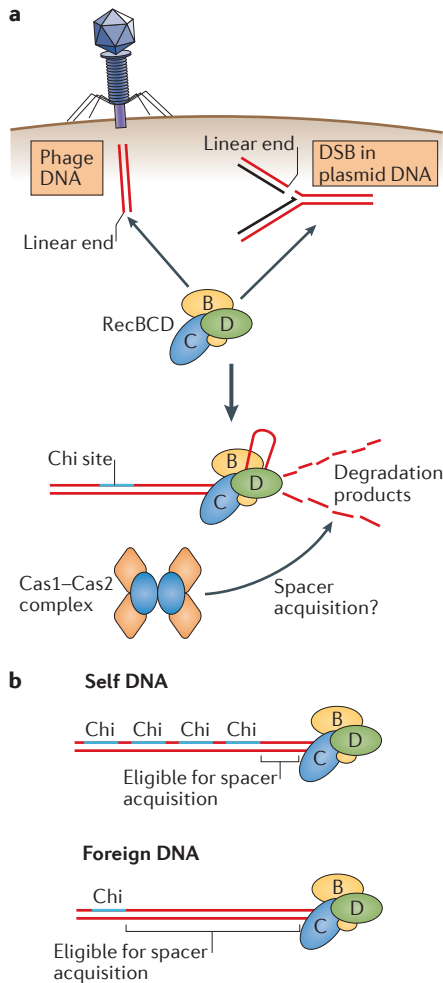


Figure 2 | Discrimination between self and non-self DNA in type I-E CRISPR-Cas system adaptation. **a** | For CRISPR-Cas systems, the source material for new spacers is suggested to be derived from the processing of linear double-stranded DNA (dsDNA) ends, which are found in phage DNA or are formed following a double-strand break (DSB). The multisubunit RecBCD nuclease enzyme processes these ends, producing single-stranded DNA (ssDNA) intermediates. DNA processing by RecBCD proceeds until the enzyme reaches the nearest instance of a specific octameric sequence known as a Chi site. **b** | The *Escherichia coli* genome is highly enriched in Chi sites, so that RecBCD processing is soon terminated and thus produces only a small amount of host genome-derived degradation material. By contrast, foreign DNA that lacks Chi site enrichment is more extensively processed, providing ample material for new spacers.

Cas2 has various DNA and RNA cleavage activities^{50–52}, but these activities are not essential for spacer acquisition²⁸; therefore, the primary role of Cas2 in spacer acquisition is currently thought not to involve its catalytic activity^{45,46}. Direct evidence for

the involvement of the Cas1–Cas2 complex in the adaptation stage was first provided in 2012 by genetic studies on the type I-E CRISPR–Cas system in *E. coli*¹⁰, for which it was shown that Cas1 and Cas2 are the only Cas proteins required for efficient spacer acquisition.

The Cas1–Cas2 complex typically inserts new spacers into the junction between the leader sequence and the first repeat of the CRISPR array²⁹. The leader sequence is a long AT-rich sequence positioned immediately upstream of the CRISPR array, and it usually contains both the promoter that drives crRNA expression and the recognition sequence for spacer insertion^{8–10}. The junction between the leader sequence and the first CRISPR repeat is the preferred site of new spacer integration, and the minimal sequence required for integration spans only a short segment at the 3' end of the leader sequence and a single repeat unit^{9,10}. Owing to the preference for integration at this junction, spacers are inserted into the CRISPR array with a polarity towards the leader sequence end of the array^{5,6}, generating a chronologically ordered array in which the most recently acquired spacer is the spacer most proximal to the leader sequence.

The sequences from which the spacers are derived are called protospacers (denoting the sequence segments residing in the foreign DNA molecule prior to integration into the CRISPR array). For type I, type II and type V CRISPR–Cas systems, a protospacer-adjacent motif (PAM) is present upstream or downstream of the protospacer in the foreign DNA^{3,53–56}. The PAM is a short (2–5 nucleotide) sequence that is essential for cleavage of the target DNA during the interference stage^{3,53,54}. During spacer acquisition, spacers are preferentially selected from protospacers that have a cognate PAM for the CRISPR–Cas system in question⁵⁴. Although the Cas1–Cas2 complex was shown to be sufficient to mediate PAM-dependent spacer acquisition in type I CRISPR–Cas systems¹⁰, PAM recognition in type II systems additionally requires Cas9 (REFS 34,35).

The source material for new spacers

At the initial stage of spacer acquisition, the foreign DNA needs to be recognized and processed to derive the substrate for spacer integration by the Cas1–Cas2 complex. A recent genome-wide study of protospacer hotspots in *E. coli* suggested that the substrates for integration are degraded DNA intermediates that are formed during

the repair of double-strand breaks (DSBs)⁵⁷. When a DSB occurs in an *E. coli* cell, the RecBCD exonuclease complex recognizes the exposed double-stranded DNA (dsDNA) end and then rapidly unwinds and degrades the DNA until it reaches an 8 bp sequence motif (5'-GCTGGTGG-3') called a Chi site⁵⁸. Using deep-sequencing analysis of millions of spacer acquisition events in Cas1–Cas2-expressing *E. coli*, it was found that protospacer hotspots are located between replication fork stalling sites (which are major sources of DSBs) and the nearest Chi site⁵⁷. This suggested that the Cas1–Cas2 complex acquires new spacers from the debris emerging from RecBCD-mediated DNA degradation (FIG. 2a). Indeed, the artificial induction of DSBs at a specific position in the *E. coli* genome resulted in the formation of a strong hotspot for spacer acquisition between the DSB site and the nearest Chi site on either side of the induced break⁵⁷. These data are also suggestive of an elegant solution to the problem of discrimination between self and non-self DNA in CRISPR–Cas-based immunity (see below).

Under native conditions, RecBCD is thought to degrade linear dsDNA into single-stranded DNA (ssDNA) molecules with sizes ranging from tens to thousands of nucleotides⁵⁸. Recent structural studies have shown that the Cas1–Cas2 complex binds to protospacers with a 23 bp dsDNA core and splayed ssDNA ends^{45,46} (see below). Presumably, therefore, RecBCD-generated ssDNA fragments reanneal in the cell to form incomplete dsDNA intermediates that are substrates for spacer acquisition by the Cas1–Cas2 complex. However, an alternative possibility is that the Cas1–Cas2 complex initially binds to ssDNA, and then DNA polymerase activity from an unknown source generates the second strand to form a dsDNA. Hence, further studies are required to elucidate the mechanism of the very early steps of spacer acquisition.

Discrimination between self and non-self DNA.

In natural settings, the accidental acquisition of spacers from 'self' DNA — that is, from the genome of the cell — instead of from invading DNA is usually detrimental, as it results in the degradation of self DNA by the CRISPR–Cas interference machinery. Such self-targeting leads to CRISPR–Cas autoimmunity⁵⁹, and it has been shown that escape from this autoimmunity usually involves the mutational inactivation of *cas* genes, mutations in the repeats next to the self-derived spacer or escape mutations in

the PAM⁵⁹. Therefore, it is necessary for CRISPR–Cas systems to avoid acquiring self DNA to minimize these harmful effects. Indeed, early observations in the *E. coli* type I-E CRISPR–Cas system showed a strong preference for spacer acquisition from foreign DNA and an avoidance of self DNA^{10,26,57}.

The involvement of the RecBCD machinery and Chi sites in generating the substrate for spacer acquisition provides a simple explanation for the strong bias against acquiring self DNA. Chi sites are highly enriched in the *E. coli* genome, occurring on average once every 4.6 kb (instead of approximately once every 65 kb, as expected by chance)⁵⁸. Therefore, when a DSB occurs in the *E. coli* genome, RecBCD degrades only a short length of self DNA before the degradation activity is halted by the nearest Chi site (which is 4.6 kb away, on average). Thus, only a small number of degraded self DNA molecules are generated as potential substrates for spacer acquisition by the Cas1–Cas2 complex. By contrast, a DSB in exogenous DNA that is not enriched for Chi sites results in long-range DNA degradation by RecBCD, generating ample substrates for new spacers (FIG. 2b). Moreover, as the genetic material of phages usually enters the host cell as linear dsDNA, the linear end is perceived by RecBCD as a DSB, promoting the degradation of phage DNA and the formation of substrates for new spacers. To counter this mechanism, some phages express RecBCD inhibitors, and others enrich their genomes with Chi sites⁶⁰.

The suggested RecBCD-based machinery also explains the preference of the Cas1–Cas2 complex for protospacers from high-copy-number plasmids, even though such plasmids are circular rather than linear. It has previously been documented that most DSBs in the cell are produced at replication forks during DNA replication^{61–63}. Importantly, two replication forks are present on the chromosome during DNA replication, but the number of replication forks on plasmid DNA is proportional to the plasmid copy number (one or two forks per copy). As a result, in cells with high-copy-number plasmids, replication forks are much more abundant on plasmid DNA than on the chromosome. This relative abundance of replication forks on plasmid DNA is therefore expected to cause more DSBs in plasmids than in the chromosome; this would yield more linear plasmid DNA molecules that form substrates for RecBCD and, ultimately, a larger number of plasmid-derived protospacers as source

material for the Cas1–Cas2 complex. Indeed, in several experimental systems in which the *E. coli* type I-E Cas1–Cas2 complex was expressed without the presence of the interference machinery, the acquisition of new spacers showed a strong bias for plasmid DNA compared with chromosomal DNA^{10,26,57}.

Interestingly, some CRISPR–Cas systems contain the protein Cas4, which has a RecB nuclease domain^{2,64,65} that has ssDNA-targeted exonuclease activity^{64,65}. One may speculate that the RecB domain of Cas4 operates as an alternative RecB nuclease in bacteria in which RecBCD is absent, or that it competes with host RecB.

It is important to note that the mechanism for discrimination between self and non-self DNA described here has to date been observed only in the type I-E CRISPR–Cas system of *E. coli*. It is possible that other systems in other organisms use alternative mechanisms to avoid self DNA during the adaptation process. For example, it has been observed that inactivation of the Cas9 nuclease activity in a type II-A CRISPR–Cas system leads to pervasive spacer acquisition from the self chromosome³⁴, indicating that a different mode of discrimination between self and non-self DNA operates in type II CRISPR–Cas systems.

Spacer acquisition

In the *E. coli* type I-E CRISPR–Cas system, Cas1 and Cas2 form a heterohexameric structural complex composed of two Cas1 dimers bound to either side of a single Cas2 dimer with a dissociation constant (K_d) of 290 nM^{28,45,46} (FIG. 3a). The Cas1–Cas2 complex seems to have a dual role in the adaptation stage, as it needs to both acquire protospacer DNA and integrate it into the CRISPR array²⁹ (FIG. 1b).

Two recent structural studies, in which the *E. coli* type I-E Cas1–Cas2 complex was crystallized bound to a protospacer, have shed light on the mechanisms involved in PAM recognition⁴⁶ and spacer size determination^{45,46} during spacer acquisition. These structural studies revealed a 33 bp protospacer bound to the Cas1–Cas2 complex such that the central 23 bp of the protospacer form a duplex with 5 bases on each side splayed into ssDNA ends^{45,46}. These 3' ssDNA ends are threaded into one of the monomers of each Cas1 dimer by an arginine-rich channel, positioning the 3'-OH group at the end of the ssDNA into the catalytic site (FIG. 3a,b); although the Cas1 monomers are all encoded by the same gene, the second Cas1 monomer of each dimer

does not have a 3'-OH within the catalytic site. The central segment of the 23 bp duplex binds to the surface of the Cas2 dimer and is mainly stabilized by the interactions formed between a set of arginine residues (denoted the arginine clamp⁴⁵) in the Cas1–Cas2 complex and the phosphate groups in the phosphodiester backbone of the protospacer.

These crystallographic studies also revealed the structural basis for spacer size determination. A conserved tyrosine residue in Cas1 (Tyr22) is responsible for bracketing the central duplex region of the protospacer, forming a wedge that terminates the dsDNA region and splays apart the ends of the DNA on each side of the duplex^{45,46}. One of the studies also revealed that the structural basis for PAM recognition is a sequence-specific pocket that is formed in the Cas1 dimer and recognizes the PAM-complementary sequence (5'-CTT-3', as the PAM of the *E. coli* type I-E CRISPR–Cas system is 5'-AAG-3'). The pocket positions the PAM-complementary sequence in the correct orientation relative to the catalytic histidine (His208) in the active site of one of the Cas1 monomers⁴⁶. This leads to cleavage of the ssDNA, leaving 5 bases of ssDNA that are terminated by the cytosine of the PAM-complementary sequence, and generating a 33 bp protospacer (FIG. 3b).

A comparison between protospacer-bound and DNA-free Cas1–Cas2 complexes reveals significant conformational changes in the Cas1–Cas2 complex upon protospacer binding (FIG. 3a). These changes involve the rotation of each of the Cas1 dimers in opposing directions around the Cas2 dimer, exposing the flat protein surface that binds to the central duplex and orienting the two Tyr22 residues so that they bracket the duplex. Furthermore, a catalytic pocket is formed in one monomer of each Cas1 dimer, enabling accurate cleavage of the 3' overhangs^{45,46}.

Spacer integration

Spacer integration in type I-E CRISPR–Cas systems. An important recent study that investigated spacer acquisition *in vitro* (BOX 1) suggests that the Cas1–Cas2 complex acts as an integrase, and provides evidence of a nicking activity that serially targets the two ends of the first repeat in the CRISPR array²⁹ (FIG. 3c). Results from this study indicate that the protospacer integrates into the CRISPR array in a two-step mechanism that resembles retroviral integration and DNA transposition²⁹. According to the suggested model, during the first step, the Cas1–Cas2 complex positions the 3'-OH

group at one end of the protospacer to catalyse a nucleophilic attack on the minus strand of the CRISPR array. This results in a branched intermediate in which the

protospacer is connected at one end to the 5'-phosphate of the minus strand of the first CRISPR repeat (FIG. 3c). During the second step, the protospacer — now covalently

attached to the minus strand of the repeat array — attacks the junction between the first CRISPR repeat and the leader sequence on the plus strand. As a result, the protospacer

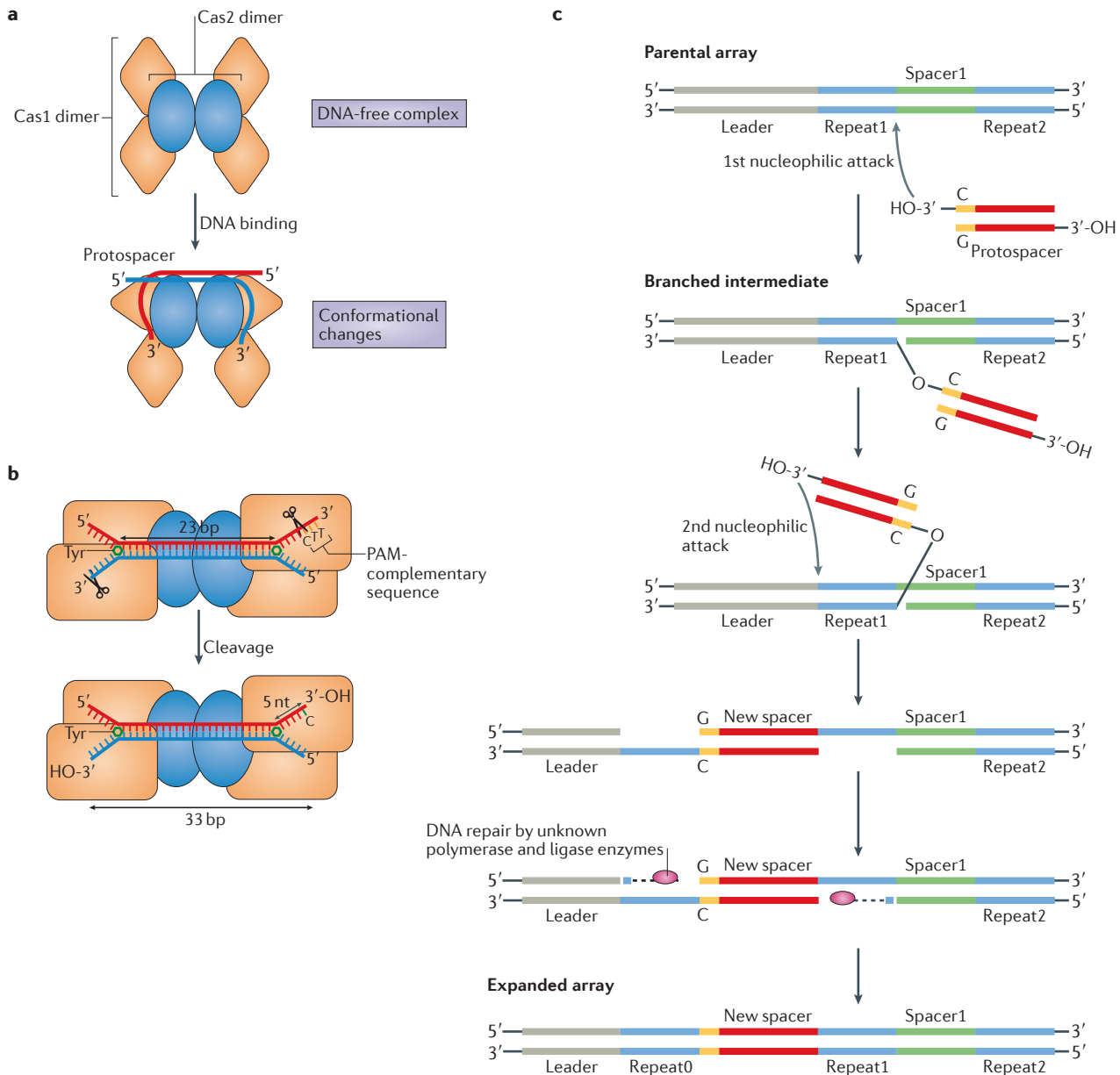


Figure 3 | Spacer integration into the CRISPR array. **a** | Schematic depiction of protospacer binding by the Cas1–Cas2 complex, comprising two Cas1 dimers and one Cas2 dimer. Conformational changes occur upon binding of protospacer DNA. The protospacer DNA, composed of a 23 bp double-stranded DNA (dsDNA) with single-stranded DNA (ssDNA) overhangs, is displayed on top of an arch that is formed by components of all six subunits. **b** | Schematic depiction of protospacer-adjacent motif (PAM)-specific protospacer cleavage by the Cas1–Cas2 complex prior to integration into the CRISPR array, showing the top view of the protospacer-bound complex. The ends of the protospacer on either side of a 23 bp dsDNA core are splayed into ssDNA by wedges formed by the conserved Tyr22 residue in two Cas1 monomers. A PAM-complementary sequence (5'-CTT-3') in the 3' ssDNA overhang is positioned within the active site (black scissors) of the catalytically active Cas1 subunit, where it is specifically recognized

and cleaved. Cleavage results in a protospacer intermediate with a precise length of 33 nucleotides, comprising the 23 bp dsDNA core and two splayed 5-nucleotide ssDNA overhangs with 3'-OH groups. It is suggested that this intermediate forms the substrate for spacer integration into the CRISPR array. **c** | A model for protospacer integration into the CRISPR array. The protospacer 3'-OH group carries out a nucleophilic attack on the 5' end of the first repeat, thus initiating spacer acquisition by forming a branched intermediate in which a single strand of the protospacer is ligated to a single strand of the CRISPR array. The 3'-OH group on the other protospacer strand generates a second nucleophilic attack on the 5' end of the opposing DNA strand of the repeat, which is juxtaposed to the leader sequence. The product of this reaction is an expanded CRISPR array with a new spacer and a duplicated repeat. The ssDNA gaps that are produced at the repeat sequences are filled and repaired by uncharacterized enzymes.

becomes a fully integrated spacer that is flanked on either side by one strand of the first CRISPR repeat, as ssDNA, and a gap on the opposing strand^{24,29} (FIG. 3c). Completion of the integration process requires both a DNA polymerase and a DNA ligase to fill the formed gaps, but the specific proteins that carry out these tasks have not yet been identified²⁹.

Notably, the nucleotide preferred by the protospacer for the nucleophilic attack on the CRISPR array is the 3'-cytosine derived from the PAM-complementary sequence²⁹. This preference generates a new spacer with a complementary guanine nucleotide as the first 5'-nucleotide²⁹ (FIG. 3c). Although this guanine nucleotide was originally considered to form part of the *E. coli* repeat sequence, it was later shown to be derived from the protospacer^{25,66,67}. The 3'-cytosine determines the spacer orientation inside the CRISPR array²⁹; when the protospacer substrate lacks a 3'-cytosine or contains a 3'-cytosine at both ends, the protospacer sequence can be integrated in either possible orientation into the array²⁹.

In the *in vitro* assay, spacer integration by the Cas1–Cas2 complex was possible when the protospacer was a dsDNA molecule with either blunt ends or 3'-overhangs, but not when the protospacer was a ssDNA molecule²⁹; this is consistent with the structure of protospacer-bound Cas1–Cas2 complexes^{45,46}. An OH group at each 3' end of the protospacer was found to be essential for the integration process. In addition, the protospacer was shown to integrate into supercoiled DNA, whereas nicked or linear DNA did not form an efficient acceptor for spacer integration²⁹. The structural basis for protospacer integration into the CRISPR array is still unclear; however, in the Cas1–Cas2–protospacer complex, there are two Cas1 monomers without protospacer DNA in their catalytic sites, and it has been hypothesized that these monomers are responsible for binding to the DNA of the CRISPR array during spacer integration⁴⁵.

A second recent *in vitro* study examined spacer integration by analysing the inverse reaction — disintegration (BOX 1) — using Cas1 and Cas2 from *E. coli* and Cas1 and

Cas2 from the type I-A CRISPR–Cas system found in *S. solfataricus*³³. Spacer disintegration was more efficient at the junction between the leader sequence and the first CRISPR repeat than at the junction between the first CRISPR repeat and the first spacer³³, leading to the proposal that the initial step of spacer integration occurs at the leader–repeat junction, rather than at the repeat–spacer junction, in a reversal of the order suggested in the study described above²⁹.

The Cas1–Cas2 complex was suggested to require, as part of its binding target, a palindromic sequence that can potentially form a cruciform DNA structure^{24,29}, which is a characteristic requirement of various integrases^{68,69}. It was shown that Cas1 can recognize such palindromic sequences⁴⁷ and that disruption of the predicted structure also arrests protospacer integration²⁴. Furthermore, an *in vitro* assay showed that spacer integration can occur at a putative cruciform structure adjacent to an AT-rich sequence (reminiscent of the leader sequence), even outside of the

Box 2 | Primed spacer acquisition

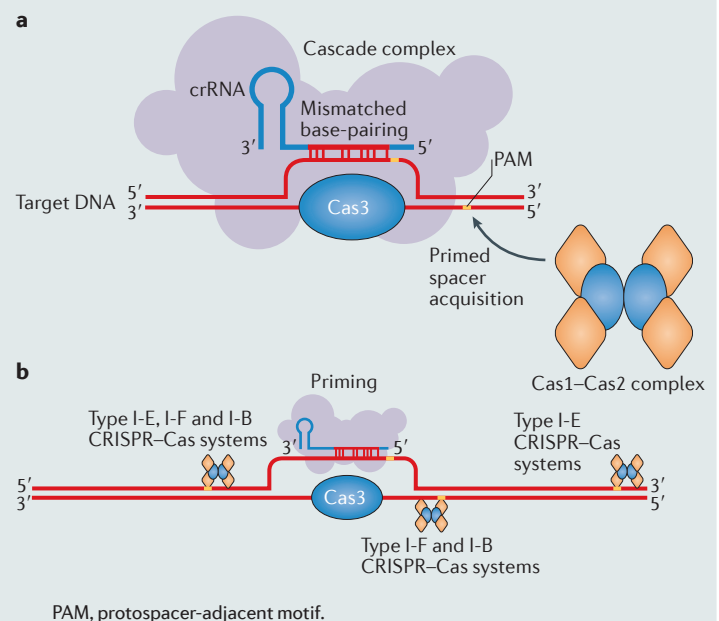
Primed spacer acquisition (or priming) denotes cases in which an existing spacer against a foreign DNA promotes the rapid and efficient acquisition of additional spacers from the same foreign DNA^{25,30,38,67,73–75}. Primed acquisition has been recorded in type I-E^{25,30,31,67}, type I-F⁷⁵ and type I-B³⁸ CRISPR–Cas systems and has been shown to require the activity of the CRISPR-associated complex for antiviral defence (Cascade) and Cas3 — which recognize and degrade foreign DNA in the interference stage — in addition to the Cas1–Cas2 adaptation complex^{25,30,38,67,75} (see the figure, part a). Intriguingly, priming occurs for both active spacers, which trigger interference^{27,67}, and inactive, mismatch-containing spacers, which cannot elicit interference^{25,74,75}.

In type I-E CRISPR–Cas systems, priming initiates strand-biased spacer acquisition, so that the additional spacers will almost always be derived from the same strand as the priming protospacer^{25,74}. By contrast, in type I-B and type I-F CRISPR–Cas systems, primed spacers are acquired from both strands, but with a skewed distribution that is clustered around the position of the priming protospacer^{38,75} (see the figure, part b). Several mechanistic models for priming have been suggested, but the molecular mechanism of this process remains unclear. The requirement of Cascade and Cas3 during priming led to the suggestion that Cas3 generates cleavage points or nicks that produce the substrate for acquisition by the Cas1–Cas2 complex^{25,30,38,67,74,75}. Interestingly, in type I-F CRISPR–Cas systems, *cas2* is frequently fused to *cas3*, which may facilitate more efficient priming⁷⁵.

A recent fluorescence resonance energy transfer (FRET)-based structural study suggested that the Cascade–CRISPR RNA (crRNA) complex uses two distinct modes of binding to its target⁷³. According to this model, the canonical binding mode, in which the target DNA is perfectly matched by the cognate crRNA, ensures interference and degradation of the target DNA. In the non-canonical mode, mismatched base-pairing between the crRNA and its target leads to low-fidelity binding and the initiation of priming. However, priming was also shown to occur for crRNAs that

perfectly match their target and that can effectively trigger interference, suggesting that interference and priming can occur simultaneously^{27,67}.

Mismatch-triggered priming provides an efficient evolutionary strategy for neutralizing phage or plasmid mutants that have evaded the initially acquired spacer. The acquisition of additional spacers from different regions of a single phage will provide additional interference targets, leading to markedly lower chances that escape mutants will evolve. Thus, it has been suggested that two types of spacer exist in nature: spacers for immediate protection and spacers that are selectively maintained for longer term, primed protection⁷⁴.



context of CRISPR arrays, which suggests that palindromic repeats are indeed an important determinant of the target site for integration²⁹. Nonetheless, the Cas1–Cas2 complex clearly recognizes the leader–repeat junction in addition to the palindromic sequence, as mutating the repeat sequence while maintaining the palindromic sequence inhibited spacer integration *in vivo*²⁴.

Spacer integration in other CRISPR–Cas systems. In type I–E CRISPR–Cas systems, Cas1 and Cas2 are both necessary and sufficient to drive spacer acquisition and integration¹⁰. By contrast, two recent studies have shown that additional Cas proteins are required for spacer acquisition and integration in type II–A CRISPR–Cas systems^{34,35}. These additional proteins include Cas9 (together with its accessory tracrRNA) and Csn2, a Cas protein that binds to dsDNA but has no identified enzymatic properties⁷⁰. On the basis of these findings, Cas9 became the first Cas protein known to be involved in all of the functional steps of CRISPR–Cas immunity — adaptation, expression and maturation, and interference.

Whereas the Cas1–Cas2 complex is responsible for PAM recognition during spacer acquisition in type I–E CRISPR–Cas systems⁴⁶, it has been demonstrated that Cas9 carries out this role in type II–A systems. Indeed, mutations in the Cas9 PAM-binding motif led to an accumulation of spacers that did not have a PAM³⁵. Although it has been shown that the nuclease activity of Cas9 is not necessary for spacer acquisition^{34,35}, the use of a Cas9 mutant lacking nuclease activity resulted in a significant bias (96%) towards spacers originating from the genome rather than a plasmid³⁴. As mentioned above, this bias suggests that Cas9 has a role in discriminating between self and non-self DNA in type II–A CRISPR–Cas systems.

Based on evidence showing that Csn2 is involved in the acquisition process, one might assume that other Cas proteins, in addition to Cas1 and Cas2, may have a role in the acquisition process in various CRISPR–Cas systems. In many CRISPR–*cas* loci, *cas1* and/or *cas2* are associated, either by gene fusion or by close genetic proximity, with specific *cas* genes. Altogether, about 7% of *cas1* genes are fused to other genes, including those encoding Cas4, Csx1, type III restriction enzymes of the COG4951 family, transcriptional regulators of the COG2378 family, reverse

transcriptases and Argonaute proteins⁷¹. The most common association of *cas1* is with *cas4*, which is adjacent to *cas1* in about one-third of cases and fused to it in another ~3% of cases⁷¹. Cas4 contains a RecB-like exonuclease domain and was shown to generate ssDNA overhangs by cleaving ssDNA (in either a 5′-to-3′ or a 3′-to-5′ direction)^{64,65,72}. As a result, it has been proposed that Cas4 may generate protospacer intermediates during spacer acquisition^{23,64,65}.

A series of recent studies have shown that, in certain cases, the Cascade and Cas3 interference machinery in type I CRISPR–Cas systems can be involved in spacer acquisition. This occurs in cases in which the CRISPR array already contains a spacer against a particular phage or plasmid, and leads to the acquisition of additional spacers from the same foreign element much more rapidly and efficiently than normal ‘naive’ (or ‘non-primed’) acquisition^{25,27,30,38,67,73–75}. This positive feedback loop between existing spacers and the acquisition of new spacers is termed primed acquisition or priming^{25,67} (BOX 2).

Evolution of the adaptation machinery

Recent studies of standalone *cas1* genes (that is, *cas1* genes located outside of CRISPR–*cas* loci) suggest a surprising evolutionary origin of Cas1 and the CRISPR–Cas system from transposable elements. Cas1 was found to be associated with a particular group of putative transposons that were named casposons⁷⁶, and it was proposed to function as the transposase of these transposons. One of the proteins that frequently flanks Cas1 in casposons, HenMarC1, has an amino-terminal domain that is related to Cas4⁷⁶. Furthermore, some casposons are flanked by palindromic terminal inverted repeats (TIRs), which conceptually resemble the CRISPR palindromic repeats^{76,77}.

On the basis of these observations, an evolutionary scenario has been proposed in which the Cascade–Cas3 system was initially an innate immune system with no capacity for adaptation. Such a system might have resembled RNAi systems, which require guide sequences but do not have the capability to acquire and store new immunological memory. The adaptation machinery possibly originated from a Cas1-containing casposon that lost one of its two TIRs and was transposed next to the Cascade–Cas3 innate immune system. Subsequently, the single remaining TIR could have undergone amplifications within the same locus and generated a CRISPR

array, forming the CRISPR–Cas system. Interestingly, some CRISPR–*cas* operons, classified as type U systems, include genes encoding Cascade complex components but do not include CRISPR arrays or *cas1* or *cas2* genes^{71,78}. The origin of Cas2 is unclear, although it has been proposed to be derived from a toxin–antitoxin module⁷⁷. Intriguingly, ancient transposons were also suggested to be the evolutionary source of the vertebrate adaptive immune system⁷⁷, indicating that mobile genetic elements might be key drivers of evolutionary innovations that lead to new defence systems^{76,77}.

Outstanding questions and outlook

In this Progress article, we have described the recent advances in our understanding of spacer acquisition. We now have a greater understanding of the biochemical steps involved in spacer integration, the roles of the individual CRISPR–Cas components, the structural basis for PAM recognition and spacer size determination, how the integration of self DNA is avoided, and the role of primed acquisition in generating enhanced immunity. Although the past 3 years have seen substantial progress in our understanding of the CRISPR–Cas adaptation process⁵⁶, key questions remain unanswered. First, additional mechanistic studies are required to understand how foreign DNA is biochemically processed to form the protospacer bound to the Cas1–Cas2 complex. In addition, the structural basis of protospacer integration into the CRISPR array is still unknown. The role of the catalytic activities of Cas2, which do not seem to be required for spacer acquisition, remains enigmatic, as are many mechanistic aspects of primed spacer acquisition. Finally, there is a paucity of mechanistic data concerning spacer acquisition in type III CRISPR–Cas systems, indicating a need to develop a well characterized model for these systems.

In the past 3 years, a thorough mechanistic understanding of CRISPR–Cas interference complexes, and particularly those involving Cas9, has yielded disruptive new tools for genome engineering and several additional biotechnological uses. By analogy, it is conceivable that a similarly comprehensive insight into the spacer acquisition process may pave the way for the creation of additional powerful tools. For example, CRISPR–Cas adaptation has a natural capacity for DNA-based information storage. If such a system could be properly manipulated to enable

on-demand acquisition, it may have the potential to form the basis of new ways to store information in living organisms. The rapid developments in the CRISPR–Cas field to date indicate that such applications may be developed sooner rather than later.

Gil Amitai and Rotem Sorek are at the Department of Molecular Genetics, Weizmann Institute of Science, Rehovot 76100, Israel.
Correspondence to R.S.
rotem.sorek@weizmann.ac.il

doi:10.1038/nrmicro.2015.14

Published online 11 Jan 2016

- Ishino, Y., Shinagawa, H., Makino, K., Amemura, M. & Nakata, A. Nucleotide sequence of the *iap* gene, responsible for alkaline phosphatase isozyme conversion in *Escherichia coli*, and identification of the gene product. *J. Bacteriol.* **169**, 5429–5433 (1987).
- Jansen, R., Embden, J. D., Gaastra, W. & Schouls, L. M. Identification of genes that are associated with DNA repeats in prokaryotes. *Mol. Microbiol.* **43**, 1565–1575 (2002).
- Bolotin, A., Quinquis, B., Sorokin, A. & Ehrlich, S. D. Clustered regularly interspaced short palindromic repeats (CRISPRs) have spacers of extrachromosomal origin. *Microbiology* **151**, 2551–2561 (2005).
- Mojica, F. J., Diez-Villasenor, C., Garcia-Martinez, J. & Soria, E. Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. *J. Mol. Evol.* **60**, 174–182 (2005).
- Pourcel, C., Salvignol, G. & Vergnaud, G. CRISPR elements in *Yersinia pestis* acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies. *Microbiology* **151**, 653–663 (2005).
- Barrangou, R. *et al.* CRISPR provides acquired resistance against viruses in prokaryotes. *Science* **315**, 1709–1712 (2007).
- Makarova, K. S., Aravind, L., Wolf, Y. I. & Koonin, E. V. Unification of Cas protein families and a simple scenario for the origin and evolution of CRISPR–Cas systems. *Biol. Direct* **6**, 38 (2011).
- Pougach, K. *et al.* Transcription, processing and function of CRISPR cassettes in *Escherichia coli*. *Mol. Microbiol.* **77**, 1367–1379 (2010).
- Wei, Y., Chesne, M. T., Terns, R. M. & Terns, M. P. Sequences spanning the leader-repeat junction mediate CRISPR adaptation to phage in *Streptococcus thermophilus*. *Nucleic Acids Res.* **43**, 1749–1758 (2015).
- Yosef, I., Goren, M. G. & Qimron, U. Proteins and DNA elements essential for the CRISPR adaptation process in *Escherichia coli*. *Nucleic Acids Res.* **40**, 5569–5576 (2012).
- Makarova, K. S. *et al.* An updated evolutionary classification of CRISPR–Cas systems. *Nat. Rev. Microbiol.* **13**, 722–736 (2015).
- Brouns, S. J. *et al.* Small CRISPR RNAs guide antiviral defense in prokaryotes. *Science* **321**, 960–964 (2008).
- Rouillon, C. *et al.* Structure of the CRISPR interference complex CSM reveals key similarities with cascade. *Mol. Cell* **52**, 124–134 (2013).
- Spilman, M. *et al.* Structure of an RNA silencing complex of the CRISPR–Cas immune system. *Mol. Cell* **52**, 146–152 (2013).
- Shmakov, S. *et al.* Discovery and functional characterization of diverse class 2 CRISPR–Cas systems. *Mol. Cell* **60**, 385–397 (2015).
- Jinek, M. *et al.* A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **337**, 816–821 (2012).
- Sapranaukas, R. *et al.* The *Streptococcus thermophilus* CRISPR/Cas system provides immunity in *Escherichia coli*. *Nucleic Acids Res.* **39**, 9275–9282 (2011).
- Zetsche, B. *et al.* Cpf1 is a single RNA-guided endonuclease of a class 2 CRISPR–Cas system. *Cell* **163**, 759–771 (2015).
- Deltcheva, E. *et al.* CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. *Nature* **471**, 602–607 (2011).
- Jiang, F. & Doudna, J. A. The structural biology of CRISPR–Cas systems. *Curr. Opin. Struct. Biol.* **30**, 100–111 (2015).
- Plagens, A., Richter, H., Charpentier, E. & Randau, L. DNA and RNA interference mechanisms by CRISPR–Cas surveillance complexes. *FEMS Microbiol. Rev.* **39**, 442–463 (2015).
- Sorek, R., Lawrence, C. M. & Wiedenheft, B. CRISPR-mediated adaptive immune systems in bacteria and archaea. *Annu. Rev. Biochem.* **82**, 237–266 (2013).
- van der Oost, J., Jore, M. M., Westra, E. R., Lundgren, M. & Brouns, S. J. CRISPR-based adaptive and heritable immunity in prokaryotes. *Trends Biochem. Sci.* **34**, 401–407 (2009).
- Arslan, Z., Hermanns, V., Wurm, R., Wagner, R. & Pul, U. Detection and characterization of spacer integration intermediates in type I-E CRISPR–Cas system. *Nucleic Acids Res.* **42**, 7884–7893 (2014).
- Datsenko, K. A. *et al.* Molecular memory of prior infections activates the CRISPR/Cas adaptive bacterial immunity system. *Nat. Commun.* **3**, 945 (2012).
- Diez-Villasenor, C., Guzman, N. M., Almendros, C., Garcia-Martinez, J. & Mojica, F. J. CRISPR-spacer integration reporter plasmids reveal distinct genuine acquisition specificities among CRISPR–Cas I-E variants of *Escherichia coli*. *RNA Biol.* **10**, 792–802 (2013).
- Fineran, P. C. & Charpentier, E. Memory of viral infections by CRISPR–Cas adaptive immune systems: acquisition of new information. *Virology* **434**, 202–209 (2012).
- Nunez, J. K. *et al.* Cas1–Cas2 complex formation mediates spacer acquisition during CRISPR–Cas adaptive immunity. *Nat. Struct. Mol. Biol.* **21**, 528–534 (2014).
- Nunez, J. K., Lee, A. S., Engelman, A. & Doudna, J. A. Integrase-mediated spacer acquisition during CRISPR–Cas adaptive immunity. *Nature* **519**, 193–198 (2015).
- Savitskaya, E., Semenova, E., Dedkov, V., Metlitskaya, A. & Severinov, K. High-throughput analysis of type I-E CRISPR/Cas spacer acquisition in *E. coli*. *RNA Biol.* **10**, 716–725 (2013).
- Shmakov, S. *et al.* Pervasive generation of oppositely oriented spacers during CRISPR adaptation. *Nucleic Acids Res.* **42**, 5907–5916 (2014).
- Yosef, I. *et al.* DNA motifs determining the efficiency of adaptation into the *Escherichia coli* CRISPR array. *Proc. Natl Acad. Sci. USA* **110**, 14396–14401 (2013).
- Rollie, C., Schneider, S., Brinkmann, A. S., Bolt, E. L. & White, M. F. Intrinsic sequence specificity of the Cas1 integrase directs new spacer acquisition. *eLife* **4**, e08716 (2015).
- Wei, Y., Terns, R. M. & Terns, M. P. Cas9 function and host genome sampling in type II-A CRISPR–Cas adaptation. *Genes Dev.* **29**, 356–361 (2015).
- Heler, R. *et al.* Cas9 specifies functional viral targets during CRISPR–Cas adaptation. *Nature* **519**, 199–202 (2015).
- Lopez-Sanchez, M. J. *et al.* The highly dynamic CRISPR1 system of *Streptococcus agalactiae* controls the diversity of its mobilome. *Mol. Microbiol.* **85**, 1057–1071 (2012).
- Cady, K. C., Bondy-Denomy, J., Heussler, G. E., Davidson, A. R. & O’Toole, G. A. The CRISPR/Cas adaptive immune system of *Pseudomonas aeruginosa* mediates resistance to naturally occurring and engineered phages. *J. Bacteriol.* **194**, 5728–5738 (2012).
- Li, M., Wang, R., Zhao, D. & Xiang, H. Adaptation of the *Haloarcula hispanica* CRISPR–Cas system to a purified virus strictly requires a priming process. *Nucleic Acids Res.* **42**, 2483–2492 (2014).
- Erdmann, S. & Garrett, R. A. Selective and hyperactive uptake of foreign DNA by adaptive immune systems of an archaeon via two distinct mechanisms. *Mol. Microbiol.* **85**, 1044–1056 (2012).
- Erdmann, S., Le Moine Bauer, S. & Garrett, R. A. Inter-viral conflicts that exploit host CRISPR immune systems of *Sulfolobus*. *Mol. Microbiol.* **91**, 900–917 (2014).
- Hooton, S. P. & Connerton, I. F. *Campylobacter jejuni* acquire new host-derived CRISPR spacers when in association with bacteriophages harboring a CRISPR-like Cas4 protein. *Front. Microbiol.* **5**, 744 (2014).
- Barrangou, R. & van der Oost, J. *CRISPR–Cas Systems* (eds Barrangou, R. & van der Oost, J.) (Springer, 2013).
- Makarova, K. S. *et al.* Evolution and classification of the CRISPR–Cas systems. *Nat. Rev. Microbiol.* **9**, 467–477 (2011).
- Makarova, K. S., Wolf, Y. I. & Koonin, E. V. Comparative genomics of defense systems in archaea and bacteria. *Nucleic Acids Res.* **41**, 4360–4377 (2013).
- Nuñez, J. K., Harrington, L. B., Kranzusch, P. J., Engelman, A. N. & Doudna, J. A. Foreign DNA capture during CRISPR–Cas adaptive immunity. *Nature* **527**, 535–538 (2015).
- Wang, J. *et al.* Structural and mechanistic basis of PAM-dependent spacer acquisition in CRISPR–Cas systems. *Cell* **163**, 840–853 (2015).
- Babu, M. *et al.* A dual function of the CRISPR–Cas system in bacterial antiviral immunity and DNA repair. *Mol. Microbiol.* **79**, 484–502 (2011).
- Kim, T. Y., Shin, M., Huynh Thi Yen, L. & Kim, J. S. Crystal structure of Cas1 from *Archaeoglobus fulgidus* and characterization of its nucleolytic activity. *Biochem. Biophys. Res. Commun.* **441**, 720–725 (2013).
- Wiedenheft, B. *et al.* Structural basis for DNase activity of a conserved protein implicated in CRISPR-mediated genome defense. *Structure* **17**, 904–912 (2009).
- Beloglazova, N. *et al.* A novel family of sequence-specific endoribonucleases associated with the clustered regularly interspaced short palindromic repeats. *J. Biol. Chem.* **283**, 20361–20371 (2008).
- Nam, K. H. *et al.* Double-stranded endonuclease activity in *Bacillus halodurans* clustered regularly interspaced short palindromic repeats (CRISPR)-associated Cas2 protein. *J. Biol. Chem.* **287**, 35943–35952 (2012).
- Samai, P., Smith, P. & Shuman, S. Structure of a CRISPR-associated protein Cas2 from *Desulfovibrio vulgaris*. *Acta Crystallogr. Sect. F Struct. Biol. Cryst. Commun.* **66**, 1552–1556 (2010).
- Deveau, H. *et al.* Phage response to CRISPR-encoded resistance in *Streptococcus thermophilus*. *J. Bacteriol.* **190**, 1390–1400 (2008).
- Mojica, F. J., Diez-Villasenor, C., Garcia-Martinez, J. & Almendros, C. Short motif sequences determine the targets of the prokaryotic CRISPR defence system. *Microbiology* **155**, 733–740 (2009).
- Shah, S. A., Erdmann, S., Mojica, F. J. & Garrett, R. A. Protospacer recognition motifs: mixed identities and functional diversity. *RNA Biol.* **10**, 891–899 (2013).
- Heler, R., Marraffini, L. A. & Bikard, D. Adapting to new threats: the generation of memory by CRISPR–Cas immune systems. *Mol. Microbiol.* **93**, 1–9 (2014).
- Levy, A. *et al.* CRISPR adaptation biases explain preference for acquisition of foreign DNA. *Nature* **520**, 505–510 (2015).
- Dillingham, M. S. & Kowalczykowski, S. C. RecBCD enzyme and the repair of double-stranded DNA breaks. *Microbiol. Mol. Biol. Rev.* **72**, 642–671 (2008).
- Stern, A., Keren, L., Wurtzel, O., Amitai, G. & Sorek, R. Self-targeting by CRISPR: gene regulation or autoimmunity? *Trends Genet.* **26**, 335–340 (2010).
- Bobay, L. M., Touchon, M. & Rocha, E. P. Manipulating or superseding host recombination functions: a dilemma that shapes phage evolvability. *PLoS Genet.* **9**, e1003825 (2013).
- Kuzminov, A. Single-strand interruptions in replicating chromosomes cause double-strand breaks. *Proc. Natl Acad. Sci. USA* **98**, 8241–8246 (2001).
- Michel, B. *et al.* Rescue of arrested replication forks by homologous recombination. *Proc. Natl Acad. Sci. USA* **98**, 8181–8188 (2001).
- Shee, C. *et al.* Engineered proteins detect spontaneous DNA breakage in human and bacterial cells. *eLife* **2**, e01222 (2013).
- Zhang, J., Kasciukovic, T. & White, M. F. The CRISPR associated protein Cas4 is a 5′ to 3′ DNA exonuclease with an iron-sulfur cluster. *PLoS ONE* **7**, e47232 (2012).
- Lemak, S. *et al.* The CRISPR-associated Cas4 protein Pcal_0546 from *Pyrobaculum caldifontis* contains a [2Fe-2S] cluster: crystal structure and nuclease activity. *Nucleic Acids Res.* **42**, 11144–11155 (2014).
- Goren, M. G., Yosef, I., Auster, O. & Qimron, U. Experimental definition of a clustered regularly interspaced short palindromic duplcon in *Escherichia coli*. *J. Mol. Biol.* **423**, 14–16 (2012).
- Swarts, D. C., Mosterd, C., van Passel, M. W. & Brouns, S. J. CRISPR interference directs strand specific spacer acquisition. *PLoS ONE* **7**, e35888 (2012).

68. Cote, A. G. & Lewis, S. M. Mus81-dependent double-strand DNA breaks at *in vivo*-generated cruciform structures in *S. cerevisiae*. *Mol. Cell* **31**, 800–812 (2008).
69. Lilley, D. M. & White, M. F. The junction-resolving enzymes. *Nat. Rev. Mol. Cell Biol.* **2**, 433–443 (2001).
70. Nam, K. H., Kurinov, I. & Ke, A. Crystal structure of clustered regularly interspaced short palindromic repeats (CRISPR)-associated Csn2 protein revealed Ca^{2+} -dependent double-stranded DNA binding activity. *J. Biol. Chem.* **286**, 30759–30768 (2011).
71. Makarova, K. S., Wolf, Y. I. & Koonin, E. V. The basic building blocks and evolution of CRISPR–CAS systems. *Biochem. Soc. Trans.* **41**, 1392–1400 (2013).
72. Lemak, S. *et al.* Toroidal structure and DNA cleavage by the CRISPR-associated [4Fe-4S] cluster containing Cas4 nuclease SSO0001 from *Sulfolobus solfataricus*. *J. Am. Chem. Soc.* **135**, 17476–17487 (2013).
73. Blosser, T. R. *et al.* Two distinct DNA binding modes guide dual roles of a CRISPR-Cas protein complex. *Mol. Cell* **58**, 60–70 (2015).
74. Fineran, P. C. *et al.* Degenerate target sites mediate rapid primed CRISPR adaptation. *Proc. Natl Acad. Sci. USA* **111**, E1629–E1638 (2014).
75. Richter, C. *et al.* Priming in the Type I-F CRISPR-Cas system triggers strand-independent spacer acquisition, bi-directionally from the primed protospacer. *Nucleic Acids Res.* **42**, 8516–8526 (2014).
76. Krupovic, M., Makarova, K. S., Forterre, P., Prangishvili, D. & Koonin, E. V. Casposons: a new superfamily of self-synthesizing DNA transposons at the origin of prokaryotic CRISPR-Cas immunity. *BMC Biol.* **12**, 36 (2014).
77. Koonin, E. V. & Krupovic, M. Evolution of adaptive immunity from transposable elements combined with innate immune systems. *Nat. Rev. Genet.* **16**, 184–192 (2015).
78. Koonin, E. V. & Makarova, K. S. CRISPR-Cas: evolution of an RNA-based adaptive immunity system in prokaryotes. *RNA Biol.* **10**, 679–686 (2013).
79. Paez-Espino, D. *et al.* Strong bias in the bacterial CRISPR elements that confer immunity to phage. *Nat. Commun.* **4**, 1430 (2013).

Acknowledgements

The authors thank P. Fineran for discussions on primed acquisition. R.S. was supported, in part, by the Israel Science Foundation (personal grant 1303/12 and Israeli Centers of Research Excellence (I-CORE) grant 1796/12), the European Research Council (ERC; Starting Grant 260432), the Human Frontier Science Program (HFSP; grant RGP0011/2013), the Abisch-Frenkel foundation, the Pasteur-Weizmann council, the Institut Merieux, the Leona M. and Harry B. Helmsley Charitable Trust, the Minerva Foundation and the Deutsche Forschungsgemeinschaft (a Deutsch-Israelische Projektkooperation (DIP) grant).

Competing interests statement

The authors declare no competing interests.