Published in final edited form as:

Methods Mol Biol. 2015 ; 1311: 47–75. doi:10.1007/978-1-4939-2687-9_4.

Annotation and Classification of CRISPR-Cas Systems

Kira S. Makarova and **Eugene V. Koonin**

Abstract

The clustered regularly interspaced short palindromic repeats (CRISPR)-Cas (CRISPR-associated proteins) is a prokaryotic adaptive immune system that is represented in most archaea and many bacteria. Among the currently known prokaryotic defense systems, the CRISPR-Cas genomic loci show unprecedented complexity and diversity. Classification of CRISPR-Cas variants that would capture their evolutionary relationships to the maximum possible extent is essential for comparative genomic and functional characterization of this theoretically and practically important system of adaptive immunity. To this end, a multipronged approach has been developed that combines phylogenetic analysis of the conserved Cas proteins with comparison of gene repertoires and arrangements in CRISPR-Cas loci. This approach led to the current classification of CRISPR-Cas systems into three distinct types and ten subtypes for each of which signature genes have been identified. Comparative genomic analysis of the CRISPR-Cas systems in new archaeal and bacterial genomes performed over the 3 years elapsed since the development of this classification makes it clear that new types and subtypes of CRISPR-Cas need to be introduced. Moreover, this classification system captures only part of the complexity of CRISPR-Cas organization and evolution, due to the intrinsic modularity and evolutionary mobility of these immunity systems, resulting in numerous recombinant variants. Moreover, most of the *cas* genes evolve rapidly, complicating the family assignment for many Cas proteins and the use of family profiles for the recognition of CRISPR-Cas subtype signatures. Further progress in the comparative analysis of CRISPR-Cas systems requires integration of the most sensitive sequence comparison tools, protein structure comparison, and refined approaches for comparison of gene neighborhoods.

Keywords

CRISPR-Cas classification; CRISPR-Cas annotation; CRISPR-Cas evolution; CRISPR; Cas bioinformatics; Cas1; RAMPs

1 Introduction

The clustered regularly interspaced short palindromic repeats (CRISPR)-Cas (CRISPR-associated proteins) modules are adaptive antivirus immunity systems that are present in most archaea and many bacteria and function on the self-nonself discrimination principle [1]. These systems incorporate fragments of alien DNA (known as spacers) into CRISPR cassettes, then transcribe the CRISPR arrays including the spacers, and process them to make a guide crRNA (CRISPR RNA) which is employed to specifically target and cleave the genome of the cognate virus or plasmid [2–5]. Numerous, highly diverse Cas (CRISPR-associated) proteins are involved in different steps of the processing of CRISPR loci transcripts, cleavage of the target DNA or RNA, and new spacer integration [5–7].

The action of the CRISPR-Cas system is usually divided into three stages: (1) adaptation or spacer integration, (2) processing of the primary transcript of the CRISPR locus (pre-crRNA) and maturation of the crRNA which includes the spacer and variable regions corresponding to 5′ and 3′ fragments of CRISPR repeats, and (3) DNA (or RNA) interference [3, 8, 9]. Two proteins, Cas1 and Cas2, that are present in the great majority of the known CRISPR-Cas systems are sufficient for the insertion of spacers into the CRISPR cassettes [10]. These two proteins form a complex that is required for this adaptation process; the endonuclease activity of Cas1 is required for spacer integration whereas Cas2 appears to perform a nonenzymatic function [11, 12]. The Cas1-Cas2 complex represents the highly conserved "information processing" module of CRISPR-Cas that appears to be quasi-autonomous from the rest of the system (see below).

The second stage, the processing of pre-crRNA into the guide crRNAs, is performed either by a dedicated RNA endonuclease complex or via an alternative mechanism that involves bacterial RNase III and an additional RNA species [13]. The mature crRNA is bound by one (type II) or several (types I and III) Cas proteins that form the effector complex, which targets the cognate DNA or RNA [14–19]. The effector complex of type I systems is known as Cascade (CRISPR-associated complex for antiviral defense) [20].

Because of the enormous diversity of CRISPR-Cas, classification of these systems and consistent annotation of the Cas proteins are major challenges [5]. Considering the complexity of the composition and architecture of the CRISPR-Cas systems and the infeasibility of a single classification criterion, a "polythetic" approach based on a combination of evidence from phylogenetic, comparative genomic, and structural analysis has been proposed [5]. Three major types of CRISPR-Cas systems are at the top of the classification hierarchy. The three types are readily distinguishable by virtue of the presence of three unique signature genes: Cas3 in type I systems, Cas9 in type II, and Cas10 in type III [5]. With several exceptions, all three CRISPR-Cas types contain full complements of components that are required for the key steps of the defense mechanism. Recently, thanks to in-depth sequence analysis and structure of the effector complexes from different variants of CRISPR-Cas systems, common principles of organization and function of the complexes have been uncovered allowing for further generalization of the CRISPR-Cas classification, at least for the systems of type I and type III [21–24].

In this chapter we present an overview of the approaches, methods, and further challenges of CRISPR-Cas subtype classification and Cas protein nomenclature taking into account recent advances in the understanding of the mechanisms and organization of CRISPR-Cas systems.

2 Phylogenomic Analysis and Classification of CRISPR-Cas Systems

2.1 Annotation of cas Genes by Comparative Analysis of the Encoded Protein Sequences

The sequences of most Cas proteins, with only a few exceptions, such as Cas1 and Cas3, are highly diverged, presumably owing to the fast evolution that is typical of defense systems. Accordingly, classification of *cas* genes on the basis of protein sequence conservation is a nontrivial task that requires careful application of the most sensitive available sequence analysis methods that typically compare frequency profiles (position-specific scoring matrices, PSSM) generated from multiple alignments of the analyzed protein families, rather than individual sequences. The most complete available set of PSSMs corresponding to the latest accepted nomenclature and classification of the *cas* gene [5] is currently available through the CDD database [25]. The list of these PSSMs and the correspondence between the CDD PSSMs and the respective Pfam and TIGR families also can be found in [5] and at ftp://ftp.ncbi.nih.gov/pub/wolf/_suppl/CRISPRclass/crisprPro.html. Several servers, widely used for sequence similarity searches, such as HHpred [26], include mirrors of the CDD database and thus can also be used for Cas protein annotation. Usually, PSI-BLAST [27] or HHpred [26] is used to identify similarity of a sequence (e.g., the sequence of a particular Cas protein) to a library of PSSMs generated from the respective multiple alignments. Different Cas profiles and different programs vary with respect to the sensitivity and selectivity when used for searches with the aforementioned programs with default parameters.

Table 1 provides information on those signature *cas* genes for individual subtypes that can be reliably identified by any program. Unfortunately, search for similarity with many other Cas families, including several signature genes, may result in a considerable number of false positives and false negatives. Specifically, such proteins as Cas3′, Cas3″, Cas10, Cas4, and Cas9 can display similarity to related proteins or domains that are not linked to CRISPR-Cas system, e.g., diverse helicases in the case of Cas3 and various polymerases and cyclases in the case of Cas10. Furthermore, many RNA recognition motif-containing proteins that function as Cas effector complex subunits are similar to each other despite being present in the loci for distinct CRISPR-Cas subtypes (*see* Subheading 2.6.2). By contrast, profiles for the small and large Cas effector complex subunits are not sensitive enough due to extreme divergence of these proteins even within a single subtype.

In order to increase sensitivity of the searches for these families, distinct profiles for each subfamily have to be generated. For this purpose, sequences that are not recognized as known *cas* genes but are present in the CRISPR-Cas loci have to be clustered using a clustering method such as BLASTCLUST [28] and then aligned using an appropriate multiple alignment. Alternatively or additionally, these sequences can be used as queries for similarity searches using PSI-BLAST and the closely related homologs found in this search can also be aligned to generate profiles that are then used as queries for a more sensitive sequence similarity search method, such as HHpred, to detect potential remote sequence similarity with known Cas families.

2.2 Classification of CRISPR-Cas Systems

Considering technical problems with the sensitivity and selectivity of Cas protein family profiles and uncertainties of Cas1 phylogeny and CRISPR-Cas subtype classification described below, fully automated identification of CRISPR-Cas subtypes in general is not currently feasible. The best approach to ensure the correct classification is to combine several sources of information such as Cas1 phylogeny, identification, and annotation of as many Cas proteins as possible in the locus in question and, for type II systems, identification of the trans-activating crRNA (tracrRNA) genes [29]. Table 1 provides a description of the key features of each subtype that help in the classification of the CRISPR-Cas systems.

Extra caution should be exercised when introducing new gene names and new subtypes, because, due to the often extreme sequence divergence of the Cas proteins, the similarity with already defined genes and subtypes can easily be overlooked. Furthermore, the abundance of associated genes that are likely to represent (quasi)independent immunity mechanism and are only loosely linked to CRISPR-Cas loci requires extra evidence to assign new *cas* names [30–32].

2.3 A Brief History of CRISPR-Cas System Classification and cas Gene Nomenclature

The original bioinformatic analysis that linked the CRISPR repeats and *cas* genes proposed four names for the most conserved and abundant *cas* genes and their products: Cas1, Cas2, Cas3, and Cas4 [33]. Subsequent analyses of proteins associated with these systems have shown that the genomes of various CRISPR- containing organisms encode approximately 65 distinct sets of orthologous Cas proteins which can be classified into either 23–45 families depending on the classification criteria (granularity of clustering) [6, 7]. Two additional core *cas* gene names were introduced at this stage, namely *cas5* and *cas6* [6].

Cas1 is the most conserved protein that is present in most of the CRISPR-Cas systems and evolves slower than other Cas proteins [34]. Accordingly, Cas1 phylogeny has been used as the guide for CRISPR-Cas system classification. Distinct operon organization of the CRISPR-Cas models in the genomes also was recognized as an important additional classification criterion [6, 7]. Eight distinct subtypes were originally proposed and named after the species whose genomes encoded a typical system of each subtype: Ecoli, Ypest, Nmeni, Dvulg, Tneap, Hmari, Apern, and Mtube [6]. In addition, the RAMP module (named after several proteins from the RAMP—repeat-associated mysterious proteins—superfamily of proteins containing the RNA recognition motif (RRM) domain) was described as a gene complex that is often present in the genomes along with CRISPR-Cas systems of one of the aforementioned subtypes but is not linked with a distinct *cas1*-*cas2* gene pair [6]. Consequently, additional protein families specific for each subtype or the RAMP module received names indicating subtype of their origin. For example, CRISPR subtype Apern named after the system present in *Aeropyrum pernix* genome encompasses several specific genes: csa1 (*C*RISPR *s*ystem *A*pern gene *1*), csa2, csa3, and csa4 [6]. Additionally, several core genes from the subtypes that are sufficiently distinct received a suffix letter, e.g., Cas5a (Cas5 superfamily genes of *A*pern subtype) [6]. Numerous gene families, for which no clear link to a particular subtype has been established, received gene symbols with the prefix “csx” [6].

The accumulating limitations and inconsistencies in the classification and nomenclature of CRISPR-Cas systems and *cas* genes, along with the pressing need to accommodate the rapidly growing data on sequence analysis, and structural and biochemical characterization of Cas proteins, prompted a team of CRISPR researchers to propose a revision of the aforementioned classification and nomenclature that is the current standard in the field [24] and is considered in detail below. However, it should be noted right away that the remarkable progress in the understanding of the molecular mechanisms of CRISPR-Cas and the structure of effector complexes as well as the growing number of genomes with numerous highly derived variants of CRISPR-Cas systems reveal further problems and challenges that call for the next update and improvement of this classification in the near future.

2.4 Three Major Types of CRISPR-Cas Systems, Their Subtypes, and cas Gene Nomenclature

The top level of the current CRISPR-Cas classification hierarchy includes the three major types (I, II, and III) [5] and the less common but clearly distinct type IV [30, 31]. The distinction between the CRISPR-Cas types is based on the respective signature genes and the typical organization of the respective loci. The current CRISPR-Cas classification is summarized in Table 1 and the description of structural and functional features of core Cas protein is given in Table 2.

2.4.1 Type I CRISPR-Cas Systems—All type I loci contain the signature gene *cas3* which encodes a large protein with a helicase possessing a single-stranded DNA (ssDNA)-stimulated ATPase activity coupled to unwinding of DNA-DNA and RNA-DNA duplexes [40]. Often, but not always, the helicase domain is fused to an HD family domain which has an endonuclease activity and is involved in the cleavage of the targeted DNA [40, 61]. Exonuclease (3′–5′) activity on single-stranded DNAs and RNAs has also been reported for the HD domain from *Methanocaldococcus jannaschii* [62]. The HD domain is located at the N-terminus of Cas3 proteins or is encoded by a separate gene within the same locus as *cas3* helicase. In the latter case, the helicase is denoted *cas3′* and the HD nuclease is denoted *cas3″* (Fig. 1 and [5]). In type I-F systems, *cas3* is additionally fused to the *cas2* gene.

Usually type I systems are encoded by a single operon containing the *cas1* and *cas2* genes, genes for the subunits of the Cascade or effector complex, including large subunit, small subunit (often fused to the large subunit), *cas5* and *cas7* genes, and *cas6* gene that is directly responsible for pre-crRNA transcript processing. Each gene in the type I system operons is usually present in a single copy. Several exceptions for effector complex organization are described in Table 1 and below in the text. Type I systems are currently divided into six subtypes, I-A to I-F, each of which has its own signature gene and distinct features of operon organization (Table 1). Unlike other subtypes, I-E and I-F lack the *cas4* gene. These subtypes are related according to the Cas1 phylogeny (Figs. 1 and 2). Subtypes I-A, I-B, I-C, I-E, and I-F mostly correspond to the originally proposed ones [6], with the exception of Hmari and Tneap subtypes that were combined into subtype I-B [5]. Recently, other diverged variants of several subtypes have been identified; these, however, share several features with the existing subtypes and thus still could be described within existing classification, e.g., several type I-C variants and a derived type I-F variant [24] (Fig. 1 and Table 1). In addition, the number and diversity of stand-alone (not associated with *cas1-cas2* gene pair) effector complexes are growing. These “solo” effector complexes are often present on plasmids and/or associated with transposon-related genes, such as TniQ/TnsD, a DNA-binding protein required for transposition [63, 64]. Many such cases are derivatives of subtype I-F (Fig. 1) and some others (e.g., Ava_3490-Ava_3493 *Anabaena variabilis* ATCC 29413, with genes encoding Cas6, Cas8, Cas5, Cas7) are derivatives of subtype I-C. If a system includes a derived variant of a known Cas protein family, this family might have an optional suffix indicating the subtype to which this protein belongs (e.g., Cas6f is a highly derived member of Cas6 superfamily specific for subtype I-F). Notably, the phylogenetic tree of the type I signature protein Cas3 seems to accurately reflect the subtype classification [65].

2.4.2 Type II CRISPR-Cas Systems—The signature gene for type II CRISPR-Cas systems is *cas9*, which encodes a multidomain protein that combines all the functions of effector complexes and the target DNA cleavage and is essential for the maturation of the crRNA [15]. The type II systems are also known as the "HNH" systems, *Streptococcus*-like or Nmeni subtype. Every CRISPR-Cas locus of this subtype, in addition to the *cas9* gene, also contains the ubiquitous *cas1* and *cas2* genes. In addition to these three protein-coding genes, the vast majority of type II loci also encompass one or two genes for tracrRNA, an RNA that is partially homologous to the cognate CRISPR [29, 66]. These systems use cellular (not encoded within the CRISPR- Cas loci) RNase III and tracrRNA for the processing of pre-crRNA [13]. The large Cas9 protein (~800–1,400 amino acids) contains two nuclease domains, namely the RuvC-like nuclease (RNase H fold) and the HNH (McrA-like) nuclease domain that is located in the middle of the protein [24]. Both nucleases are required for target DNA cleavage [15, 58].

Recently, several crystal structures of Cas9 have been solved including one with an artificial single-guide RNA (sgRNA) and a target DNA [59, 60]. It has been shown that Cas9 forms a two- lobed structure, with the target DNA and sgRNA positioned in the interface between the two lobes. Two loops in both lobes contribute to the recognition of the PAM. A conserved arginine cluster at the N-terminus of Cas9 belongs to a bridge helix which is critical for sgRNA: DNA recognition [59]. Outside the RuvC and HNH domains, the Cas9 structure shows no apparent structural similarity to other proteins. The part of the Cas9 protein including both nuclease domains and the arginine-rich cluster probably originated from mobile genes that are not associated with CRISPR repeats [29]. These mobile genes themselves appear to descend from a transposon gene known as ORF-B whose role in the transposon life cycle remains unknown [29]. Due to the significant sequence similarity between Cas9 and its homologs that are unrelated to CRISPR-Cas, Cas9 cannot be used as the only marker for identification of type II systems.

Type II CRISPR-Cas systems are currently classified into three subtypes (II-A, II-B, and II-C), two of which were introduced in the updated classification [5] and one was proposed recently on the basis of the distinct operon organization [29, 66, 67] (Table 1). Type II-A systems encompass an additional gene, known as *csn2* (Fig. 1), which is considered a signature gene for this subtype. The Csn2 protein is not required for interference but apparently has an unclear role in spacer integration [56]. The Csn2 proteins form homotetrameric rings that bind linear double-stranded DNA through the central hole [68–71]. This protein has been shown to adopt a highly derived P-loop ATPase fold in which the ATP- binding site appears to be inactivated [68, 69, 71]. Several highly diverged Csn2 subfamilies have been identified [29], in particular short [68, 69] and long forms [71] for which structures and biochemical characterization are available [68–71]. Type II-B systems do not encode the *csn2* gene but possess a distinct fourth gene that belongs to the Cas4 family which is also associated with subtypes I-A to I-D (but not I-E and I-F) [5]. The Cas4 proteins possess 5′-single-stranded DNA exonuclease activity [42] and belong to the PD-

EDxK family of nucleases [7]. The actual role of the Cas4 proteins in the CRISPR-Cas systems remains unknown. The recently proposed type II-C CRISPR-Cas systems possess only three protein-coding genes (*cas1*, *cas2*, and *cas9*) and are common in sequenced bacterial genomes [29, 30, 66]. Recently, type II systems have been developed into a powerful genome editing and engineering tool with a major biotechnological potential [72, 73].

2.4.3 Type III CRISPR- Cas Systems—All type III systems possess the signature gene *cas10* which encodes a multidomain protein containing a palm domain similar to that in cyclases and polymerases of the PolB family [74, 75]. Thus, this protein originally was predicted to be a polymerase [76]. Recently, the structure of Cas10 has been solved and four distinct domains have been identified [53, 77]: the N-terminal cyclase-like domain that adopts the same RRM fold as the palm domain but is not predicted to possess enzymatic activity, a helical domain containing the Zn-binding treble clef motif, the palm domain that retains the catalytic residues and is predicted to be active, and the C-terminal alpha helical domain resembling the thumb domain of A-family DNA polymerase and Cmr5, a small alpha helical protein present in some of the type III CRISPR-Cas systems. Cas10 is the large subunit of effector complexes of type III systems. Each type III locus also encodes other subunits of effector complexes such as one gene for the small subunit, one gene for a Cas5 group RAMP protein, and usually several genes for RAMP proteins of the Cas7 group (Fig. 1 and *see* Subheading 2.6.2). Often Cas10 is fused to an HD family nuclease domain that is distinct from the HD domains of type I CRISPR-Cas systems and, unlike the latter, contains a circular permutation of the conserved motifs [7, 76]. Type III CRISPR-Cas systems often do not encode their own *cas1* and *cas2* genes but use crRNAs produced from CRISPR arrays associated with type I or type II systems [14, 78]. Nevertheless, in many genomes that lack type I and type II systems, *cas1*, *cas2*, and *cas6* genes are linked to a type III system that accordingly is predicted to be fully functional [31]. Currently, there are two subtypes within type III, III-A (former Mtube subtype or Csm module), and III-B (former RAMP module or Cmr module), which are clearly related but could be distinguished by the presence of distinct genes for small subunits of effector complexes, *csm2* and *cmr5*, respectively (Fig. 1, Table 1, and Subheading 2.6.3). The subtype III-A loci often possess *cas1*, *cas2*, and *cas6* [31] and have been shown to target DNA [79], whereas most of the III-B systems lack these genes and therefore depend on other CRISPR-Cas systems present in the same genome. The type III-B CRISPR-Cas systems have been shown to target RNA [14, 23, 47].

The composition and organization of type III CRISPR-Cas systems are much more diverse compared with type I systems. The diversity is achieved by gene duplications and deletions, domain insertions and fusions, and the presence of additional, poorly characterized domains that presumably are involved in either effector complexes or associated immunity. At least two of the type III variants (one of type III-A and the other of type III-B) are relatively common (Fig. 1 and Table 1). The distinguishing feature of the type III-B variant is the apparent inactivation of the palm/cyclase domain of Cas10 whereas the type III-A variants typically encompass a Cas10 gene lacking the HD domain and additionally contain an uncharacterized gene homologous to all1473 from *Nostoc* sp. PCC 7120 [24]. Both type III variants are typically present in a genome along with other CRISPR-Cas systems.

2.4.4 Type IV CRISPR- Cas Systems—Type IV CRISPR-Cas systems, found in several bacterial genomes, often on plasmids, can be typified by the CRISPR-Cas locus in *Acidithiobacillus ferrooxidans* ATCC 23270 (operon AFE_1037- AFE_1040). Similar to subtype III-A, this system lacks *cas1* and *cas2* genes and often is not associated with CRISPR arrays. Moreover, in many bacteria, this is the only CRISPR-Cas system, with no CRISPR cassette detectable in the genome. The type IV systems possess an effector complex that consists of a highly reduced large subunit (*csf1*), two genes for RAMP proteins of the Cas5 (*csf3*) and Cas7 (*csf2*) groups, and, in some cases, a gene for a predicted small subunit [24]. The *csf1* gene could be considered a signature gene for this system (Fig. 1 and Table 1). There are two distinct subtypes of type IV systems, one of which contains a DinG family helicase *csf4* [80], whereas the second subtype lacks DinG but typically contains a gene for a small alpha helical protein, presumably a small subunit [24]. Type IV CRISPR-Cas systems could be mobile modules that, similar to type III systems, could utilize crRNA from different CRISPR arrays once these become available. However, other mechanisms such as generation of crRNA directly from alien RNA, without incorporation of spacers in CRISPR cassettes, cannot be ruled out.

The classification of CRISPR-Cas systems outlined above more or less adequately covers the representation of these systems in sequenced bacterial and archaeal genomes. However, considering the rapid evolution of CRISPR-Cas, these variants might represent only the proverbial a tip of the iceberg with respect to the true diversity of prokaryotic adaptive immunity. As a case in point, two novel CRISPR-Cas systems have been recently identified in the genomes of *Thermococcus onnurineus* and *Ignisphaera aggregans* [81]. Based on some marginal similarities, these loci could be tentatively assigned to type I and type III, respectively; however, they do not contain any signature genes described above that would allow one to classify them into any known subtype. Similarly, classification of certain type I systems, such as the one from *Microcystis aeruginosa* (MAE_30760-MAE_30790) and several other species [82], is hampered by the apparent absence of signature genes of the known type I subtypes. Accumulation of such “unclassifiable” variants raises the possibility that the current principles of CRISPR- Cas system classification might have to be reconsidered to take into account the challenge of the ever-increasing diversity.

2.5 Phylogeny and Genomic Associations of Cas1

The endonuclease Cas1 is an essential Cas protein that ensures the unique ability of CRISPR systems to keep memory of previous encounters with infectious agents. Cas1 and Cas2 form a hetero- hexameric complex that is necessary and sufficient for spacer integration [10–12, 83]. However, only the enzymatic activity of Cas1 is required for spacer integration by the Cas1-Cas2 complex whereas the activity of Cas2 is dispensable indicating that this protein has a structural role in spacer acquisition [11].

To date, three Cas1 proteins, from *Escherichia coli*, *Pseudomonas aeruginosa*, and *Archaeoglobus fulgidus*, have been experimentally characterized and their structures have been solved [35, 37, 84]. It has been shown that Cas1 protein forms a homodimer and is a metal-dependent nuclease that cleaves ssDNA and dsDNA. The Cas1 monomer consists of two domains, with the C-terminal α-helical catalytic domain and the mostly beta-stranded

N-terminal domain that is probably involved in dimerization and interaction with other proteins, in particular Cas2 [35, 37, 84].

Cas1 is the most conserved Cas protein and its phylogeny generally correlates with the organization of CRISPR-Cas system loci; accordingly, until recently, Cas1 has been considered the signature for the presence of CRISPR-Cas systems in a genome [5–7]. However, as pointed out above, recently it has been found that many genomes that lack a *cas1* gene possess Cas loci that encode apparently active effector complexes and thus might function in a Cas1-independent fashion. The examples of systems lacking *cas1* include the type IV systems, described above, subtype III-B, and a variant of subtype I-F (Fig. 1).

Conversely, it has been recently shown that *cas1* is a component of predicted self-synthesizing transposable elements, dubbed casposons, where it is always associated with a DNA polymerase of the B family and variable sets of diverse genes [85]. Furthermore, in some other archaeal genomes from the Methanomicrobiales lineage, *cas1* is linked neither to casposons nor to CRISPR-Cas system and its function in these organisms remains obscure [31, 85]. Figure 2a presents a scheme of the Cas1 phylogeny published before [31]. The two groups of Cas1 that are not associated with CRISPR-Cas systems form two separate branches deep in the Cas1 tree. Their relationships with branches that correspond to Cas1 groups associated with known CRISPR-Cas systems are not resolved. Consistent with previous analyses, most of the known type I and type II subtypes form distinct branches. However, only subtypes I-E, I-F, II-A, and II-B are strictly monophyletic whereas the other subtypes show multiple deviations from the classification scheme [31]. In contrast, Cas1 proteins associated with both type III subtypes do not form monophyletic groups suggesting that these systems are compatible with a wide range of Cas1 proteins acquired from other CRISPR-Cas types. The number of genomes that possess only subtype III-A CRISPR-Cas is growing fast whereas subtype III-B systems associated with Cas1 and Cas2 remain rare [31]. Accordingly, much of the diversity of Cas1 is concentrated within subtype III-A (Fig. 2a) [31]. Another important observation is the polyphyly of the type II systems whereby Cas1 sequences of type II-B form a clade within the type I-A branch. The origin of the other type II-B genes from within type I is also supported by phylogenetic analysis of Cas2 and Cas4 proteins [29]. Generally, these findings confirm that effector complexes of CRISPR-Cas can function in association with “information processing” modules of different origin.

The c*as1* gene is often found either fused or located in the same predicted operons with a number of enzymatically active domains or predicted transcriptional regulators (Fig. 2b). Many enzymatic domains linked to Cas1 do not belong to any Cas families and are known components of various defense systems that possess either RNase or DNase activity (*see* Subheading 2.5). Thus, it appears that the expression and activity of Cas1 proteins are tightly controlled and coupled to programmed cell death/dormancy mechanisms [1, 30, 86].

One of the most puzzling connections of Cas1 is to a gene called *cpf1* (see description at http://www.jcvi.org/cgi-bin/tigrfams/HmmReportPage.cgi?acc=TIGR04330), which encodes a large protein (about 1,300 amino acids), an uncharacterized protein. This gene is found in several diverse bacterial genomes, typically in the same locus with *cas1*, *cas2*, and *cas4* genes and a CRISPR cassette (for example, FNFX1_1431-FNFX1_1428 of *Francisella* cf.

novicida Fx1). Thus, the layout of this putative novel CRISPR- Cas system appears to be similar to that of type II-B. Furthermore, similar to Cas9, the Cpf1 protein contains a readily identifiable C-terminal region that is homologous to the transposon ORF-B and includes an active RuvC-like nuclease, an arginine-rich region, and a Zn finger (absent in Cas9). However, unlike Cas9, Cpf1 is also present in several genomes without a CRISPR-Cas context and its relatively high similarity with ORF-B suggests that it might be a transposon component. If however this is a genuine CRISPR-Cas system and Cpf1 is a functional analog of Cas9 it would be a novel CRISPR-Cas type, namely type V. Hopefully this interesting system will be experimentally studied in the near future.

2.6 Principles of Organization of CRISPR-Cas Surveillance and Effector Complexes

2.6.1 General Features of Effector Complex Organization—The effector (surveillance) complex of CRISPR-Cas systems is involved in pre-crRNA processing (except in type III) and crRNA- guided targeting of foreign DNA or RNA. This complex contains from one (type II CRISPR-Cas systems) to several proteins and binds crRNA and DNA. Specific recognition of the DNA sequence matching the spacer (termed the protospacer) within the respective crRNA is necessary for the Cascade to form an R-loop and recruit a nuclease to cleave the target DNA. In addition, in type I and II systems recognition of a flanking PAM (protospacer adjacent sequence) is required. To date, the structure and organization of several effector complexes from different CRISPR-Cas systems of both type I and type III have been studied in detail. These include the Cascade complex from *E. coli* (subtype I-E) [20, 43, 44], Csy complex (subtype I-F) from *P. aeruginosa* [19], a(rchaeal) Cascade from *S. solfataricus* (subtype I-A)[87], subtype I-C complex *from Bacillus halodurans* [39], Csm-complex from *S. solfataricus* (subtype III-A) [45], and Cmr complexes (subtype III-B) from *Thermus thermophilus* [22] and *Pyrococcus furiosus* [23]. The analysis of these complexes revealed striking similarities in their organization despite the absence of sequence similarity between the majority of the constituents. These findings are consistent with previous predictions that have been made using comparative genomic methods [24, 30].

A general scheme of effector complex organization related to type I and type III systems is shown in Fig. 3a and reflects the following observations. Effector complexes consist of one large subunit, several small subunits, one Cas5 family protein, and several Cas7 family proteins. Cas5 and large and small subunits are usually encoded by a single gene each in both type I and type III systems, although large and small subunits are likely fused in several type I subtypes (Fig. 1) [24]. The Cas7 group proteins are encoded by a single gene in the respective type I system loci and by several separate genes in type III systems (Fig. 1). Functionally, Cas7 is involved in crRNA binding, and Cas5 in binding the 5′-handle of crRNA and interaction with the large subunit and the proximal Cas7 protein. The large subunit participates in DNA binding and recognition of the PAM sequence [17, 18, 22, 23, 43, 45, 88]. The Cas6 proteins, which are directly involved in pre-crRNA processing, usually do not belong to the effector complex but could be loosely associated with some of them [20]. A strong interaction has been detected between the Cas7 and Cas5 proteins and loose association has been identified between the large and small subunits when encoded by separate genes [17, 19, 45, 88]. It has been proposed that, in addition to crRNA-guided DNA targeting, type I-E Cascade can migrate along the DNA molecule, facilitating the selection of fragments to be incorporated into the CRISPR locus [83].

Four distinct subunits of the effector complexes of type I and type III CRISPR-Cas contain RRM domains, which consists of a four-stranded antiparallel β-sheet (arranged as β4β1β3β2), with two β-helices located after β1 and β3 in a βαββαβ ferredoxin-like fold [7, 54]. The type III large subunit, Cas10, contains two RRM domains, one of which is a polymerase/cyclase palm domain predicted to be active, whereas the other one is an inactivated version of the palm domain [53, 77]. The Cas5 proteins typically contain two RRM domains, with the C-terminal domain degraded in several subfamilies; the Cas7 proteins possess a single RRM domain; and the Cas6 proteins encompass two RRM domains [24, 30]. It has been hypothesized that the RAMP proteins evolved from the large subunit by duplication and specialization [24, 31]. The type II effector complex consists of a single multidomain protein, Cas9, that binds crRNA and tracrRNA. Similarly to the type I and type III complexes, the type II effector complex (Cas9) scans DNA, recognizes PAM, and forms an R-loop (*see* Subheading 2.4.2).

2.6.2 Three Major Families of RAMPs—Exhaustive sequence analysis, supported by the analysis of the growing collection of structural data, indicates that RAMPs can be classified into three families, the largest of which includes the Cas7 proteins (Fig. 3b). In the majority of CRISPR-Cas systems, processing of pre-crRNA is catalyzed by dedicated endoribonucleases that belong to the Cas6 family of RAMPs. Among all RAMP families, the Cas6 family has been characterized in most detail, both structurally and biochemically. This protein shows remarkable plasticity of the catalytic mechanism and RNA recognition modes [48–50]. The type member of the Cas6 family is the protein from the archaeon *Pyrococcus furiosus* [47, 51, 54]. The *P. furiosus* Cas6 contains two RRM domains with a G-rich loop located at the C-terminus of the second RRM domain and the catalytic triad consisting of histidine, tyrosine, and lysine located within the first, N-terminal RRM domain [51]. The conserved catalytic histidine is located within the alpha helix that follows the first core beta strand of the N-terminal RRM domain. Many Cas6 subfamilies contain the catalytic histidine in the same position but other arrangements of catalytic residues have been detected as well [48, 49, 54]. The cleavage of the pre-crRNA occurs within a CRISPR repeat at the 5′ side of the phosphodiester bond, generating a 5′ end hydroxyl group and either a 3′ phosphate (Cas6 from *Pseudomonas aerugi-nosa*) or a 2′, 3′ end cyclic phosphate group (Cas6e), and yields a crRNA of approximately 60 nt in size [19, 52, 89]. The majority of the Cas6 proteins show substantial sequence conservation and belong to the core of the Cas6 family (COG1853/COG5551) but several are highly divergent, e.g., those associated with I-E (Cas6e) and I-F (Cas6f) CRISPR-Cas subtypes (Fig. 3b). The latter is the most derived Cas6 protein with a severely degraded C-terminal RRM domain [24, 52, 54].

The Cas7 proteins form the backbone of effector complexes that play the key role in binding and protecting the crRNA guide sequence. The Cas7 family proteins typically contain one RRM domain that is structurally similar to the N-terminal RRM domain of Cas6 and two distinct, albeit highly variable, subdomains [54, 87, 90]. The majority of the Cas7 family proteins associated with type III systems contain the characteristic G-rich loop, the structural marker of the RAMP superfamily (Fig. 3b). These proteins are diverse and could be present in several copies in the type III loci (Fig. 1). The Cas7 family proteins associated with type III systems are apparently prone to aggregation, forming multidomain proteins (e.g., Cmr1 family or Psta_1142 from *Pirellula staleyi* or HMPREF9137_2396 *Prevotella denticola*). Several subfamilies of the Cas7 family (Cmr4, Csm5, and Csm3) possess a conserved histidine that is structurally equivalent to the catalytic histidine of Cas6 required for pre-crRNA cleavage, suggesting that these Cas7 proteins are active RNases (Fig. 3b) [24]. This hypothesis is compatible with the demonstration of the RNA cleavage activity of the Cmr complex of *T. thermophilus* [22].

The Cas5 family proteins bind the 5′-handle of crRNA and provide the interaction interface for the large subunit and the Cas7 proteins. Similarly to the Cas6 family, most of the Cas5 proteins contain two RRM domains [24, 46, 54, 91] although the C-terminal domain is severely deteriorated in many Cas5 proteins associated with type I systems (Fig. 3b) [24]. The G-rich loop is easily detectable in the first RRM domain and often is present also in the second RRM domain in those Cas5 group RAMPs that are associated with type III systems (Fig. 3b) [24, 54]. Usually, only one Cas5 protein is encoded in a CRISPR-Cas locus (Fig. 1). These proteins, especially those associated with type III systems, are prone to structural rearrangements and fusions (e.g., Rcas_3293 from *Roseiflexus castenholzii* represents fusion of Cas5 and Cas7 group RAMPs). Some proteins are assigned to the Cas5 family provisionally, based on the general principles of organization of effector complexes because they do not share any similarity with known Cas proteins (Fig. 3b) [24]. Among the Cas5 family members, there are also proteins with a conserved N-terminal histidine (e.g., Csm4 subfamily). However, RNase activity has been experimentally demonstrated only for the Cas5 proteins that are associated with type I-C systems; these proteins have a different set of catalytic residues compared with Cas6 and are directly involved in pre-crRNA processing [39, 46].

2.6.3 Small Subunits—Typically, the small subunit is an alpha-helical protein containing up to eight predicted alpha helices. The small subunits are encoded by a separate gene in all type III CRISPR-Cas systems, in some type I systems, such as I-A and I-E, and one variant of type IV (Fig. 1). Analogous to the large subunit, the small subunits are highly diverse, such that different families often show no detectable sequence similarity to each other. In the majority of type I systems, large subunits contain a 4–6 alpha helical C-terminal extension that appears to complement the absence of a small subunit gene. Accordingly, it has been hypothesized that these proteins represent a fusion of the large and small subunits (Figs. 1 and 3b, d) [24]. In the effector complexes that include the small subunit as a separate component, there are usually several small subunit genes [17, 18, 20, 22, 23, 43, 45, 87, 92]. Recently, the structure of the small subunit Csa5 of type I-A from *S. solfataricus* has been reported [54, 55]. Comparison of these structures sheds new light on the evolution of the small subunits by demonstrating the evolutionary connection between small subunits of type I and type III systems [54, 55]. Structural comparison revealed that Cmr5, the small subunit of subtype III-B, corresponds to the N-terminal domain of Cse2, the small subunit of type I-E, whereas Csa5 partially corresponds to the C-terminal domain of Cse2 [54, 55]. In addition, Csa5 has a unique beta-stranded extension which is absent even in the proteins that belong to the same protein family (Fig. 3b). The relationships of these proteins to other distinct families of predicted small subunits, such as Csm2 (type III-A) and RHA1_ro10070 (type IV), remain to be elucidated (Fig. 3b).

2.6.4 Large Subunits—Multiple lines of evidence coming from *in silico* analysis of Cas proteins suggest that, the absence of significant sequence similarity notwithstanding, the large subunits present in most of the type I CRISPR-Cas systems could be homologous to Cas10 proteins, which contain two palm/cyclase domains, one of which is predicted to be enzymatically active (Fig. 3d) [24, 31]. Recent structures of effector and surveillance complexes from both major subtypes are compatible with this inference [19, 22, 23, 39, 45, 87, 88]. The crystal structures of two distinct large subunits have been solved, namely those of Cas10, the large subunit of type III systems [53, 77], and CasA (Cse1 or Cas8e) of subtype I-E [16]. The core domains of Cas10 are the N-terminal cyclase domain, Zn finger containing the treble clef domain, cyclase (palm) domain with the characteristic catalytic motif "GGDD," and the C-terminal alpha helical bundle (Fig. 3d). This arrangement is reminiscent to "Fingers," "Palm," and "Thumb" domains present in DNA polymerases of different families [93]. In Cas10, these four regions are arranged into four distinct domains, whereas Cse1 displays a much compact architecture where traces of the putative ancestral domain architecture are barely identifiable [31]. Major structural rearrangements of this type could have been anticipated even before these structures became available because the large subunits of different type I systems substantially differ in size and even seem to be missing in some systems (e.g., I-C variant, Table 1 and Fig. 3d) [24, 31].

Variations of the domain architecture and inactivation of the catalytic palm domain could be identified even within the Cas10 family. Some representatives lack the HD domain, a nuclease that is typically fused to Cas10 at the N-terminus (e.g., Caur_2291 from *Chloroflexus aurantiacus*), and the palm domain appears to be inactivated in many subtype III-B variants, e.g., MTH326 from *Methanothermobacter thermautotrophicus*) (see details in [24]).

An exhaustive comparison of multiple alignments and predicted secondary structures of the large subunits of type I and type III systems revealed several shared features, such as a Zn finger in the middle of the protein sequence in the Cas8a, Cas8b, Cse1, and Csf1 families, conserved beta-hairpin in the region roughly corresponding to the palm domain, and an alpha helical region at the C-terminus of the proteins compatible with the alpha helical bundle of Cas10 (Fig. 3d) [24]. However, direct comparisons detect no sequence similarity between Cas10 and any large subunits of type I systems, and moreover, many large type I subunits share no significant sequence similarity with each other, suggestive of extremely fast divergence of these proteins. An unusual variant of the large subunit denoted that Cas10d is associated with type I-D. This protein appears to be structurally similar to Cas10 and even contains an N-terminal HD domain but the latter has a circular permutation of the catalytic motifs which seems suggestive of its origin from the HD domain of Cas3 protein (Fig. 2) [24].

2.7 Potential Associated Immunity Genes and Regulatory Components

Many DNA-targeting defense systems contain previously overlooked components that are implicated in programmed cell death (PCD)/dormancy [1, 30, 31, 86, 94]. Experimental data on coupling between immunity and PCD is scarce, having been demonstrated only for the *Escherichia coli* anticodon nuclease (ACNase) PrrC which contributes to the T4 phage exclusion mechanism as a component of the RM type Ic system PrrI [95]. A general hypothesis for the apparent integration of the two defense strategies has been proposed [1, 30, 31, 86, 94]. Specifically, a toxin associated with an immunity system, such as CRISPR-Cas, could act either as a dormancy inducer, which prevents fast virus propagation and could "buy the time" for the activation of the primary immune system, or, alternatively, as a toxin that causes altruistic suicide when immunity fails [86]. CRISPR-Cas systems are especially rich in genes encoding proteins associated with PCD [30, 31]. In particular, two core Cas proteins, Cas2 and Cas4, belong to families of nucleases that commonly function as toxins in toxin-antitoxin systems which are responsible for PCD in prokaryotes [1, 30, 86]. It has been shown that Cas2 forms a specific complex with Cas1 that is required for spacer acquisition by CRISPR-Cas but mutation of a residue predicted to be required for nuclease activity of Cas2 did not affect this step [11]. Thus, it appears likely that Cas1 and Cas2 modulate each other's activities, where the putative toxic nuclease activity of Cas2 is unleashed only under genotoxic stress caused by infection that is not controlled by immunity. Both Cas2 and Cas4 are fused with several other Cas proteins but not with effector complex components (Fig. 4). In several bacteria, an apparently inactivated Cas2 is fused to a 3′–5′ exonuclease of the DEDDh family [7], suggesting that the lost nuclease activity of Cas2 could be replaced by an unrelated enzyme. Cas4 proteins of two distinct families are often present in the same operon (e.g., PAE0079- PAE0082 in *Pyrobaculum aerophilum* str. IM2) suggestive of functional differentiation (Fig. 4). A variety of other Cas proteins and proteins that are sporadically present within CRISPR-Cas loci are fused to or encoded next to Cas1, indicating that there are multiple ways to control the activity of this key Cas protein (Fig. 2b) [31]. One of such Cas1 fusions involves a reverse transcriptase for which cell toxicity has been recently demonstrated [96] (Fig. 2b).

Many more genes seem to be specifically associated with type III CRISPR-Cas systems. However, all these families have been identified in other genomic contexts as well. The largest superfamily of such proteins includes members of COG1517 that typically consist of a CARF domain (CRISPR-Cas-associated Rossmann- fold domain), an HTH domain [97, 98], and various effector domains, most of which are predicted to be active RNases and DNases [5, 7, 32] (Fig. 4). The HEPN domain containing a characteristic RxxxxH motif, the most abundant effector domain present in such families as Csm6 and Csx1, is a predicted ribonuclease [94]. Notably, the great majority of the COG1517 members associated with type III CRISPR-Cas systems contain effector domains [32]. The COG1517-related genes are often found in the same operon with other genes encoding uncharacterized proteins that contain conserved potential catalytic residues and could represent novel families of nucleases (Fig. 4) [32]. Another abundant uncharacterized family typically contains the WYL domain (named after the respective conserved amino acid residues) and an HTH domain and thus reminds the core domain organization of COG1517. Both WYL and CARF domains are predicted to bind yet unidentified ligands, most likely nucleotides, and thus regulate the expression and/or activity of CRISPR-Cas systems via an allosteric mechanism [32]. Indeed, one of the WYL domain-containing proteins has been shown to regulate the expression of the CRISPR- Cas locus in *Synechocystis* sp. PCC6803 [99]. A plausible possibility seems to be that these regulators mediate the functional coupling of the CRISPR-Cas immunity with dormancy induction and PCD.

3 Conclusions and Outlook

The advances of comparative genomic analysis reveal unprecedented complexity of the CRISPR-Cas systems. The classification of CRISPR-Cas systems into three types and ten subtypes introduced some order into this striking diversity and provides the essential template for genome annotation and evolutionary studies. However, it is already perfectly clear that new types and subtypes of CRISPR-Cas have to be introduced. Moreover, this classification system, however refined and improved, can capture only part of the complexity of CRISPR-Cas organization and evolution, due to the intrinsic modularity and evolutionary mobility of these immunity systems, resulting in numerous recombinant variants. In particular, although Cas1 is the most conserved Cas protein, in terms of both presence in the great majority of CRISPR-Cas loci and sequence conservation, Cas1 phylogeny is of limited utility of CRISPR-Cas classification because of the extensive shuffling of the "informational" and "executive" modules. One possible way to achieve greater flexibility in CRISPR-Cas classification is to analyze these modules separately and explicitly recognize recombinants. However, gene and domain shuffling is extensive also within the modules so that we expect CRISPR classification to remain challenging for the foreseeable future. Above and beyond this organizational complexity of CRISPR-Cas systems, most of the *cas* genes evolve rapidly, which complicates the family assignment for many Cas proteins and the use of family profiles for the recognition of CRISPR-cas subtype signatures. Clearly, to achieve progress in the comparative analysis of CRISPR-Cas systems integration of the most sensitive sequence comparison tools with protein structure comparison is essential.

References

1. Makarova KS, Wolf YI, Koonin EV. Comparative genomics of defense systems in archaea and bacteria. Nucleic Acids Res. 2013; 41:4360–4377. [PubMed: 23470997]
2. Barrangou R, Horvath P. CRISPR: new horizons in phage resistance and strain identification. Annu Rev Food Sci Technol. 2012; 3:143–162. [PubMed: 22224556]
3. Wiedenheft B, Sternberg SH, Doudna JA. RNA-guided genetic silencing systems in bacteria and archaea. Nature. 2012; 482:331–338. [PubMed: 22337052]
4. van der Oost J, Jore MM, Westra ER, Lundgren M, Brouns SJ. CRISPR-based adaptive and heritable immunity in pro-karyotes. Trends Biochem Sci. 2009; 34:401–407. [PubMed: 19646880]
5. Makarova KS, Haft DH, Barrangou R, Brouns SJ, Charpentier E, Horvath P, Moineau S, Mojica FJ, Wolf YI, Yakunin AF, van der Oost J, Koonin EV. Evolution and classification of the CRISPR-Cas systems. Nat Rev Microbiol. 2011; 9:467–477. [PubMed: 21552286]
6. Haft DH, Selengut J, Mongodin EF, Nelson KE. A guild of 45 CRISPR-associated (Cas) protein families and multiple CRISPR/Cas subtypes exist in prokaryotic genomes. PLoS Comput Biol. 2005; 1:e60. [PubMed: 16292354]
7. 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]

8. Barrangou R. CRISPR-Cas systems and RNA-guided interference. Wiley Interdiscip Rev RNA. 2013; 4:267–278. [PubMed: 23520078]
9. Westra ER, Swarts DC, Staals RH, Jore MM, Brouns SJ, van der Oost J. The CRISPRs, they are a-changin': how prokaryotes generate adaptive immunity. Annu Rev Genet. 2012; 46:311–339. [PubMed: 23145983]
10. 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]
11. Nunez JK, Kranzusch PJ, Noeske J, Wright AV, Davies CW, Doudna JA. Cas1-Cas2 complex formation mediates spacer acquisition during CRISPR-Cas adaptive immunity. Nat Struct Mol Biol. 2014; 21:528–534. [PubMed: 24793649]
12. Richter C, Gristwood T, Clulow JS, Fineran PC. In vivo protein interactions and complex formation in the *Pectobacterium atro-septicum* subtype I-F CRISPR/Cas System. PLoS One. 2012; 7:e49549. [PubMed: 23226499]
13. Deltcheva E, Chylinski K, Sharma CM, Gonzales K, Chao Y, Pirzada ZA, Eckert MR, Vogel J, Charpentier E. CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. Nature. 2011; 471:602–607. [PubMed: 21455174]
14. Hale CR, Majumdar S, Elmore J, Pfister N, Compton M, Olson S, Resch AM, Glover CV III, Graveley BR, Terns RM, Terns MP. Essential features and rational design of CRISPR RNAs that function with the Cas RAMP module complex to cleave RNAs. Mol Cell. 2012; 45:292–302. [PubMed: 22227116]
15. Jinek M, Chylinski K, Fonfara I, Hauer M, Doudna JA, Charpentier E. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. Science. 2012; 337:816–821. [PubMed: 22745249]
16. Sashital DG, Wiedenheft B, Doudna JA. Mechanism of foreign DNA selection in a bacterial adaptive immune system. Mol Cell. 2012; 46:606–615. [PubMed: 22521690]
17. van Duijn E, Barbu IM, Barendregt A, Jore MM, Wiedenheft B, Lundgren M, Westra ER, Brouns SJ, Doudna JA, van der Oost J, Heck AJ. Native tandem and ion mobility mass spectrometry highlight structural and modular similarities in clustered-regularly-interspaced shot-palindromic-repeats (CRISPR)-associated protein complexes from escherichia coli and pseudomonas aeruginosa. Mol Cell Proteomics. 2012; 11:1430–1441. [PubMed: 22918228]
18. Zhang J, Rouillon C, Kerou M, Reeks J, Brugger K, Graham S, Reimann J, Cannone G, Liu H, Albers SV, Naismith JH, Spagnolo L, White MF. Structure and mechanism of the CMR complex for CRISPR-mediated antiviral immunity. Mol Cell. 2012; 45:303–313. [PubMed: 22227115]
19. Wiedenheft B, van Duijn E, Bultema JB, Waghmare SP, Zhou K, Barendregt A, Westphal W, Heck AJ, Boekema EJ, Dickman MJ, Doudna JA. RNA-guided complex from a bacterial immune system enhances target recognition through seed sequence interactions. Proc Natl Acad Sci U S A. 2011; 108:10092–10097. [PubMed: 21536913]
20. Brouns SJ, Jore MM, Lundgren M, Westra ER, Slijkhuis RJ, Snijders AP, Dickman MJ, Makarova KS, Koonin EV, van der Oost J. Small CRISPR RNAs guide antiviral defense in prokaryotes. Science. 2008; 321:960–964. [PubMed: 18703739]
21. Zhang Y, Heidrich N, Ampattu BJ, Gunderson CW, Seifert HS, Schoen C, Vogel J, Sontheimer EJ. Processing-independent CRISPR RNAs limit natural transformation in *Neisseria meningitidis*. Mol Cell. 2013; 50:488–503. [PubMed: 23706818]
22. Staals RH, Agari Y, Maki-Yonekura S, Zhu Y, Taylor DW, van Duijn E, Barendregt A, Vlot M, Koehorst JJ, Sakamoto K, Masuda A, Dohmae N, Schaap PJ, Doudna JA, Heck AJ, Yonekura K, van der Oost J, Shinkai A. Structure and activity of the RNA-targeting Type III-B CRISPR-Cas complex of *Thermus thermophilus*. Mol Cell. 2013; 52:135–145. [PubMed: 24119403]
23. Spilman M, Cocozaki A, Hale C, Shao Y, Ramia N, Terns R, Terns M, Li H, Stagg S. Structure of an RNA silencing complex of the CRISPR-Cas immune system. Mol Cell. 2013; 52:146–152. [PubMed: 24119404]
24. 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]

25. Marchler-Bauer A, Anderson JB, Chitsaz F, Derbyshire MK, DeWeese-Scott C, Fong JH, Geer LY, Geer RC, Gonzales NR, Gwadz M, He S, Hurwitz DI, Jackson JD, Ke Z, Lanczycki CJ, Liebert CA, Liu C, Lu F, Lu S, Marchler GH, Mullokandov M, Song JS, Tasneem A, Thanki N, Yamashita RA, Zhang D, Zhang N, Bryant SH. CDD: specific functional annotation with the conserved domain database. Nucleic Acids Res. 2009; 37:D205–D210. [PubMed: 18984618]
26. Soding J, Biegert A, Lupas AN. The HHpred interactive server for protein homology detection and structure prediction. Nucleic Acids Res. 2005; 33:W244–W248. [PubMed: 15980461]
27. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997; 25:3389–3402. [PubMed: 9254694]
28. Wheeler D, Bhagwat M. BLAST QuickStart: example-driven web-based BLAST tutorial. Methods Mol Biol. 2007; 395:149–176. [PubMed: 17993672]
29. Chylinski K, Makarova KS, Charpentier E, Koonin EV. Classification and evolution of type II CRISPR-Cas systems. Nucleic Acids Res. 2014; 42:6091–6105. [PubMed: 24728998]
30. Koonin EV, Makarova KS. CRISPR- Cas: evolution of an RNA-based adaptive immunity system in prokaryotes. RNA Biol. 2013; 10:679–686. [PubMed: 23439366]
31. 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]
32. Makarova KS, Anantharaman V, Grishin NV, Koonin EV, Aravind L. CARF and WYL domains: ligand-binding regulators of prokaryotic defense systems. Front Genet. 2014; 5:102. [PubMed: 24817877]
33. Jansen R, Embden JD, Gaastra W, Schouls LM. Identification of genes that are associated with DNA repeats in prokaryotes. Mol Microbiol. 2002; 43:1565–1575. [PubMed: 11952905]
34. Takeuchi N, Wolf YI, Makarova KS, Koonin EV. Nature and intensity of selection pressure on CRISPR-associated genes. J Bacteriol. 2012; 194:1216–1225. [PubMed: 22178975]
35. Wiedenheft B, Zhou K, Jinek M, Coyle SM, Ma W, Doudna JA. Structural basis for DNase activity of a conserved protein implicated in CRISPR-mediated genome defense. Structure. 2009; 17:904–912. [PubMed: 19523907]
36. Han D, Lehmann K, Krauss G. SSO1450–a CAS1 protein from Sulfolobus solfataricus P2 with high affinity for RNA and DNA. FEBS Lett. 2009; 583:1928–1932. [PubMed: 19427858]
37. Babu M, Beloglazova N, Flick R, Graham C, Skarina T, Nocek B, Gagarinova A, Pogoutse O, Brown G, Binkowski A, Phanse S, Joachimiak A, Koonin EV, Savchenko A, Emili A, Greenblatt J, Edwards AM, Yakunin AF. A dual function of the CRISPR-Cas system in bacterial antivirus immunity and DNA repair. Mol Microbiol. 2011; 79:484–502. [PubMed: 21219465]
38. Beloglazova N, Brown G, Zimmerman MD, Proudfoot M, Makarova KS, Kudritska M, Kochinyan S, Wang S, Chruszcz M, Minor W, Koonin EV, Edwards AM, Savchenko A, Yakunin AF. 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]
39. Nam KH, Haitjema C, Liu X, Ding F, Wang H, DeLisa MP, Ke A. Cas5d protein processes pre-crRNA and assembles into a cascade-like interference complex in subtype I-C/Dvulg CRISPR-Cas system. Structure. 2012; 20:1574–1584. [PubMed: 22841292]
40. Sinkunas T, Gasiunas G, Fremaux C, Barrangou R, Horvath P, Siksnys V. Cas3 is a single-stranded DNA nuclease and ATP-dependent helicase in the CRISPR/Cas immune system. EMBO J. 2011; 30:1335–1342. [PubMed: 21343909]
41. Han D, Krauss G. Characterization of the endonuclease SSO2001 from *Sulfolobus solfataricus* P2. FEBS Lett. 2009; 583:771–776. [PubMed: 19174159]
42. Zhang J, Kasciukovic T, White MF. The CRISPR associated protein Cas4 Is a 5′ to 3′ DNA exonuclease with an iron-sulfur cluster. PLoS One. 2012; 7:e47232. [PubMed: 23056615]
43. Wiedenheft B, Lander GC, Zhou K, Jore MM, Brouns SJ, van der Oost J, Doudna JA, Nogales E. Structures of the RNA-guided surveillance complex from a bacterial immune system. Nature. 2011; 477:486–489. [PubMed: 21938068]
44. Jore MM, Lundgren M, van Duijn E, Bultema JB, Westra ER, Waghmare SP, Wiedenheft B, Pul U, Wurm R, Wagner R, Beijer MR, Barendregt A, Zhou K, Snijders AP, Dickman MJ, Doudna JA,

Boekema EJ, Heck AJ, van der Oost J, Brouns SJ. Structural basis for CRISPR RNA-guided DNA recognition by Cascade. Nat Struct Mol Biol. 2011; 18:529–536. [PubMed: 21460843]

45. Rouillon C, Zhou M, Zhang J, Politis A, Beilsten-Edmands V, Cannone G, Graham S, Robinson CV, Spagnolo L, White MF. Structure of the CRISPR interference complex CSM reveals key similarities with cascade. Mol Cell. 2013; 52:124–134. [PubMed: 24119402]

46. Koo Y, Ka D, Kim EJ, Suh N, Bae E. Conservation and variability in the structure and function of the Cas5d endoribonuclease in the CRISPR-mediated microbial immune system. J Mol Biol. 2013; 425:3799–3810. [PubMed: 23500492]

47. Hale CR, Zhao P, Olson S, Duff MO, Graveley BR, Wells L, Terns RM, Terns MP. RNA-guided RNA cleavage by a CRISPR RNA-Cas protein complex. Cell. 2009; 139:945–956. [PubMed: 19945378]

48. Niewoehner O, Jinek M, Doudna JA. Evolution of CRISPR RNA recognition and processing by Cas6 endonucleases. Nucleic Acids Res. 2014; 42:1341–1353. [PubMed: 24150936]

49. Reeks J, Sokolowski RD, Graham S, Liu H, Naismith JH, White MF. Structure of a dimeric crenarchaeal Cas6 enzyme with an atypical active site for CRISPR RNA processing. Biochem J. 2013; 452:223–230. [PubMed: 23527601]

50. Richter H, Lange SJ, Backofen R, Randau L. Comparative analysis of Cas6b processing and CRISPR RNA stability. RNA Biol. 2013; 10:700–707. [PubMed: 23392318]

51. Carte J, Wang R, Li H, Terns RM, Terns MP. Cas6 is an endoribonuclease that generates guide RNAs for invader defense in pro-karyotes. Genes Dev. 2008; 22:3489–3496. [PubMed: 19141480]

52. Haurwitz RE, Jinek M, Wiedenheft B, Zhou K, Doudna JA. Sequence- and structure- specific RNA processing by a CRISPR endonuclease. Science. 2010; 329:1355–1358. [PubMed: 20829488]

53. Cocozaki AI, Ramia NF, Shao Y, Hale CR, Terns RM, Terns MP, Li H. Structure of the Cmr2 subunit of the CRISPR-Cas RNA silencing complex. Structure. 2012; 20:545–553. [PubMed: 22405013]

54. Reeks J, Naismith JH, White MF. CRISPR interference: a structural perspective. Biochem J. 2013; 453:155–166. [PubMed: 23805973]

55. Reeks J, Graham S, Anderson L, Liu H, White MF, Naismith JH. Structure of the archaeal Cascade subunit Csa5: relating the small subunits of CRISPR effector complexes. RNA Biol. 2013; 10:762–769. [PubMed: 23846216]

56. Barrangou R, Fremaux C, Deveau H, Richards M, Boyaval P, Moineau S, Romero DA, Horvath P. CRISPR provides acquired resistance against viruses in prokaryotes. Science. 2007; 315:1709–1712. [PubMed: 17379808]

57. Garneau JE, Dupuis ME, Villion M, Romero DA, Barrangou R, Boyaval P, Fremaux C, Horvath P, Magadan AH, Moineau S. The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. Nature. 2010; 468:67–71. [PubMed: 21048762]

58. Sapranauskas R, Gasiunas G, Fremaux C, Barrangou R, Horvath P, Siksnys V. The *Streptococcus thermophilus* CRISPR/Cas system provides immunity in *Escherichia coli*. Nucleic Acids Res. 2011; 39:9275–9282. [PubMed: 21813460]

59. Nishimasu H, Ran FA, Hsu PD, Konermann S, Shehata SI, Dohmae N, Ishitani R, Zhang F, Nureki O. Crystal structure of Cas9 in complex with guide RNA and target DNA. Cell. 2014; 156:935–949. [PubMed: 24529477]

60. Jinek M, Jiang F, Taylor DW, Sternberg SH, Kaya E, Ma E, Anders C, Hauer M, Zhou K, Lin S, Kaplan M, Iavarone AT, Charpentier E, Nogales E, Doudna JA. Structures of Cas9 endonucleases reveal RNA-mediated conformational activation. Science. 2014; 343:1247997. [PubMed: 24505130]

61. Mulepati S, Bailey S. Structural and biochemical analysis of nuclease domain of clustered regularly interspaced short palindromic repeat (CRISPR)-associated protein 3 (Cas3). J Biol Chem. 2011; 286:31896–31903. [PubMed: 21775431]

62. Beloglazova N, Petit P, Flick R, Brown G, Savchenko A, Yakunin AF. Structure and activity of the Cas3 HD nuclease MJ0384, an effector enzyme of the CRISPR interference. EMBO J. 2011; 30:4616–4627. [PubMed: 22009198]

63. Chakrabarti A, Desai P, Wickstrom E. Transposon Tn7 protein TnsD binding to *Escherichia coli* attTn7 DNA and its eukaryotic orthologs. Biochemistry. 2004; 43:2941–2946. [PubMed: 15005630]

64. Kholodii GY, Mindlin SZ, Bass IA, Yurieva OV, Minakhina SV, Nikiforov VG. Four genes, two ends, and a res region are involved in transposition of Tn5053: a paradigm for a novel family of transposons carrying either a mer operon or an integron. Mol Microbiol. 1995; 17:1189–1200. [PubMed: 8594337]

65. Jackson RN, Lavin M, Carter J, Wiedenheft B. Fitting CRISPR-associated Cas3 into the helicase family tree. Curr Opin Struct Biol. 2014; 24:106–114. [PubMed: 24480304]

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

67. Fonfara I, Le Rhun A, Chylinski K, Makarova KS, Lecrivain AL, Bzdrenga J, Koonin EV, Charpentier E. Phylogeny of Cas9 determines functional exchangeability of dual-RNA and Cas9 among orthologous type II CRISPR-Cas systems. Nucleic Acids Res. 2014; 42:2577–2590. [PubMed: 24270795]

68. Nam KH, Kurinov I, Ke A. Crystal structure of clustered regularly interspaced short palindromic repeats (CRISPR)-associated Csn2 protein revealed Ca2 +-dependent double-stranded DNA binding activity. J Biol Chem. 2011; 286:30759–30768. [PubMed: 21697083]

69. Koo Y, Jung DK, Bae E. Crystal structure of *Streptococcus pyogenes* Csn2 reveals calcium-dependent conformational changes in its tertiary and quaternary structure. PLoS One. 2012; 7:e33401. [PubMed: 22479393]

70. Arslan Z, Wurm R, Brener O, Ellinger P, Nagel-Steger L, Oesterhelt F, Schmitt L, Willbold D, Wagner R, Gohlke H, Smits SH, Pul U. Double-strand DNA end- binding and sliding of the toroidal CRISPR- associated protein Csn2. Nucleic Acids Res. 2013; 41:6347–6359. [PubMed: 23625968]

71. Lee KH, Lee SG, Eun Lee K, Jeon H, Robinson H, Oh BH. Identification, structural, and biochemical characterization of a group of large Csn2 proteins involved in CRISPR-mediated bacterial immunity. Proteins. 2012; 80:2573–2582. [PubMed: 22753072]

72. Wei C, Liu J, Yu Z, Zhang B, Gao G, Jiao R. TALEN or Cas9 - rapid, efficient and specific choices for genome modifications. J Genet Genomics. 2013; 40:281–289. [PubMed: 23790627]

73. Pennisi E. The CRISPR craze. Science. 2013; 341:833–836. [PubMed: 23970676]

74. Anantharaman V, Iyer LM, Aravind L. Presence of a classical RRM-fold palm domain in Thg1-type 3′-5′ nucleic acid polymerases and the origin of the GGDEF and CRISPR polymerase domains. Biol Direct. 2010; 5:43. [PubMed: 20591188]

75. Pei J, Grishin NV. GGDEF domain is homologous to adenylyl cyclase. Proteins. 2001; 42:210–216. [PubMed: 11119645]

76. Makarova KS, Aravind L, Grishin NV, Rogozin IB, Koonin EV. A DNA repair system specific for thermophilic Archaea and bacteria predicted by genomic context analysis. Nucleic Acids Res. 2002; 30:482–496. [PubMed: 11788711]

77. Zhu X, Ye K. Crystal structure of Cmr2 suggests a nucleotide cyclase-related enzyme in type III CRISPR-Cas systems. FEBS Lett. 2012; 586:939–945. [PubMed: 22449983]

78. Nickel L, Weidenbach K, Jager D, Backofen R, Lange SJ, Heidrich N, Schmitz RA. Two CRISPR-Cas systems in Methanosarcina mazei strain Go1 display common processing features despite belonging to different types I and III. RNA Biol. 2013; 10:779–791. [PubMed: 23619576]

79. Marraffini LA, Sontheimer EJ. CRISPR interference limits horizontal gene transfer in staphylococci by targeting DNA. Science. 2008; 322:1843–1845. [PubMed: 19095942]

80. White MF. Structure, function and evolution of the XPD family of iron-sulfur- containing 5′ → 3′ DNA helicases. Biochem Soc Trans. 2009; 37:547–551. [PubMed: 19442249]

81. Makarova KS, Wolf YI, Forterre P, Prangishvili D, Krupovic M, Koonin EV. Dark matter in archaeal genomes: a rich source of novel mobile elements, defense systems and secretory complexes. Extremophiles. 2014; 18:877–893. [PubMed: 25113822]

82. Makarova KS, Wolf YI, Snir S, Koonin EV. Defense islands in bacterial and archaeal genomes and prediction of novel defense systems. J Bacteriol. 2011; 193:6039–6056. [PubMed: 21908672]

83. Datsenko KA, Pougach K, Tikhonov A, Wanner BL, Severinov K, Semenova E. Molecular memory of prior infections activates the CRISPR/Cas adaptive bacterial immunity system. Nat Commun. 2012; 3:945. [PubMed: 22781758]

84. Kim TY, Shin M, Huynh Thi Yen L, Kim JS. Crystal structure of Cas1 from Archaeoglobus fulgidus and characterization of its nucleolytic activity. Biochem Biophys Res Commun. 2013; 441:720–725. [PubMed: 24211577]

85. 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]

86. 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]

87. Lintner NG, Kerou M, Brumfield SK, Graham S, Liu H, Naismith JH, Sdano M, Peng N, She Q, Copie V, Young MJ, White MF, Lawrence CM. Structural and functional characterization of an archaeal clustered regularly interspaced short palindromic repeat (CRISPR)-associated complex for antiviral defense (CASCADE). J Biol Chem. 2011; 286:21643–21656. [PubMed: 21507944]

88. Shao Y, Cocozaki AI, Ramia NF, Terns RM, Terns MP, Li H. Structure of the Cmr2- Cmr3 subcomplex of the Cmr RNA silencing complex. Structure. 2013; 21:376–384. [PubMed: 23395183]

89. Jore MM, Brouns SJ, van der Oost J. RNA in defense: CRISPRs protect prokaryotes against mobile genetic elements. Cold Spring Harb Perspect Biol. 2012; 4 pii: a003657.

90. Hrle A, Su AA, Ebert J, Benda C, Randau L, Conti E. Structure and RNA-binding properties of the type III-A CRISPR-associated protein Csm3. RNA Biol. 2013; 10:1670–1678. [PubMed: 24157656]

91. Osawa T, Inanaga H, Numata T. Crystal structure of the Cmr2-Cmr3 subcom-plex in the CRISPR-Cas RNA silencing effector complex. J Mol Biol. 2013; 425:3811–3823. [PubMed: 23583914]

92. Quax TE, Wolf YI, Koehorst JJ, Wurtzel O, van der Oost R, Ran W, Blombach F, Makarova KS, Brouns SJ, Forster AC, Wagner EG, Sorek R, Koonin EV, van der Oost J. Differential translation tunes uneven production of operon-encoded proteins. Cell Rep. 2013; 4:938–944. [PubMed: 24012761]

93. Steitz TA. The structural basis of the transition from initiation to elongation phases of transcription, as well as translocation and strand separation, by T7 RNA polymerase. Curr Opin Struct Biol. 2004; 14:4–9. [PubMed: 15102443]

94. Anantharaman V, Makarova KS, Burroughs AM, Koonin EV, Aravind L. Comprehensive analysis of the HEPN superfamily: identification of novel roles in intra- genomic conflicts, defense, pathogenesis and RNA processing. Biol Direct. 2013; 8:15. [PubMed: 23768067]

95. Penner M, Morad I, Snyder L, Kaufmann G. Phage T4-coded Stp: double-edged effector of coupled DNA and tRNA-restriction systems. J Mol Biol. 1995; 249:857–868. [PubMed: 7791212]

96. Wang C, Villion M, Semper C, Coros C, Moineau S, Zimmerly S. A reverse transcriptase-related protein mediates phage resistance and polymerizes untemplated DNA in vitro. Nucleic Acids Res. 2011; 39:7620–7629. [PubMed: 21676997]

97. Kim YK, Kim YG, Oh BH. Crystal structure and nucleic acid-binding activity of the CRISPR-associated protein Csx1 of *Pyrococcus furiosus*. Proteins. 2013; 81:261–270. [PubMed: 22987782]

98. Lintner NG, Frankel KA, Tsutakawa SE, Alsbury DL, Copie V, Young MJ, Tainer JA, Lawrence CM. The structure of the CRISPR-associated protein Csa3 provides insight into the regulation of the CRISPR/Cas system. J Mol Biol. 2011; 405:939–955. [PubMed: 21093452]

99. Hein S, Scholz I, Voss B, Hess WR. Adaptation and modification of three CRISPR loci in two closely related cyanobacteria. RNA Biol. 2013; 10:852–864. [PubMed: 23535141]

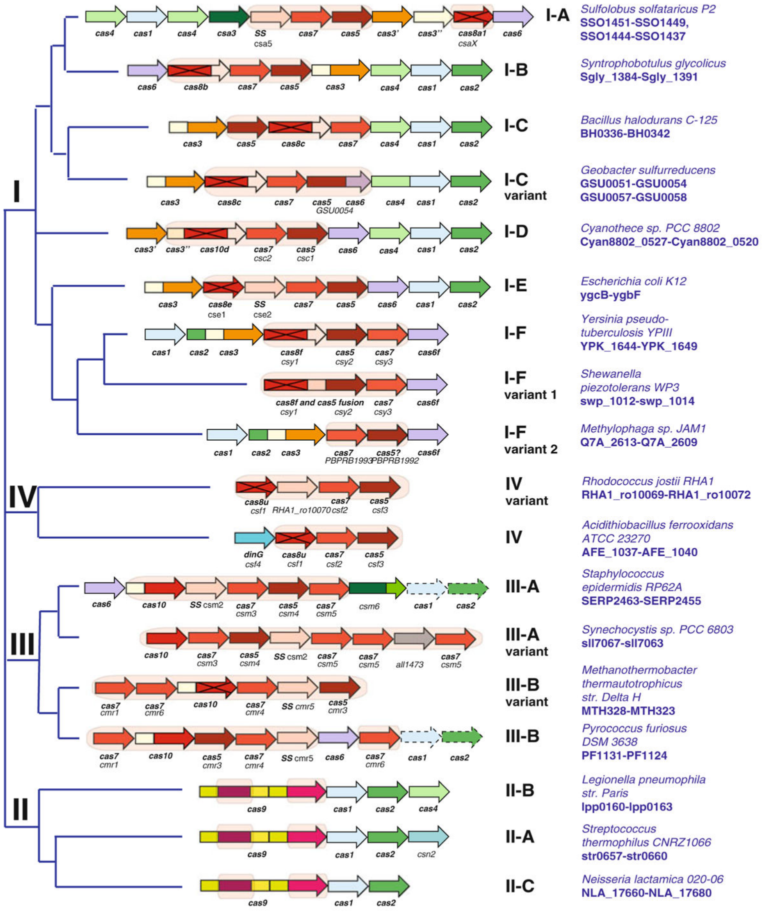

Fig. 1.

Classification and organization of CRISPR-Cas systems. Typical operon organization is shown for each CRISPR-Cas subtype. For each CRISPR-Cas subtype, a representative genome and the respective gene locus tag names are indicated. Homologous genes are *color coded* and identified by a family name. Names follow the classification from [5]. See also details in [30]. Names in *bold* are proposed systematic names; "legacy names" are in regular font. Abbreviations: *LS* large subunit (including subfamilies of Cas10, Cas8, Cse1, Csy1), *SS* small subunit (including Cmr2, Cmr5, Cse2). Genes coding for inactivated large subunits are indicated by *crosses*. Genes and domain components for effector complexes are highlighted by *pink background*

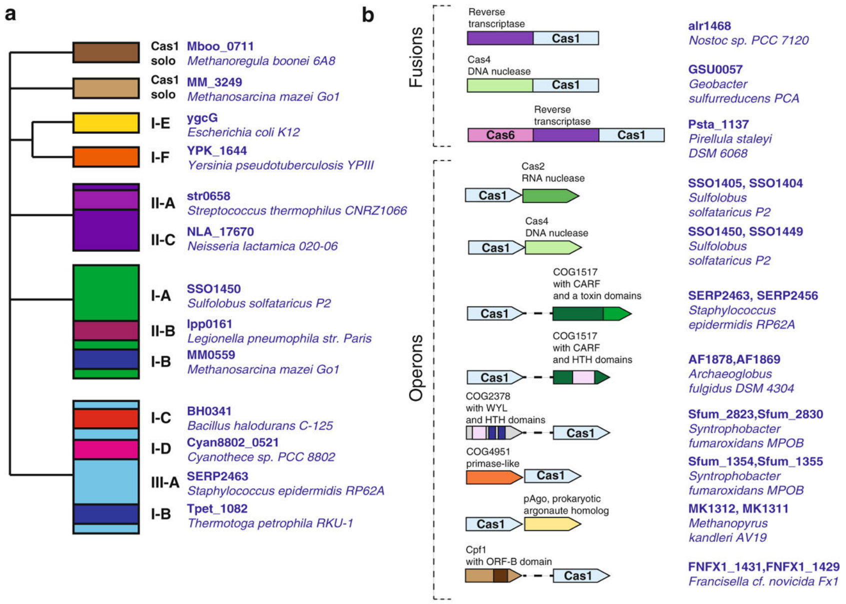
a
Cas1 solo
Mboo_0711
Methanoregula boonei 6A8
Cas1 solo
MM_3249
Methanosarcina mazei Go1
I-E
ygcG
Escherichia coli K12
I-F
YPK_1644
Yersinia pseudotuberculosis YPIII
II-A
str0658
Streptococcus thermophilus CNRZ1066
II-C
NLA_17670
Neisseria lactamica 020-06
I-A
SSO1450
Sulfolobus solfataricus P2
II-B
lpp0161
Legionella pneumophila str. Paris
I-B
MM0559
Methanosarcina mazei Go1
I-C
BH0341
Bacillus halodurans C-125
I-D
Cyan8802_0521
Cyanothece sp. PCC 8802
III-A
SERP2463
Staphylococcus epidermidis RP62A
I-B
Tpet_1082
Thermotoga petrophila RKU-1
b
Fusions
Reverse transcriptase
Cas1
alr1468
Nostoc sp. PCC 7120
Cas4 DNA nuclease
Cas1
GSU0057
Geobacter sulfurreducens PCA
Reverse transcriptase
Cas6
Cas1
Psta_1137
Pirellula staleyi DSM 6068
Operons
Cas2 RNA nuclease
Cas1
SSO1405, SSO1404
Sulfolobus solfataricus P2
Cas4 DNA nuclease
Cas1
SSO1450, SSO1449
Sulfolobus solfataricus P2
COG1517 with CARF and a toxin domains
Cas1
SERP2463, SERP2456
Staphylococcus epidermidis RP62A
COG1517 with CARF and HTH domains
Cas1
AF1878,AF1869
Archaeoglobus fulgidus DSM 4304
COG2378 with WYL and HTH domains
Cas1
Sfum_2823,Sfum_2830
Syntrophobacter fumaroxidans MPOB
COG4951 primase-like
Cas1
Sfum_1354,Sfum_1355
Syntrophobacter fumaroxidans MPOB
pAgo, prokaryotic argonaute homolog
Cas1
MK1312, MK1311
Methanopyrus kandleri AV19
Cpf1 with ORF-B domain
Cas1
FNFX1_1431,FNFX1_1429
Francisella cf. novicida Fx1

Fig. 2.

Phylogeny of Cas1 and its associations with other genes. (**a**) Schematic representation of Cas1 phylogeny (the complete tree and details of the tree reconstruction are available in [31]). The branches are *colored* according to the automatic assignment of *cas1* genes to CRISPR-Cas subtypes based on the analysis of ten up- and ten downstream genes. (**b**) Cas1 fusions and operonic associations

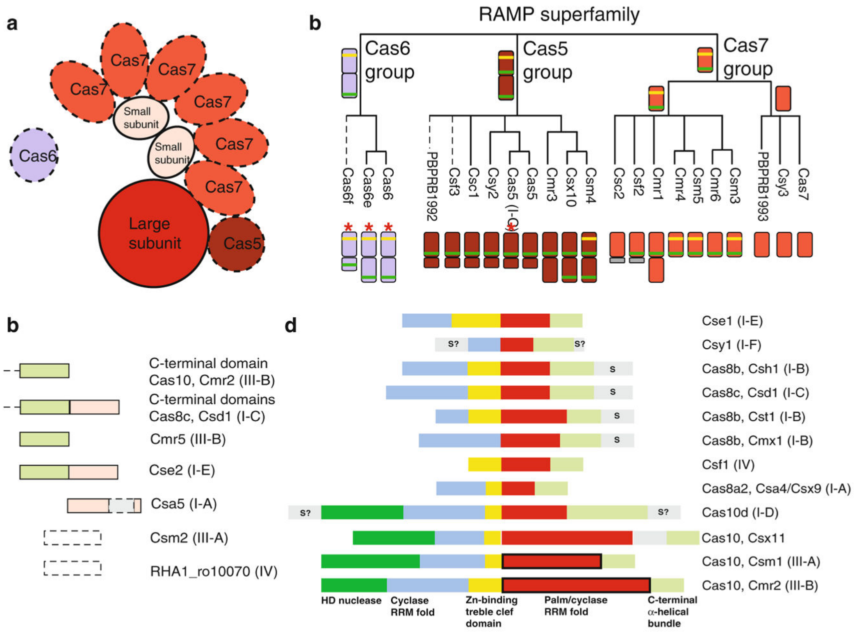

Fig. 3.

General organization of effector complexes in different types of CRISPR-Cas systems. (**a**) The generalized model of subunit composition of effector complexes of type I, III, and IV systems. Color coding is the same as in Fig. 2. The subunits that belong to the RAMP superfamily are shown by *dashed circles*. (**b**) Classification of the RAMP superfamily into three families. The tree-like scheme of RAMP relationships is based on the sequence similarity, structural features, and neighborhood analysis. Unresolved relationships are shown as multifurcations and tentative assignments are shown by *broken lines*. Glycine-rich loops are shown by *green lines*. The conserved histidines, suggesting catalytic activity of some of the RAMP proteins, are shown by *yellow lines*. Protein families shown to be active ribonucleases are marked by an *asterisk*. The deteriorated RRM domain is shown by the *gray rectangle* (i.e., Csf2 and Csc2). The predicted ancestral domain configuration is shown for each major node. (**c**) Domain organization of the small subunits of different subtypes of type I, III, and IV CRISPR-Cas systems. Homologous domains are color coded. *Dashed outline empty boxes* show that the structure similarity of these small subunits is unknown. The *gray box* shows a unique beta-stranded insertion. (**d**) Domain organization of the large subunits of different type I and III CRISPR-Cas systems. The palm-like domains of Cas10 proteins with intact cyclase/polymerase catalytic motifs are shown with a *black outline*. The letter “S” marks the regions that could be homologous to small subunits of Cascade complexes encoded as separated genes in type III systems, I-E subtype, and some systems of the I-A subtype

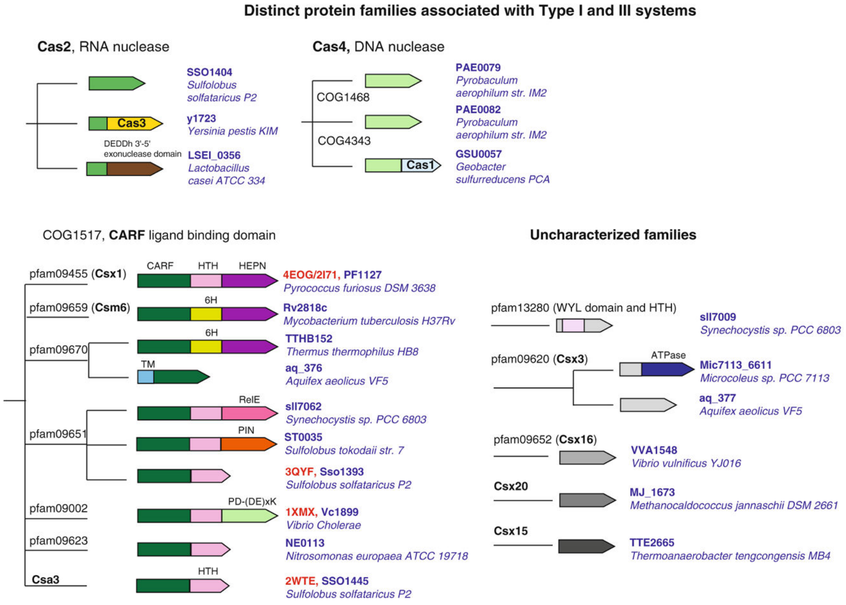
Distinct protein families associated with Type I and III systems
Cas2, RNA nuclease
SSO1404
Sulfolobus solfataricus P2
Cas3
y1723
Yersinia pestis KIM
DEDDh 3'-5' exonuclease domain
LSEI_0356
Lactobacillus casei ATCC 334
Cas4, DNA nuclease
COG1468
PAE0079
Pyrobaculum aerophilum str. IM2
COG4343
PAE0082
Pyrobaculum aerophilum str. IM2
Cas1
GSU0057
Geobacter sulfurreducens PCA
COG1517, CARF ligand binding domain
pfam09455 (Csx1)
CARF
HTH
HEPN
4EOG/2I71, PF1127
Pyrococcus furiosus DSM 3638
pfam09659 (Csm6)
6H
Rv2818c
Mycobacterium tuberculosis H37Rv
pfam09670
6H
TTHB152
Thermus thermophilus HB8
TM
aq_376
Aquifex aeolicus VF5
RelE
sll7062
Synechocystis sp. PCC 6803
pfam09651
PIN
ST0035
Sulfolobus tokodaii str. 7
3QYF, Sso1393
Sulfolobus solfataricus P2
pfam09002
PD-(DE)xK
1XMX, Vc1899
Vibrio Cholerae
pfam09623
NE0113
Nitrosomonas europaea ATCC 19718
Csa3
HTH
2WTE, SSO1445
Sulfolobus solfataricus P2
Uncharacterized families
pfam13280 (WYL domain and HTH)
sll7009
Synechocystis sp. PCC 6803
pfam09620 (Csx3)
ATPase
Mic7113_6611
Microcoleus sp. PCC 7113
aq_377
Aquifex aeolicus VF5
pfam09652 (Csx16)
VVA1548
Vibrio vulnificus YJ016
Csx20
MJ_1673
Methanocaldococcus jannaschii DSM 2661
Csx15
TTE2665
Thermoanaerobacter tengcongensis MB4

Fig. 4.

Associated immunity components of CRISPR-Cas systems. CRISPR-Cas gene names follow the nomenclature and classification from [5] and are shown in *bold*. An identifier of the sequence profile from COG and PFAM databases is provided when available. Examples of proteins for each distinct family are provided in the form of locus tag and organism name. The PDB code is shown in *red* when available. The genes are depicted as *block arrows* and domains as *block squares* . Homologous genes and domains are shown by *arrows* of the same color. The following domain names are indicated above the corresponding shape when shown for the first time: PD-(D/E)xK, restriction endonuclease superfamily protein, predicted DNA nuclease; HTH, DNA-binding helix-turn- helix domain; HEPN, HEPN domain, see details in [94]; RelE, RelE superfamily protein, predicted ribonuclease; PIN, PIN superfamily ribonuclease; 6H, helical middle domain in some of the COG1517 superfamily proteins

Table 1

Classification of CRISPR-Cas subtypes

System subtype	Mono-phyletic on Cas1 tree	Signature genes: *strong*/ weak^a (other name)	Comment
I-A	No	Cas8a2, Csa5	Cas3 is often split into helicase domain Cas3′ and HD nuclease domain Cas3″ and a separate gene for small subunit Csa5 is often present. There are hybrid systems having a Cas1-Cas2-Cas4 module similar to I-A but a Cascade module more similar to type I-B. They are paraphyletic to I-A clade. The latter systems often have the large subunits described as Cas8a1; however, they should be rather classified as I-B.
I-B	No	Cas8b	I-B systems belong to two distinct clades on the Cas1 tree. One is paraphyletic group to I-A on the Cas1 tree, and another is paraphyletic to III-A. Usually the Cas3 gene is not split.
I-C	No	Cas8c	These systems usually do not have a *cas6* gene. Cas5 is catalytically active and replaces Cas6 function.
I-C variant	No	**GSU0054** (Cas5 group RAMP)	These systems usually do not have *cas6*. Cas5 has several specific insertions or fusions, but is likely to be catalytically active. There are systems with different subfamilies of the large subunit, which are often severely deteriorated and sometimes even missing.
I-D	No	**Cas10d**	The HD domain is associated with the large subunit rather than with Cas3, although it does not have the circular permutation of the motifs like the HD domain fused with Cas10 in type III systems.
I-E	Yes	**Cse1, Cse2**	The *cas4* gene is not associated with this system.
I-F	Yes	**Csy1, Csy2, Csy3, Cas6f**	The *cas4* gene is not associated with this system, and *cas2* is fused to *cas3*. There is no separate gene for a small subunit, which is either missing or fused to the large subunit.
I-F variant 1	N/A	**Csy1/Csy2 fusion**	The *cas1-cas2-cas3* genes are not present. Usually three genes (*csy1/csy2* fusion, *csy3*, and *cas6f*) are present in an operon, which is often found next to *tniQ*/*tnsD* family genes. These are potentially mobile effector complexes.
I-F variant 2	Yes	**PBPRB1993** **PBPRB1992**	A derived variant of I-F different from predicted group 5 (PBPRB1992) and group 7 genes (PBPRB1993).
II-A	Yes	Csn2	Monophyletic group on Cas9 tree. There are four genes in these operons with *csn2* gene in addition to *cas1_2_9*. There are at least five distinct families of Csn2.
II-B	Yes	Cas9 (Csx12 subfamily)	Monophyletic group on Cas9 tree with four gene operons containing *cas4* in addition to *cas1_2_9*.
II-C	No	N/A	Only three genes are present in the II-C operon—*cas1_2_9*.
III-A	No	Csm2 (small subunit)	Also known as the Csm module and Cas10 usually has active catalytic motifs. III-A is associated with several Cas7 group RAMPs and is often linked to *csm6* which has CARF and C-terminal HEPN domain. Might be associated with Cas1-Cas2 pairs of different origin.
III-A variant	No	Csx10 all1473	These have many modifications. Csx10 is a fusion of Cas5 and Cas7 proteins. The specific gene all1473 is likely to be a component of Cascade but is not similar to any known Cas proteins. The large subunit is often lacking the HD domain. Csx10 could be fused to the small subunit and Cas7 group RAMPs are often fused and have large insertions.
III-B	No	Cmr5 (small subunit)	Also known as the Cmr (or RAMP) module and Cas10 often has active catalytic motifs. These systems are usually associated with several Cas7 group RAMPs and are rarely present in a genome as a stand-alone system They are usually not linked to *cas1-cas2* gene pair and Cmr1 has a duplication of RAMPs both from the Cas7 group.
III-B variant	No	MTH326 (Cas10 or Csx11)	The large subunit is often inactivated and some Cmr1 family proteins possess only one RAMP domain.

System subtype	Mono-phyletic on Cas1 tree	Signature genes: *strong*/ weak[a] (other name)	Comment
IV	N/A	DinG (Csf4)	These systems possess a gene for a very reduced large subunit *csf1.* This variant, in addition to predicted large subunit and Cas7 and Cas5 group RAMPs, often encodes a gene for a DinG-like helicase.
IV-variant	N/A	**RHA1_ro10070**	These variants possess a gene for a very reduced large subunit *csf1* and RHA1_ro10070-like proteins are predicted small subunits of effector complexes. These systems are found mostly in Actinobacteria and often on plasmids.

[*] **Strong**/weak—is the characteristic of the signature protein family with respect to subtype recognition/classification ability using the respective profile. Strong means that it has a relatively high specificity and high selectivity, i.e., a reliable signature, whereas weak means that search for this family yields a high level of either false positives or false negatives, but nevertheless the family remains the best available signature for a particular subtype

Table 2

Structures, domain architectures, and functions of the core components of CRISPR-Cas systems

Family	Biochemical/*in silico* evidence	Examples of available structures and structural features
Cas1	Metal-dependent deoxyribonuclease [35, 36]; deletion of Cas1 in *E. coli* results in increased sensitivity to DNA damage and impaired chromosomal segregation [37].	PDB: 3GOD, 3LFX, 2YZS Unique fold with two domains: N-terminal β-stranded domain and catalytic C-terminal α-helical domain
Cas2	RNase specific to U-rich regions [38]. Double-stranded DNase [39].	PDB: 2IVY, 2I8E, and 3EXC RRM (ferredoxin) fold
Cas3 (helicase and HD domain)	Single-stranded DNA nuclease (HD domain) and ATP- dependent helicase [40]; required for interference [20].	
Cas3″ stand-alone HD nuclease	Metal-dependent deoxyribonuclease specific for double- stranded oligonucleotides [41].	PDB: 3S4L and 3SKD
Cas4	PD-(DE)xK superfamily nuclease with three-cysteine C-terminal cluster [7]; possesses 5′-ssDNA exonuclease activity [42].	PDB: 4IC1
Cas5	Subunit of Cascade complex interacting with large subunit and Cas7 subunit and binding the 5′-handle of crRNA [20, 22, 23, 43–45]. In the subtype I-C system Cas5 is the ribonuclease that replaces Cas6 function [46].	PDB: 3KG4; 3VZI; 3VZH Two domains of RRM (ferredoxin) fold, the C-terminal domain is deteriorated in many Cas5 protein of type I; RAMP superfamily
Cas6	Metal-independent endoribonuclease that generates crRNAs[20, 44, 47–52].	PDB: 2XLJ, 1WJ9,3I4H, 4C8Z, 4DZD Two domains of RRM (ferredoxin) fold, RAMP superfamily
Cas7	Subunit of Cascade complexes binding crRNA [20, 22, 23, 43, 45]; often present in Cascade complexes in several copies.	PDB: 3PS0, 4N0L RRM (ferredoxin) fold with subdomains, RAMP superfamily
Cas8abcef (large subunit)	Subunit of Cascade complex, involved in PAM recognition [16–18, 20, 43].	PDB: 4AN8
Cas10 (large subunit)	Subunit of Cascade (Cmr) complex [22, 23, 45, 47].	PDB: 3UNG, 4DOZ Two domains homologous to palm domain polymerases and cyclases, both belonging to RRM (ferredoxin) fold; Zn finger containing domain and C-terminal alpha helical domain [53]; Fusion: HD nuclease domain
Small subunit	Small, mostly alpha helical protein, subunit of Cascade complex [20, 22, 23, 44, 45, 47, 54, 55].	PDB: 2ZCA (Cse2); 2ZOP, 2OEB (Cmr5); 3ZC4 (Csa5) Cse2 has two alpha helical bundle-like domains; Cmr5 has a domain matching N-terminal domain of Cse1 and Csa5 has a domain matching C-terminal domain of Cse2
Cas9	In type II CRISPR-Cas systems, Cas9 is sufficient both to generate crRNA and to cleave the target DNA [56, 57], although it requires the help of a housekeeping gene coding for RNase III and a special gene tracrRNA encoded in the respective CRISPR-Cas locus [13]. Both the RuvC and HNH nuclease domains of Cas9 are involved in the cleavage of the target DNA [15, 58].	PDB: 4OGC, 4OO8, 4CMP Cas9 has several subdomains and adopts a two-lobed general structure. Beyond two catalytic nuclease domain its subdomains do not appear to be similar to other known protein structures [59, 60]