Evolution of adaptive immunity from transposable elements combined with innate immune systems

Eugene V. Koonin and

National Center for Biotechnology Information, National Library of Medicine, Bethesda, Maryland 20894, USA

Mart Krupovic

Institut Pasteur, Unité Biologie Moléculaire du Gène chez les Extrêmophiles, 25 rue du Docteur Roux, 75015 Paris, France

Abstract

Adaptive immune systems in prokaryotes and animals give rise to long-term memory through modification of specific genomic loci, such as by insertion of foreign (viral or plasmid) DNA fragments into clustered regularly interspaced short palindromic repeat (CRISPR) loci in prokaryotes and by V(D)J recombination of immunoglobulin genes in vertebrates. Strikingly, recombinases derived from unrelated mobile genetic elements have essential roles in both prokaryotic and vertebrate adaptive immune systems. Mobile elements, which are ubiquitous in cellular life forms, provide the only known, naturally evolved tools for genome engineering that are successfully adopted by both innate immune systems and genome-editing technologies. In this Opinion article, we present a general scenario for the origin of adaptive immunity from mobile elements and innate immune systems.

All cellular organisms persist and evolve under a perennial onslaught of mobile genetic elements (MGEs), such as transposons, viral sequences and plasmids. Many, if not most, of these diverse, 'selfish' elements insert into the chromosomes of the cellular hosts, either as an obligate part of their life cycles or at least sporadically. In multicellular eukaryotes, MGEs constitute a substantial proportion of the host genome, for example, >50% of the genome in mammals and >70% of the genome in some plants[1–3]. Integrated MGEs are also present in the genomes of most bacteria and archaea[4,5]; although they are not as abundant as those in eukaryotes, these elements account for up to 30% of some bacterial genomes[6,7].

Transposons are DNA segments that move from one location in the host genome to another. Most of the transposons can be grouped into two classes[8,9]. Class I elements (also known as retrotransposons) transpose via an RNA intermediate which, prior to integration, is copied back to the DNA form by the element-encoded reverse trans-criptase. Class II DNA transposons move in the host genome via the 'cut-and-paste' mechanism, whereby the transposon is excised from its initial location and inserted into a new locus. Most of the class II transposons have characteristic terminal inverted repeats (TIRs) but differ widely with

---

respect to the element size and gene content, the mechanisms of transposition and the transposases encoded[8,10,11]. The majority of the transposases belong to the DDE superfamily (which is named after two aspartate residues and one glutamate residue that form the catalytic triad), but several other unrelated families of transposases have been identified[8,10,11]. Some transposons encode transposases that are homologous to the rolling-circle replication initiation endonucleases found in single-stranded DNA viruses and plasmids[12–14], whereas other transposases are homologous to bacteriophage tyrosine or serine recombinases[15,16], or to eukaryotic APE1-like DNA repair endonucleases (which function in conjunction with reverse transcriptases)[17]. Such diversity of transposases strongly suggests that transposons have emerged on multiple independent occasions via recruitment of non-homologous endonucleases.

Owing to the ubiquity and high abundance of MGEs, their co-evolution with cellular hosts is a perennial parasite–host 'arms race' in which the two sides evolved extremely diverse and elaborate systems of defence and counter-defence[18–22]. Notably, many defence systems — including restriction–modification enzymatic modules, toxin–antitoxin and the clustered regularly interspaced short palindromic repeat–CRISPR-associated protein (CRISPR–Cas) systems in prokaryotes, and the apoptosis machinery in eukaryotes — seem to be 'guns for hire'; that is, they are also recruited by viruses and other MGEs for counter-defence[23,24].

All organisms have a plethora of innate immunity mechanisms, and many also have adaptive immunity[25–27]. In general, innate immunity covers all systems of defence against a broad range of pathogens, whereas adaptive immunity is tailored towards a specific pathogen, and its essential feature is immunological memory, whereby an organism that survives an encounter with a particular pathogen is specifically protected from that pathogen for the long term (often for the lifetime of the individual). Adaptive immunity is highly specific and extremely efficient against many pathogens, despite numerous powerful counter-defence strategies evolved by the pathogens[18–22].

In prokaryotes, innate immunity mechanisms include the well-studied restriction–modification enzymatic modules and multiple less thoroughly characterized systems[28]. Notable among the latter is the recently described mechanism that uses bacterial homologues of the eukaryotic Argonaute proteins — the key enzymes of RNA interference (RNAi) — to generate guide RNA or DNA molecules that are then used to inactivate foreign genomes[29–31]. Until recently, prokaryotes have not been thought to have adaptive immunity. However, this perception was overturned by the discovery of the CRISPR–Cas systems that are represented in most archaea and many bacteria (FIG. 1a). CRISPR–Cas is an immunity mechanism that functions by incorporating fragments of foreign (viral or plasmid) DNA into CRISPR cassettes and then using the transcripts of these unique spacers to target and inactivate the cognate genomes[28–38]. Although the immunological memory of the CRISPR–Cas system is short-lived by evolutionary standards, extremely efficient and specific immunity can be transmitted across many thousands of generations[39]. Thus, the CRISPR–Cas system fully satisfies the definition of adaptive immunity and is also a mechanism of bonafide Lamarckian adaptive evolution[40].

Eukaryotes encompass a variety of innate and adaptive immunity mechanisms of their own; some of these mechanisms seem to have their roots in prokaryotes, whereas others are eukaryote-specific. All eukaryotes seem to have some form of RNAi, a powerful defence system that uses RNA guides to inactivate invading nucleic acids, primarily those of RNA viruses[41–44]. In addition, animals encompass the paradigmatic system of antibacterial innate immunity centred around Toll-like receptors[45,46], and vertebrates (and possibly other deuterostomes) also have the equally well-characterized interferon antiviral response[47,48]. Historically, the most well-known form of anti-parasitic defence is adaptive (acquired) immunity, which is prominent in mammals and is also represented in all other vertebrates[49,50]. The specificity of the adaptive immunity of jawed vertebrates is achieved via proliferation of lymphocyte clones that carry immunoglobulin receptors for antigens of the given pathogen and that are selected accordingly from an enormous pre-existing repertoire of cells with diverse receptors[51,52]. In contrast to the prokaryotic adaptive immune system, in vertebrates, immunological memory is limited to somatic cells and has no transgenerational inheritance. Instead, the vast repertoire of immunoglobulin genes is generated via dedicated diversification processes known as V(D)J recombination — in which variable (V), diversity (D) and joining (J) segments are recombined — and hypermutation[53,54] (FIG. 1b).

Paradoxically, insights into the origins of adaptive immune systems in both eukaryotes and prokaryotes come from the least expected field of research — namely, studies on MGEs. It was demonstrated that *RAG1*, which encodes the key enzyme of V(D)J recombination, is derived from a eukaryotic transposon[55,56]. More recent studies on bacterial and archaeal mobilomes have provided clues regarding the origin of the CRISPR–Cas system[57].

Specifically, we identified a novel family of archaeal and bacterial MGEs that were named 'casposons' because they encode Cas1 homologues that are implicated as the transposase of these elements[57]. The discovery of casposons puts a new twist on the origin of CRISPR–Cas, especially given that in phylogenetic trees casposon Cas1 does not cluster with any particular group of CRISPR-associated Cas1 proteins, which is compatible with a basal position of casposons in the phylogenetic tree of the Cas1 family. We proposed that casposons could have been at the 'root' of CRISPR–Cas[57]. Below, we develop this proposal into a complete evolutionary scenario in which CRISPR–Cas was derived from a casposon and an innate immune system, and discuss the striking parallels with the evolution of adaptive immunity in animals, as well as general implications of the naturally evolved genome engineering capacity of MGE-encoded recombinases.

Evolutionary origin of CRISPR–Cas

Prokaryotes have evolved two analogous mechanisms of immunity — namely CRISPR–Cas and Argonaute-based systems — that rely on short guide RNA or DNA molecules for targeting and inactivating of the nucleic acids of invading MGEs[29,30]. Despite similar mechanisms of action, the CRISPR–Cas system is adaptive, whereas the Argonaute-centred system is an embodiment of innate immunity and is homologous to eukaryotic RNAi. The key distinction between these adaptive and innate immune systems lies in the ability of the CRISPR–Cas system to keep a record of past infections by incorporating spacer sequences

derived from MGEs into the dedicated CRISPR loci[28–38] (FIG. 1a). The immunization process, known as adaptation, is mediated by the concerted action of two proteins, Cas1 and Cas2 (REFS 34,58,59). These two proteins are conserved in the three major types of functionally characterized CRISPR–Cas systems (FIG. 2) and can be considered the signature proteins of the systems[32,60,61]. By contrast, other CRISPR–Cas components are mostly type-specific. These other components include Cascade (CRISPR-associated complex for antiviral defence), which mediates the processing of primary CRISPR transcripts, generates the mature guide CRISPR RNAs (crRNAs), and loads them on the target DNA, and 'executor' nucleases that are directly involved in the cleavage of the target DNA (FIGS 1a,2). Consequently, it seems that the CRISPR–Cas immunity mechanism emerged via the fusion of originally independent functional modules — the block of genes encoding an RNA- or DNA-guided innate immune system, and a module responsible for the adaptation process[60]. The 'last piece in the puzzle' is the source of the CRISPR loci. Tracing the origins of these distinct components of CRISPR–Cas is thus expected to shed light on the emergence of adaptive immunity in prokaryotes.

Cascades: the effector modules of CRISPR–Cas

The Cascade complexes of the Type I and Type III CRISPR–Cas systems consist of the Cas5 and Cas7 proteins, the large-subunit protein Cas8 (in Type I) or Cas10 (in Type III) and the small-subunit proteins; in some Type I systems, Cas6 proteins, the RNases directly responsible for guide RNA precursor processing, are also subunits of the Cascade complexes[61–63]. At the heart of the Cascade complexes are RNA recognition motif (RRM) domains, which are common RNA-binding domains in all cellular organisms[64–66]. The Cas5, Cas6, Cas7 and Cas10 proteins all contain one or two RRM domains[28,32,61]. It has been proposed that the Type III Cascade complex is the ancestral form that could have evolved from a simple double-RRM protein through fusion with a histidine–aspartate (HD) nuclease domain and a series of RRM domain duplications[61,67]. Once in existence, the Cascade complex could initially function as an innate immune system, analogous to the extant Argonaute proteins, although sequence analysis has unequivocally showed that the Argonaute-based system is evolutionarily unrelated to the Cascade complexes[68]. The credence to this hypothesis is given by the fact that many Type III CRISPR–Cas loci, in particular those of subtype IIIB, are not associated with CRISPR cassettes or the Cas1–Cas2 module and apparently use, in *trans,* the adaptation machinery of other Type I or Type III systems present in the same genomes[61,67]. Furthermore, a recent comparative genomic analysis has uncovered a growing variety of Type IV (formerly Type U) CRISPR–Cas systems[61] (FIG. 2). Similar to some of the Type III systems, Type IV systems lack Cas1, Cas2 and the CRISPR cassettes but, in this case, the respective genomes do not typically encompass any other CRISPR–Cas loci[61,67]. Thus, although none of these 'minimal' CRISPR–Cas variants have been functionally characterized so far, they clearly cannot provide adaptive immunity via genome manipulation that is characteristic of canonical CRISPR–Cas, but are most likely to represent a distinct innate immunity mechanism. In a close analogy to the small interfering RNA (siRNA) branch of the eukaryotic RNAi system and the Argonaute-based bacterial innate immune systems[30,68–70], the 'solo- Cascade' modules might generate small guide RNAs from transcripts of invading MGE genomes or guide DNAs directly from such genomes, and use these guide molecules for the inactivation

of the cognate foreign DNA. This putative form of innate immunity could resemble the ancestral state of the Cascade complex that was a key contributor in the evolution of CRISPR–Cas.

Cas1–Cas2: the immunization (informational) module

Recently, the likely source of the CRISPR–Cas immunization module has also been uncovered. Cas1 and Cas2 are endonucleases that form a heterohexamer involved in the acquisition of the protospacer sequences from the invading MGEs and insertion of the spacers into the CRISPR loci[58]. It has been demonstrated that the nuclease activity of Cas2 (REFS 71,72) is not required for this process, whereas the activity of Cas1 is essential; thus, Cas1 is the primary enzyme involved in immunization[58]. The endonuclease and DNA strand-rejoining activities of Cas1 mechanistically resemble the respective activities of MGE-encoded integrases and transposases, although Cas1 is not homologous to any of the known recombinase families[58,59]. Indeed, transposon-like elements of the casposon superfamily encode Cas1 and apparently use its endonuclease activity for integration into and excision out of the cellular genome[57] — a role strongly reminiscent of that postulated for the Cas1–Cas2 complex during spacer sequence acquisition in CRISPR–Cas. Deep branching of the casposon Cas1 sequences within the global Cas1 phylogeny has led to the proposal that casposons could have played a pivotal part in the emergence of CRISPR–Cas, specifically by providing the ancestral *cas1* gene[57]. Under the proposed scenario, CRISPR–Cas would emerge when a casposon inserted into an archaeal genome next to a solo-Cascade operon (FIG. 3).

Casposons are large MGEs that, in addition to the genes encoding Cas1 and a family B DNA polymerase (PolB) that are present in each of them, encompass a broad diversity of protein-coding genes that are found among different casposons[57]. Several of these dispensable casposon genes encode various nucleases and helicases — enzymes that are common in CRISPR–Cas systems, including a homologue of Cas4, a nuclease present in the majority of CRISPR–Cas systems (FIG. 2). Notably, in some CRISPR–Cas systems Cas4 is fused to Cas1, suggesting that it could play a part in spacer sequence acquisition, although a role of Cas4 in programmed cell death has also been proposed[60,67]. We hypothesize that the casposon at the origin of CRISPR–Cas incorporated several ancestral *cas* genes, in particular, *cas1* and *cas4* (FIG. 3).

The Cas2 protein is a homologue of VapD— a typical prokaryotic toxin that has the activity of an mRNA interferase, which is a nuclease that specifically cleaves ribosome-associated mRNAs to induce dormancy or to kill the cell[28,73,74]. Accordingly, although either RNase or DNase activity has been reported for Cas2 proteins from different prokaryotes[71,72], Cas2 is most likely to have originated from a typical toxin–antitoxin module, which could have already been present in the casposon that gave rise to CRISPR–Cas (FIG. 3). Although none of the currently known casposons carry recognizable toxin–antitoxin systems, toxin–antitoxin modules are common in other bacterial and archaeal MGEs[29,75,76].

CRISPR cassettes

The key Cas proteins might not be the only contribution of casposons to the emergence of CRISPR–Cas; the CRISPR cassettes, which are perhaps the most enigmatic component of the CRISPR–Cas systems, might have also been derived from casposons. By definition, CRISPRs are clusters of short palindromic repeats that are interspersed with unique spacer sequences. Although not universal, the palindromic character of the repeats is widespread in the CRISPR cassettes from different organisms[77]. These repeats are thought to be recognized by the Cas1–Cas2 complex that introduces a staggered cut to allow the incorporation of new spacer sequences into the CRISPR arrays[78,79]. In the case of the casposons, according to the proposed model[57], the Cas1 recognition site lies within the TIRs that are present at the extremities of all casposons. Similar to CRISPR repeats, TIRs from some casposons display a palindromic feature[57] and even share sequence similarity with CRISPR repeats from certain organisms (FIG. 4a,b). Although TIRs are variable in size (25–602 bp)[57], their median length is around 50 bp, which is within the reported size range of CRISPR repeats (20–50 bp)[35]. Thus, casposon TIRs are similar to CRISPR repeats with regard to the sequence, size, secondary structure and the (postulated) ability to bind to Cas1. Inactivation of a TIR at one of the extremities of an integrated casposon would immobilize the inserted casposon genes and produce a palindromic sequence that is reminiscent of a CRISPR unit.

Emergence of CRISPR–Cas

Importantly, the casposon Cas1 is expected to be capable of recognizing and acting on its TIR substrate in *trans*, similarly to the way in which the Cas1–Cas2 complex operates on CRISPR cassettes. Indeed, physical coupling of the target sequence (casposon TIR) with the gene encoding the protein that recognizes it (casposon Cas1) is not necessary, as indicated by the ability of transposases to mobilize non-autonomous MGEs that contain the cognate transposase-binding sites within their TIRs[8,10]. Consequently, recognition of such 'solo-TIRs', and their subsequent amplification within the same locus, would eventually result in arrays of palindromic repeats, the putative ancestors of CRISPRs (FIG. 3). Indeed, such physical uncoupling of the recombinase from its target could have been a prerequisite for the emergence of a stably inheritable immune system. The scenario of the emergence of a CRISPR–Cas system from a casposon and a solo-Cascade then becomes less complex, requiring only integration of the casposon next to the Cascade complex, proliferation of the repeats originating from the casposon TIR and deletion of some of the casposon genes, in particular, the *polB* gene (FIG. 3).

Type II CRISPR–Cas systems differ substantially from Type I and Type III systems in terms of the organization of the processing-executive module, which in Type II systems consists of a single large Cas9 protein. This protein binds to the crRNA (which is processed with the help of bacterial RNase III), mediates its annealing to the target DNA and cleaves the target via its two nuclease domains, RuvC and HNH[32,80,81]. Strikingly, the Cas9 protein is homologous to a family of transposon-encoded proteins known as TnpB (also known as Fanzors) that contain the RuvC-like nuclease domain but that are not required for transposition[81]. Several transposons encode only the TnpB protein and use a transposase in *trans*[82]. The Type II CRISPR–Cas system is most likely to have evolved when a transposon

encoding a Cas9 ancestor inserted into a type I CRISPR–Cas locus and replaced the genes for the Cascade subunits. Thus, the major components of Type II CRISPR–Cas, the type that is used for genome engineering[83], apparently evolved through two transposon insertion events such that this system seems to consist entirely of transposon-derived genes.

MGEs in vertebrate adaptive immunity

Similar to CRISPR–Cas, the classic vertebrate adaptive immunity also involves genome manipulation — namely, V(D) J recombination (FIG. 1b) that, along with somatic hypermutation, generates the diversity of the T cell receptors (TCRs) and B cell receptors (BCRs). The three segments (V, D and J) of the variable portions of the TCRs and BCRs are each encoded in several dozens of genomic copies. However, V(D)J recombination brings them together in a single exon and, in the process, generates numerous small insertions and deletions at the junctions, creating the enormous combinatorial diversity that is required to match the vast diversity of antigens[84,85]. This process is mediated by the RAG1–RAG2 recombinase complex (FIG. 1b), in which RAG1 is the enzymatically active subunit[86,87], whereas RAG2 acts as a regulatory subunit and is superficially similar to the function of Cas2 in the Cas1–Cas2 duet.

Strikingly, the recombinase domain of RAG1 is derived from the recombinase of animal transposons of the Transib family[56,88]. The small Transib transposons are found among diverse animal species but are absent in vertebrates, and encode a transposase that belongs to a distinct family within the DDE superfamily of transposases and that is homologous to the catalytic core domain of RAG1 (REF. 56). The RAG1 protein has undergone substantial evolution since the proposed recruitment from a Transib transposon, including an amino-terminal fusion with a domain containing a RING finger ubiquitin ligase, but the mechanism of DNA cleavage and rejoining during V(D)J recombination displays a striking similarity to the transposition mechanism[56]. Moreover, the target site duplications (TSDs) generated during the transposition reactions mediated by Transibs and RAG1–RAG2 are similar, and there is significant sequence similarity between the TIRs of Transib and the recombination signal sequences (RSSs) of the immunoglobulin genes (FIG. 4c), indicating that the RSS evolved via Transib insertion[56,88,89].

The tight linkage between the *RAG1* and *RAG2* genes in animal genomes suggests that this gene pair was already present in the ancestral Transib-like transposon, although no such gene combination has been detected in the currently identified transposons[56]. Several variations of the 'RAG transposon' hypothesis for the origin of adaptive immunity in animals have been proposed[55,90–92]. Typically, these scenarios postulate two independent transposition events, whereby insertion of one transposon (the RAG transposon) gave rise to the *RAG1–RAG2* gene pair, whereas insertion of a related non-autonomous element introduced the RSS into an ancestral immunoglobulin gene, which was an element of innate immunity. In such models, two conditions would have to be met. First, the TIRs of the hypothetical non-autonomous transposable elements would have to be identical to those of the RAG transposon in order to be specifically recognized by the RAG1–RAG2 transposase. Second, the TIRs of the RAG transposon would have to be obliterated to ensure the immobilization of the *RAG1–RAG2* gene pair. However, non-autonomous transposons

carrying RSS-like TIRs have not been discovered so far[55]. We propose that a more parsimonious scenario for the origin of V(D)J recombination would involve a single insertion of a fully functional RAG transposon into an ancestral immunoglobulin gene, followed by externalization of the *RAG1–RAG2* gene block while leaving the native RSS-like TIRs within the immunoglobulin gene. An important consequence of the disconnection of the TIRs from the transposon genes is that the TIRs would be immobilized within the genome, thus ensuring stable inheritance of the immune system (FIG. 5).

A general analogy to the scenario of CRISPR–Cas evolution is apparent; in both cases, the potential to form immunological memory, which is the essence of adaptive immunity, is conferred on a pre-existing innate immune system by insertion of a transposon that donates both the recombination sites and the recombinase (FIG. 5). The emergence of vertebrate adaptive immunity from innate immunity mechanisms following the recruitment of the rearrangement machinery (RAG1–RAG2) has been previously discussed[93,94].

Although V(D)J recombination, to the best of the current knowledge, is limited to jawed vertebrates, the discovery of the *RAG1–RAG2* locus in the sea urchin genome implies that this pair of closely linked genes was already present in the genome of the last common ancestor of the Deuterostomes[95]. Moreover, homologues of the RAG1 catalytic domain, which might be evolutionary intermediates between the Transib transposases and RAG1 proteins of Deuterostomes, have been detected in cnidarian genomes[56,96]. The function (or functions) of RAG1–RAG2 before the emergence of adaptive immunity is intriguing: is it possible that there may be additional mechanisms of naturally evolved genome engineering that remain to be discovered?

MGEs and natural genome engineering

Insertion of MGEs into the host genome, by definition, modifies the content of these genomes. Given that MGEs are ubiquitous in cellular life forms and account for the majority of the DNA in some genomes — particularly those of many vertebrates and plants — the consequences of the genome modifications caused by MGEs are diverse and fundamental for cellular organisms[97,98]. It is well known that MGE sequences are often recruited for various cellular functions, typically as regulatory elements and, in some cases, as novel protein-coding sequences[99–101]. Besides the recruitment of MGE sequences for diverse functions, cellular organisms directly exploit the capacity of MGEs to modify the host genome. A major case in point is the catalytic subunit of telomerase, a reverse transcriptase derived from prokaryotic retroelements (group II introns) and responsible for the replication of chromosomal ends (telomeres) in most eukaryotes[102,103]. Remarkably, in organisms that have lost the ancestral telomerase, such as insects, telomere replication is mediated by reverse transcriptases of other retrotransposons[104].

Evolution of adaptive immunity can perhaps be considered as the pinnacle of this strategy that makes exquisite use of the ability of recombinases (transposases or integrases) to insert foreign DNA into specific sites in the host genome. The two best-characterized adaptive immune systems, the prokaryotic CRISPR–Cas system and the immunoglobulin-centred adaptive immunity of jawed vertebrates, seem to have evolved completely independently, yet

through strikingly similar scenarios (FIG. 5). In both cases, the 'executive' part of the system that is responsible for the mechanics of the interaction with the target (the Cascade complex and immunoglobulins) is derived from pre-existing innate immune systems. By contrast, transposons give rise to the 'informational' module that consists of the specific integration sites and the enzymatic machinery of recombination and/or integration. Further research on the molecular mechanisms at play during the CRISPR–Cas action and casposon mobility should provide key details to help to refine the proposed model on the origin of the prokaryotic CRISPR–Cas immunity.

The finding that two unrelated classes of MGEs apparently gave rise to adaptive immune systems in prokaryotes and animals along strikingly similar evolutionary routes suggests that these two systems might not be the only versions of adaptive immunity or, more broadly, of genome manipulation mechanisms that make use of MGE-derived recombinases as naturally evolved devices for genome engineering. Given that genomes of almost all cellular organisms are replete with integrated MGEs, some of which are domesticated and conserved through long evolutionary timespans, it seems unlikely that these two adaptive immune systems are unique. For example, the HARBI1 protein that is conserved across vertebrates is a derivative of the transposase of the widespread Harbinger transposons[105,106]. The function of HARBI1 remains obscure, but a role in a yet unknown genome manipulation pathway cannot be ruled out. Focused searches for novel genome manipulation systems that exploit MGE-encoded recombinases could be a promising research direction.

The high specificity and genome engineering capacity of adaptive immune systems translate into almost unlimited potential for experimental tool development. The utility of antibodies as tools for protein detection is obvious. Recently, the immense promise of CRISPR–Cas has been realized, in particular for Type II systems, in which the transposon-derived Cas9 protein is the only protein required for target recognition and cleavage[83,107,108]. This property of CRISPR–Cas is a direct extension of its extremely high specificity achieved via genome manipulation by an MGE-derived recombinase.

Acknowledgments

References

1. Huda A, Jordan IK. Epigenetic regulation of mammalian genomes by transposable elements. Ann NY Acad Sci. 2009; 1178:276–284. [PubMed: 19845643]

2. Lopez-Flores I, Garrido-Ramos MA. The repetitive DNA content of eukaryotic genomes. Genome Dyn. 2012; 7:1–28. [PubMed: 22759811]

3. Defraia C, Slotkin RK. Analysis of retrotransposon activity in plants. Methods Mol Biol. 2014; 1112:195–210. [PubMed: 24478016]

4. Cortez D, Forterre P, Gribaldo S. A hidden reservoir of integrative elements is the major source of recently acquired foreign genes and ORFans in archaeal and bacterial genomes. Genome Biol. 2009; 10:R65. [PubMed: 19531232]

5. Makarova KS, et al. Dark matter in archaeal genomes: a rich source of novel mobile elements, defense systems and secretory complexes. Extremophiles. 2014; 18:877–893. [PubMed: 25113822]

6. Casjens S. Prophages and bacterial genomics: what have we learned so far? Mol Microbiol. 2003; 49:277–300. [PubMed: 12886937]

7. Busby B, Kristensen DM, Koonin EV. Contribution of phage-derived genomic islands to the virulence of facultative bacterial pathogens. Environ Microbiol. 2013; 15:307–312. [PubMed: 23035931]

8. Wicker T, et al. A unified classification system for eukaryotic transposable elements. Nature Rev Genet. 2007; 8:973–982. [PubMed: 17984973]

9. Kapitonov VV, Jurka J. A universal classification of eukaryotic transposable elements implemented in Repbase. Nature Rev Genet. 2008; 9:411–412. [PubMed: 18421312]

10. Jurka J, Kapitonov VV, Kohany O, Jurka MV. Repetitive sequences in complex genomes: structure and evolution. Annu Rev Genomics Hum Genet. 2007; 8:241–259. [PubMed: 17506661]

11. Curcio MJ, Derbyshire KM. The outs and ins of transposition: from mu to kangaroo. Nature Rev Mol Cell Biol. 2003; 4:865–877. [PubMed: 14682279]

12. Chandler M, et al. Breaking and joining single-stranded DNA: the HUH endonuclease superfamily. Nature Rev Microbiol. 2013; 11:525–538. [PubMed: 23832240]

13. Ilyina TV, Koonin EV. Conserved sequence motifs in the initiator proteins for rolling circle DNA replication encoded by diverse replicons from eubacteria, eucaryotes and archaebacteria. Nucleic Acids Res. 1992; 20:3279–3285. [PubMed: 1630899]

14. Krupovic M. Networks of evolutionary interactions underlying the polyphyletic origin of ssDNA viruses. Curr Opin Virol. 2013; 3:578–586. [PubMed: 23850154]

15. Goodwin TJ, Poulter RT. A new group of tyrosine recombinase-encoding retrotransposons. Mol Biol Evol. 2004; 21:746–759. [PubMed: 14963102]

16. Boocock MR, Rice PA. A proposed mechanism for IS607-family serine transposases. Mob DNA. 2013; 4:24. [PubMed: 24195768]

17. Weichenrieder O, Repanas K, Perrakis A. Crystal structure of the targeting endonuclease of the human LINE-1 retrotransposon. Structure. 2004; 12:975–986. [PubMed: 15274918]

18. Aswad A, Katzourakis A. Paleovirology and virally derived immunity. Trends Ecol Evol. 2012; 27:627–636. [PubMed: 22901901]

19. Feschotte C, Gilbert C. Endogenous viruses: insights into viral evolution and impact on host biology. Nature Rev Genet. 2012; 13:283–296. [PubMed: 22421730]

20. Duggal NK, Emerman M. Evolutionary conflicts between viruses and restriction factors shape immunity. Nature Rev Immunol. 2012; 12:687–695. [PubMed: 22976433]

21. Forterre P, Prangishvili D. The major role of viruses in cellular evolution: facts and hypotheses. Curr Opin Virol. 2013; 3:558–565. [PubMed: 23870799]

22. Koonin EV, Dolja VV. A virocentric perspective on the evolution of life. Curr Opin Virol. 2013; 3:546–557. [PubMed: 23850169]

23. Labrie SJ, Samson JE, Moineau S. Bacteriophage resistance mechanisms. Nature Rev Microbiol. 2010; 8:317–327. [PubMed: 20348932]

24. Seed KD, Lazinski DW, Calderwood SB, Camilli A. A bacteriophage encodes its own CRISPR/Cas adaptive response to evade host innate immunity. Nature. 2013; 494:489–491. [PubMed: 23446421]

25. Boehm T. Evolution of vertebrate immunity. Curr Biol. 2012; 22:R722–732. [PubMed: 22975003]

26. Rimer J, Cohen IR, Friedman N. Do all creatures possess an acquired immune system of some sort? Bioessays. 2014; 36:273–281. [PubMed: 24421210]

27. Akira S, Uematsu S, Takeuchi O. Pathogen recognition and innate immunity. Cell. 2006; 124:783–801. [PubMed: 16497588]

28. Makarova KS, Grishin NV, Shabalina SA, Wolf YI, Koonin EV. A putative RNA-interference-based immune system in prokaryotes: computational analysis of the predicted enzymatic machinery, functional analogies with eukaryotic RNAi, and hypothetical mechanisms of action. Biol Direct. 2006; 1:7. [PubMed: 16545108]

29. Makarova KS, Wolf YI, Koonin EV. Comparative genomics of defense systems in archaea and bacteria. Nucleic Acids Res. 2013; 41:4360–4377. [PubMed: 23470997]

30. Hur JK, Olovnikov I, Aravin AA. Prokaryotic Argonautes defend genomes against invasive DNA. Trends Biochem Sci. 2014; 39:257–259. [PubMed: 24836995]

31. Swarts DC, et al. The evolutionary journey of Argonaute proteins. Nature Struct Mol Biol. 2014; 21:743–753. [PubMed: 25192263]

32. Makarova KS, et al. Evolution and classification of the CRISPR–Cas systems. Nature Rev Microbiol. 2011; 9:467–477. [PubMed: 21552286]

33. van der Oost J, Jore MM, Westra ER, Lundgren M, Brouns SJ. CRISPR-based adaptive and heritable immunity in prokaryotes. Trends Biochem Sci. 2009; 34:401–407. [PubMed: 19646880]

34. Wiedenheft B, Sternberg SH, Doudna JA. RNA-guided genetic silencing systems in bacteria and archaea. Nature. 2012; 482:331–338. [PubMed: 22337052]

35. Sorek R, Lawrence CM, Wiedenheft B. CRISPR-mediated adaptive immune systems in bacteria and archaea. Annu Rev Biochem. 2013; 82:237–266. [PubMed: 23495939]

36. Barrangou R, Marraffini LA. CRISPR–Cas systems: prokaryotes upgrade to adaptive immunity. Mol Cell. 2014; 54:234–244. [PubMed: 24766887]

37. Manica A, Schleper C. CRISPR-mediated defense mechanisms in the hyperthermophilic archaeal genus Sulfolobus. RNA Biol. 2013; 10:671–678. [PubMed: 23535277]

38. van der Oost J, Westra ER, Jackson RN, Wiedenheft B. Unravelling the structural and mechanistic basis of CRISPR–Cas systems. Nature Rev Microbiol. 2014; 12:479–492. [PubMed: 24909109]

39. Weinberger AD, et al. Persisting viral sequences shape microbial CRISPR-based immunity. PLoS Comput Biol. 2012; 8:e1002475. [PubMed: 22532794]

40. Koonin EV, Wolf YI. Is evolution Darwinian or/and Lamarckian? Biol Direct. 2009; 4:42. [PubMed: 19906303]

41. Wang XH, et al. RNA interference directs innate immunity against viruses in adult *Drosophila*. Science. 2006; 312:452–454. [PubMed: 16556799]

42. Shabalina SA, Koonin EV. Origins and evolution of eukaryotic RNA interference. Trends Ecol Evol. 2008; 23:578–587. [PubMed: 18715673]

43. Carthew RW, Sontheimer EJ. Origins and mechanisms of miRNAs and siRNAs. Cell. 2009; 136:642–655. [PubMed: 19239886]

44. Zhou R, Rana TM. RNA-based mechanisms regulating host–virus interactions. Immunol Rev. 2013; 253:97–111. [PubMed: 23550641]

45. Medzhitov R. Approaching the asymptote: 20 years later. Immunity. 2009; 30:766–775. [PubMed: 19538928]

46. Kawai T, Akira S. The role of pattern-recognition receptors in innate immunity: update on Toll-like receptors. Nature Immunol. 2010; 11:373–384. [PubMed: 20404851]

47. Malmgaard L. Induction and regulation of IFNs during viral infections. J Interferon Cytokine Res. 2004; 24:439–454. [PubMed: 15320958]

48. Le Page C, Genin P, Baines MG, Hiscott J. Interferon activation and innate immunity. Rev Immunogenet. 2000; 2:374–386. [PubMed: 11256746]

49. Cooper MD, Alder MN. The evolution of adaptive immune systems. Cell. 2006; 124:815–822. [PubMed: 16497590]

50. Boehm T. Design principles of adaptive immune systems. Nature Rev Immunol. 2011; 11:307–317. [PubMed: 21475308]

51. Davis MM. The evolutionary and structural 'logic' of antigen receptor diversity. Semin Immunol. 2004; 16:239–243. [PubMed: 15522622]

52. Cannon JP, Haire RN, Rast JP, Litman GW. The phylogenetic origins of the antigen-binding receptors and somatic diversification mechanisms. Immunol Rev. 2004; 200:12–22. [PubMed: 15242392]

53. Market E, Papavasiliou FN. V(D)J recombination and the evolution of the adaptive immune system. PLoS Biol. 2003; 1:e16. [PubMed: 14551913]

54. Jung D, Alt FW. Unraveling V(D)J recombination; insights into gene regulation. Cell. 2004; 116:299–311. [PubMed: 14744439]

55. Fugmann SD. The origins of the *Rag* genes — from transposition to V(D)J recombination. Semin Immunol. 2010; 22:10–16. [PubMed: 20004590]

56. Kapitonov VV, Jurka J. RAG1 core and V(D)J recombination signal sequences were derived from Transib transposons. PLoS Biol. 2005; 3:e181. [PubMed: 15898832]

57. Krupovic M, Makarova KS, Forterre P, Prangishvili D, Koonin EV. Casposons: a new superfamily of self-synthesizing DNA transposons at the origin of prokaryotic CRISPR–Cas immunity. BMC Biol. 2014; 12:36. [PubMed: 24884953]

58. Nunez JK, et al. Cas1–Cas2 complex formation mediates spacer acquisition during CRISPR–Cas adaptive immunity. Nature Struct Mol Biol. 2014; 21:528–534. [PubMed: 24793649]

59. Wiedenheft B, et al. Structural basis for DNase activity of a conserved protein implicated in CRISPR-mediated genome defense. Structure. 2009; 17:904–912. [PubMed: 19523907]

60. Makarova KS, Aravind L, Wolf YI, Koonin EV. Unification of Cas protein families and a simple scenario for the origin and evolution of CRISPR–Cas systems. Biol Direct. 2011; 6:38. [PubMed: 21756346]

61. Makarova KS, Wolf YI, Koonin EV. The basic building blocks and evolution of CRISPR–Cas systems. Biochem Soc Trans. 2013; 41:1392–1400. [PubMed: 24256226]

62. Brouns SJ, et al. Small CRISPR RNAs guide antiviral defense in prokaryotes. Science. 2008; 321:960–964. [PubMed: 18703739]

63. Rouillon C, et al. Structure of the CRISPR interference complex CSM reveals key similarities with cascade. Mol Cell. 2013; 52:124–134. [PubMed: 24119402]

64. Anantharaman V, Koonin EV, Aravind L. Comparative genomics and evolution of proteins involved in RNA metabolism. Nucleic Acids Res. 2002; 30:1427–1464. [PubMed: 11917006]

65. Anantharaman V, Aravind L, Koonin EV. Emergence of diverse biochemical activities in evolutionarily conserved structural scaffolds of proteins. Curr Opin Chem Biol. 2003; 7:12–20. [PubMed: 12547421]

66. Clery A, Blatter M, Allain FH. RNA recognition motifs: boring? Not quite. Curr Opin Struct Biol. 2008; 18:290–298. [PubMed: 18515081]

67. Koonin EV, Makarova KS. CRISPR–Cas: evolution of an RNA-based adaptive immunity system in prokaryotes. RNA Biol. 2013; 10:679–686. [PubMed: 23439366]

68. Makarova KS, Wolf YI, van der Oost J, Koonin EV. Prokaryotic homologs of Argonaute proteins are predicted to function as key components of a novel system of defense against mobile genetic elements. Biol Direct. 2009; 4:29. [PubMed: 19706170]

69. Olovnikov I, Chan K, Sachidanandam R, Newman DK, Aravin AA. Bacterial Argonaute samples the transcriptome to identify foreign DNA. Mol Cell. 2013; 51:594–605. [PubMed: 24034694]

70. Swarts DC, et al. DNA-guided DNA interference by a prokaryotic Argonaute. Nature. 2014; 507:258–261. [PubMed: 24531762]

71. Beloglazova N, et al. A novel family of sequence-specific endoribonucleases associated with the clustered regularly interspaced short palindromic repeats. J Biol Chem. 2008; 283:20361–20371. [PubMed: 18482976]

72. Han D, Krauss G. Characterization of the endonuclease SSO2001 from *Sulfolobus solfataricus* P2. FEBS Lett. 2009; 583:771–776. [PubMed: 19174159]

73. Makarova KS, Anantharaman V, Aravind L, Koonin EV. Live virus-free or die: coupling of antivirus immunity and programmed suicide or dormancy in prokaryotes. Biol Direct. 2012; 7:40. [PubMed: 23151069]

74. Kwon AR, et al. Structural and biochemical characterization of HP0315 from *Helicobacter pylori* as a VapD protein with an endoribonuclease activity. Nucleic Acids Res. 2012; 40:4216–4228. [PubMed: 22241770]

75. Unterholzner SJ, Poppenberger B, Rozhon W. Toxin–antitoxin systems: biology, identification, and application. Mob Genet Elements. 2013; 3:e26219. [PubMed: 24251069]

76. Krupovic M, Gonnet M, Hania WB, Forterre P, Erauso G. Insights into dynamics of mobile genetic elements in hyperthermophilic environments from five new *Thermococcus* plasmids. PLoS ONE. 2013; 8:e49044. [PubMed: 23326305]

77. Kunin V, Sorek R, Hugenholtz P. Evolutionary conservation of sequence and secondary structures in CRISPR repeats. Genome Biol. 2007; 8:R61. [PubMed: 17442114]

78. Datsenko KA, et al. Molecular memory of prior infections activates the CRISPR–Cas adaptive bacterial immunity system. Nature Commun. 2012; 3:945. [PubMed: 22781758]

79. Yosef I, Goren MG, Qimron U. Proteins and DNA elements essential for the CRISPR adaptation process in *Escherichia coli*. Nucleic Acids Res. 2012; 40:5569–5576. [PubMed: 22402487]

80. Chylinski K, Le Rhun A, Charpentier E. The tracrRNA and Cas9 families of typeII CRISPR–Cas immunity systems. RNA Biol. 2013; 10:726–737. [PubMed: 23563642]

81. Chylinski K, Makarova KS, Charpentier E, Koonin EV. Classification and evolution of typeII CRISPR–Cas systems. Nucleic Acids Res. 2014; 42:6091–6105. [PubMed: 24728998]

82. Bao W, Jurka J. Homologues of bacterial TnpB_IS605 are widespread in diverse eukaryotic transposable elements. Mob DNA. 2013; 4:12. [PubMed: 23548000]

83. Kim H, Kim JS. A guide to genome engineering with programmable nucleases. Nature Rev Genet. 2014; 15:321–334. [PubMed: 24690881]

84. Tonegawa S. Somatic generation of antibody diversity. Nature. 1983; 302:575–581. [PubMed: 6300689]

85. Papavasiliou FN, Schatz DG. Somatic hypermutation of immunoglobulin genes: merging mechanisms for genetic diversity. Cell. 2002; 109:S35–S44. [PubMed: 11983151]

86. Oettinger MA, Schatz DG, Gorka C, Baltimore D. *RAG-1* and *RAG-2*, adjacent genes that synergistically activate V(D)J recombination. Science. 1990; 248:1517–1523. [PubMed: 2360047]

87. Swanson PC. The bounty of RAGs: recombination signal complexes and reaction outcomes. Immunol Rev. 2004; 200:90–114. [PubMed: 15242399]

88. Panchin Y, Moroz LL. Molluscan mobile elements similar to the vertebrate recombination-activating genes. Biochem Biophys Res Commun. 2008; 369:818–823. [PubMed: 18313399]

89. Sakano H, Huppi K, Heinrich G, Tonegawa S. Sequences at the somatic recombination sites of immunoglobulin light-chain genes. Nature. 1979; 280:288–294. [PubMed: 111144]

90. Thompson CB. New insights into V(D)J recombination and its role in the evolution of the immune system. Immunity. 1995; 3:531–539. [PubMed: 7584143]

91. Roth DB, Craig NL. VDJ recombination: a transposase goes to work. Cell. 1998; 94:411–414. [PubMed: 9727482]

92. Schatz DG. Antigen receptor genes and the evolution of a recombinase. Semin Immunol. 2004; 16:245–256. [PubMed: 15522623]

93. Du Pasquier L. Speculations on the origin of the vertebrate immune system. Immunol Lett. 2004; 92:3–9. [PubMed: 15081520]

94. Du Pasquier L. Innate immunity in early chordates and the appearance of adaptive immunity. C R Biol. 2004; 327:591–601. [PubMed: 15330258]

95. Fugmann SD, Messier C, Novack LA, Cameron RA, Rast JP. An ancient evolutionary origin of the *Rag1/2* gene locus. Proc Natl Acad Sci USA. 2006; 103:3728–3733. [PubMed: 16505374]

96. Hemmrich G, Miller DJ, Bosch TC. The evolution of immunity: a low-life perspective. Trends Immunol. 2007; 28:449–454. [PubMed: 17855167]

97. Bowen NJ, Jordan IK. Transposable elements and the evolution of eukaryotic complexity. Curr Issues Mol Biol. 2002; 4:65–76. [PubMed: 12074196]

98. Kazazian HH Jr. Mobile elements: drivers of genome evolution. Science. 2004; 303:1626–1632. [PubMed: 15016989]

99. Jordan IK, Rogozin IB, Glazko GV, Koonin EV. Origin of a substantial fraction of human regulatory sequences from transposable elements. Trends Genet. 2003; 19:68–72. [PubMed: 12547512]

100. Conley AB, Piriyapongsa J, Jordan IK. Retroviral promoters in the human genome. Bioinformatics. 2008; 24:1563–1567. [PubMed: 18535086]

101. Jurka J. Conserved eukaryotic transposable elements and the evolution of gene regulation. Cell Mol Life Sci. 2008; 65:201–204. [PubMed: 18030428]

102. Nakamura TM, Cech TR. Reversing time: origin of telomerase. Cell. 1998; 92:587–590. [PubMed: 9506510]

103. Gladyshev EA, Arkhipova IR. A widespread class of reverse transcriptase-related cellular genes. Proc Natl Acad Sci USA. 2011; 108:20311–20316. [PubMed: 21876125]

104. Pardue ML, De Baryshe PG. Retrotransposons provide an evolutionarily robust non-telomerase mechanism to maintain telomeres. Annu Rev Genet. 2003; 37:485–511. [PubMed: 14616071]

105. Kapitonov VV, Jurka J. Harbinger transposons and an ancient *HARBI1* gene derived from a transposase. DNA Cell Biol. 2004; 23:311–324. [PubMed: 15169610]

106. Sinzelle L, et al. Transposition of a reconstructed Harbinger element in human cells and functional homology with two transposon-derived cellular genes. Proc Natl Acad Sci USA. 2008; 105:4715–4720. [PubMed: 18339812]

107. Mali P, et al. RNA-guided human genome engineering via Cas9. Science. 2013; 339:823–826. [PubMed: 23287722]

108. Sternberg SH, Redding S, Jinek M, Greene EC, Doudna JA. DNA interrogation by the CRISPR RNA-guided endonuclease Cas9. Nature. 2014; 507:62–67. [PubMed: 24476820]
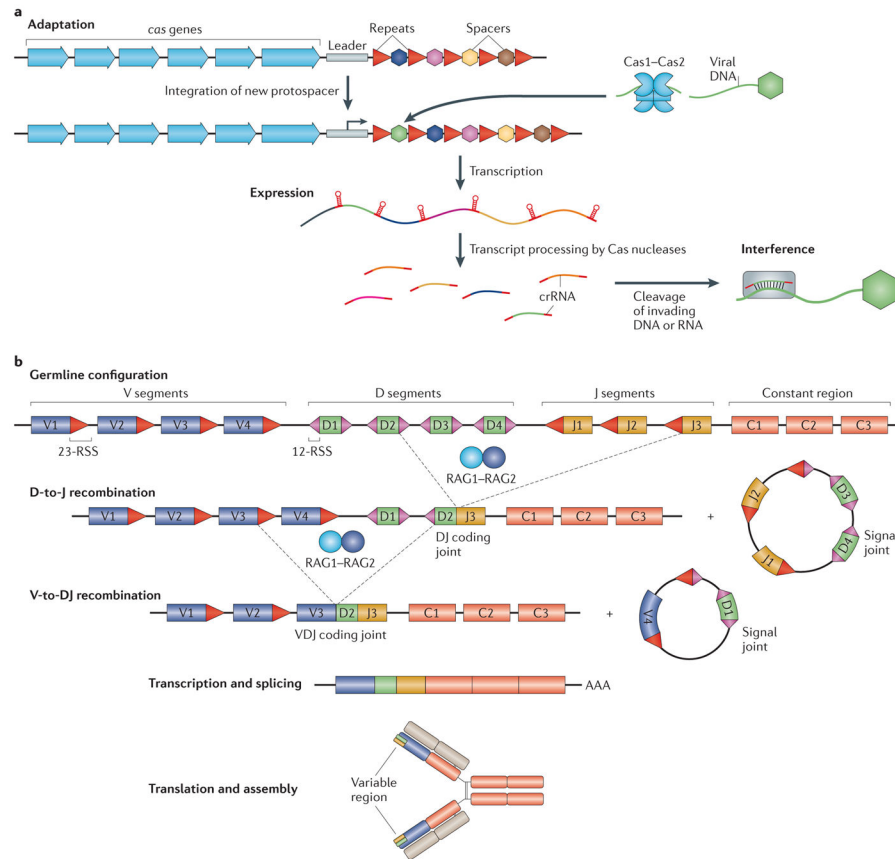
Figure 1. Adaptive immune systems of prokaryotes and eukaryotes

a | The prokaryotic clustered regularly interspaced short palindromic repeat–CRISPR-associated protein (CRISPR–Cas) locus consists of *cas* genes (blue arrows) that encode different Cas proteins, and CRISPR arrays composed of variable spacers (coloured hexagons) interspersed with direct repeats (red triangles). The leader sequence (grey rectangle) contains a promoter for the transcription of the CRISPR array and marks the end where new spacers are incorporated. Three stages of CRISPR–Cas immunity are depicted. During the adaptation stage, a Cas1–Cas2 heterohexamer uptakes a protospacer from the invading plasmid or viral DNA (green) and incorporates it at the leader-proximal end of the CRISPR array. During the expression stage, the CRISPR array is transcribed, and the transcript is processed into small CRISPR RNAs (crRNAs) by different Cas nucleases in a CRISPR–Cas type-dependent manner. During the interference stage, crRNAs act as guides for the cleavage of invading viral or plasmid DNA or RNA that contains regions complementary to the crRNA. **b** | Lymphocyte antigen receptor diversification by V(D)J recombination is shown. The variable region of the immunoglobulin heavy chain is assembled by V(D)J recombination from V (variable; purple rectangle), D (diversity; green rectangle) and J (joining; brown rectangle) gene segments. The immunoglobulin light chain is assembled from V and J segments by VJ recombination (not shown). Multiple V, D, J and C (constant region; red rectangles) gene segments are available for recombination in the germline genome. The recombination is carried out by the RAG1–RAG2 recombinase complex and involves two types of recombination signal sequences (RSSs), 23-RSS (red

triangles) and 12-RSS (pink triangles), which flank each gene segment. Joining of the DNA ends requires non-homologous end-joining (NHEJ) proteins (not shown). Two rounds of recombination, D to J and V to DJ, produce a VDJ coding joint and two circular molecules (signal joints); the latter do not have any further role and are discarded. Transcription across the VDJ coding joint, followed by splicing, produces the mature transcript of the immunoglobulin heavy chain. Subsequent translation of the transcript, assembly of the heavy chain and association with the light chain (beige rectangles) complete the assembly of the immunoglobulin receptor.
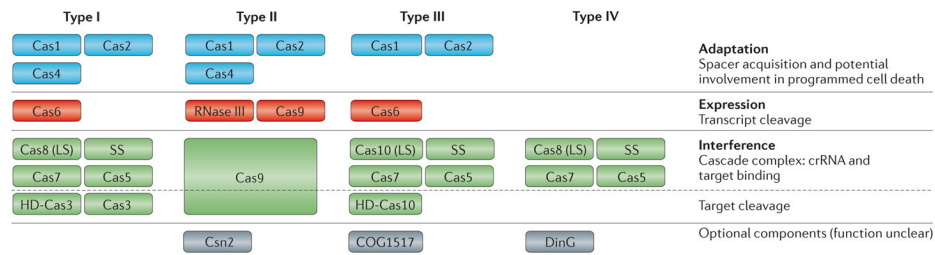
Figure 2. A general scheme of the organization of CRISPR–Cas systems

Protein names follow the current nomenclature and classification[32]. The general functions and the stages of the clustered regularly interspaced short palindromic repeat–CRISPR-associated protein (CRISPR–Cas) immunity are shown on the right; the corresponding proteins in each type of CRISPR–Cas system are shown on the left and are colour coded. Cas9 of Type II CRISPR Cas is a multifunctional protein involved in several stages of the immune response, including processing of the primary CRISPR transcript into CRISPR RNAs (crRNAs), target binding and target cleavage. Similarly, in Type I and Type III CRISPR–Cas systems, Cas6 is a subunit of the Cascade (CRISPR-associated complex for antiviral defence) complex that is involved in both pre-crRNA processing, as well as target recognition and inactivation. Note that RNase III, which participates in cleavage of Type II CRISPR transcripts, has other roles in the processing of cellular RNA, particularly ribosomal RNA. Csn2 is predicted to be functionally analogous (but not homologous) to Cas4 and participates in spacer acquisition[33]. HD, histidine–aspartate family nuclease; LS, large subunit; SS, small subunit.
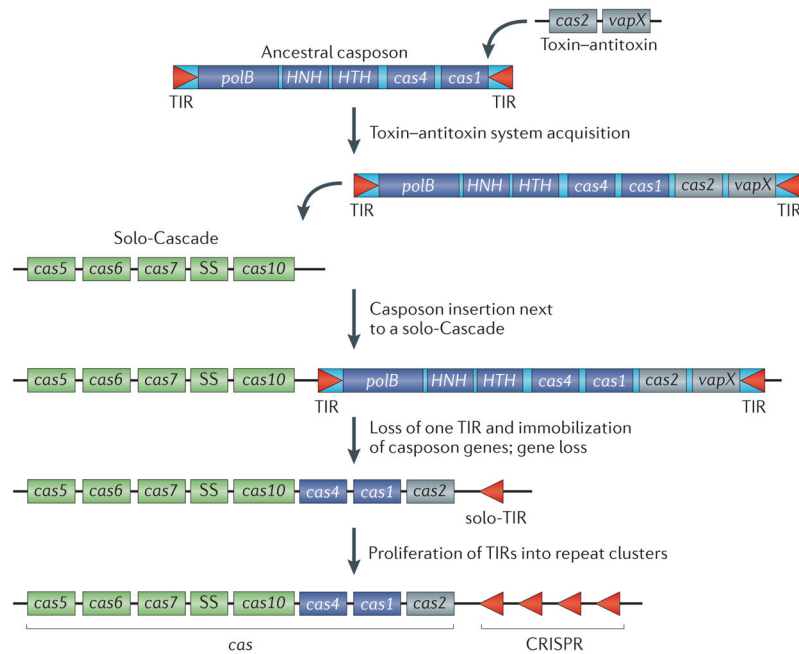
Figure 3. A scenario for the evolution of the CRISPR–Cas system from a casposon, a toxin–antitoxin module and a solo-Cascade innate immune system

Casposon-derived genes are shown as dark blue rectangles, toxin–antitoxin genes are depicted in grey and 'solo-Cascade' (CRISPR-associated complex for antiviral defence) genes are shown in green. A generic organization of a Type III Cascade operon is shown that does not depict any particular genomic locus. Most of the Cascade genes encode proteins that are distinct arrangements of one or two RNA recognition motif (RRM) domains and that might have evolved from a simple double-RRM protein through a series of RRM domain duplications and a fusion with a histidine–aspartate nuclease domain in Cas10 (REF. 61). Terminal inverted repeats (TIRs) are palindromic, which is reminiscent of the CRISPR unit. *polB*, family B DNA polymerase; SS, small subunit.

Figure 4. Comparison between TIR, CRISPR and RSS

a | Schematic organization of the casposon from *Aciduliprofundum boonei* T469 (NC_013926, nucleotide coordinates: 380320–389403) is shown at the top, whereas the clustered regularly interspaced short palindromic repeat–CRISPR-associated protein (CRISPR–Cas) system of *Thermotoga thermarum* DSM 5069 (NC_015707, nucleotide coordinates: 1706198–1717565) is shown at the bottom. *cas* genes are colour-coded according to the scheme provided in FIG. 2. The alignment of the corresponding casposon terminal inverted repeat (TIR) and CRISPR sequence is shown in the middle. Identical nucleotides are indicated by the black background. **b** | Predicted secondary structures of the *A. boonei* casposon TIR (left) and *T. thermarum* CRISPR repeat (right) are shown. **c** | Comparison between the Transib TIRs and recombination signal sequences (RSSs) is shown. The Transib5 transposon from *Drosophila melanogaster* (top) is flanked by TIRs that consist of conserved heptamer and nonamer sequences separated by a variable spacer of either 13 bp (pink triangle) or 23 bp (red triangle). Sequence alignment of the Transib5 TIRs and the consensus recombination recognition sequence (RSS) is depicted. The variable spacers in RSSs are marked by 'n'. The most conserved nucleotides in the RSS heptamer and nonamer, which are necessary for efficient V(D)J recombination, are highlighted by the red background. The RSS and TIR sequences data are derived from REF. 56. *HD*, histidine–aspartate family nuclease; *polB*, family B DNA polymerase.

Figure 5. Comparison of the proposed evolutionary paths to the prokaryotic and eukaryotic versions of adaptive immunity

Terminal inverted repeats (TIRs) and recombination signal sequences (RSSs) are depicted as triangles. In the case of V(D)J recombination, TIRs and RSSs consist of conserved heptamer and nonamer sequences separated by a variable spacer of either 12 bp (pink triangles) or 23 bp (red triangles). V and J represent variable and joining gene segments of the immunoglobulin (Ig) gene, respectively. The dashed line indicates that RAG1 and RAG2 proteins are not encoded in proximity of the cognate recombination sites. CRISPR–Cas, clustered regularly interspaced short palindromic repeat–CRISPR-associated protein.