The CRISPR Journal
Volume 1, Number 5, 2018
© Mary Ann Liebert, Inc.
DOI: 10.1089/crispr.2018.0033

The CRISPR Journal

REVIEW ARTICLE

# Classification and Nomenclature of CRISPR-Cas Systems: Where from Here?

Kira S. Makarova, Yuri I. Wolf, and Eugene V. Koonin*

## Abstract

As befits an immune mechanism, CRISPR-Cas systems are highly variable with respect to Cas protein sequences, gene composition, and organization of the genomic loci. Optimal classification of CRISPR-Cas systems and rational nomenclature for CRISPR-associated genes are essential for further progress of CRISPR research. These are highly challenging tasks because of the complexity of CRISPR-Cas and their fast evolution, including frequent module shuffling, as well as the lack of universal markers for a consistent evolutionary classification. The complexity and variability of CRISPR-Cas systems necessitate a multipronged approach to classification and nomenclature. We present a brief summary of the current state of the art and discuss further directions in this area.

## Outline of CRISPR Diversity and Classification

Typical of antivirus defense mechanisms that are locked into a perennial evolutionary arms race with pathogens,[1,2] CRISPR*-Cas systems show remarkable diversity in terms of gene composition, genomic locus architecture, and the actual sequences, even in the core genes shared by many CRISPR-Cas variants.[3–5] To elucidate the origins and evolution of this diversity and, more practically, to keep track of new variants and to achieve coherent annotation of CRISPR-*cas* loci in microbial genomes, a rational and relatively simple classification of CRISPR-Cas systems is essential. This is, however, much easier said than done. The repeats themselves show some clustering in the sequence space[6,7] but clearly do not contain enough information to serve as the basis of a robust classification. There are no universal *cas* genes, so a classification based on a single phylogenetic tree is out of the question. Worse yet, a unified approach is hampered by the pronounced modularity of CRISPR-*cas* evolution, in particular the frequent shuffling of the adaptation and effector modules.[3,8] Accordingly, the efforts on CRISPR-Cas classification have adopted a combined, semi-formal approach that takes into account the signature *cas* genes that are specific for individual types and subtypes of CRISPR-Cas, sequence similarity between multiple shared Cas proteins, the phylogeny of Cas1 (the most highly conserved Cas protein), and the organization of the genes in the CRISPR-*cas* loci.[3,9,10] The combined application of these criteria resulted in the current classification that partitions the CRISPR-Cas systems into two distinct classes that differ in the design principles of the effector module (Fig. 1).

Class 1 encompasses the most common and diversified type I, type III that also includes diverse variants and is represented in numerous archaea but is less common in bacteria, and the comparatively rare type IV that consists of rudimentary CRISPR-*cas* loci that lack the effector nuclease and in most case the adaptation module as well. The effector modules of type I and type III CRISPR-Cas are elaborate complexes that consist of multiple Cas protein subunits. The effector complexes of type I and type III share little readily detectable sequence conservation but nevertheless consist of analogous and in many cases homologous protein subunits and possess highly similar architectures, as revealed by cryo-electron microscopy. The backbones of all these complexes are composed of paralogous repeat-associated mysterious proteins (RAMPs), such as Cas7 and Cas5, which show minimum sequence conservation but all contain the RNA recognition motif (RRM) fold and diagnostic sequence, the C-terminal glycine-rich loop, along with additional ‘‘large’’ and ‘‘small’’ subunits.[11–18] Another RAMP, Cas6, is loosely associated with the effector complex and in most Class 1 systems is the repeat-specific

*Clustered Regularly Interspaced Short Palindromic Repeats.

National Center for Biotechnology Information, National Library of Medicine, Bethesda, Maryland.

*Address correspondence to: Eugene V. Koonin, 8600 Rockville Pike, Bethesda, MD 20894, E-mail: koonin@ncbi.nlm.nih.gov

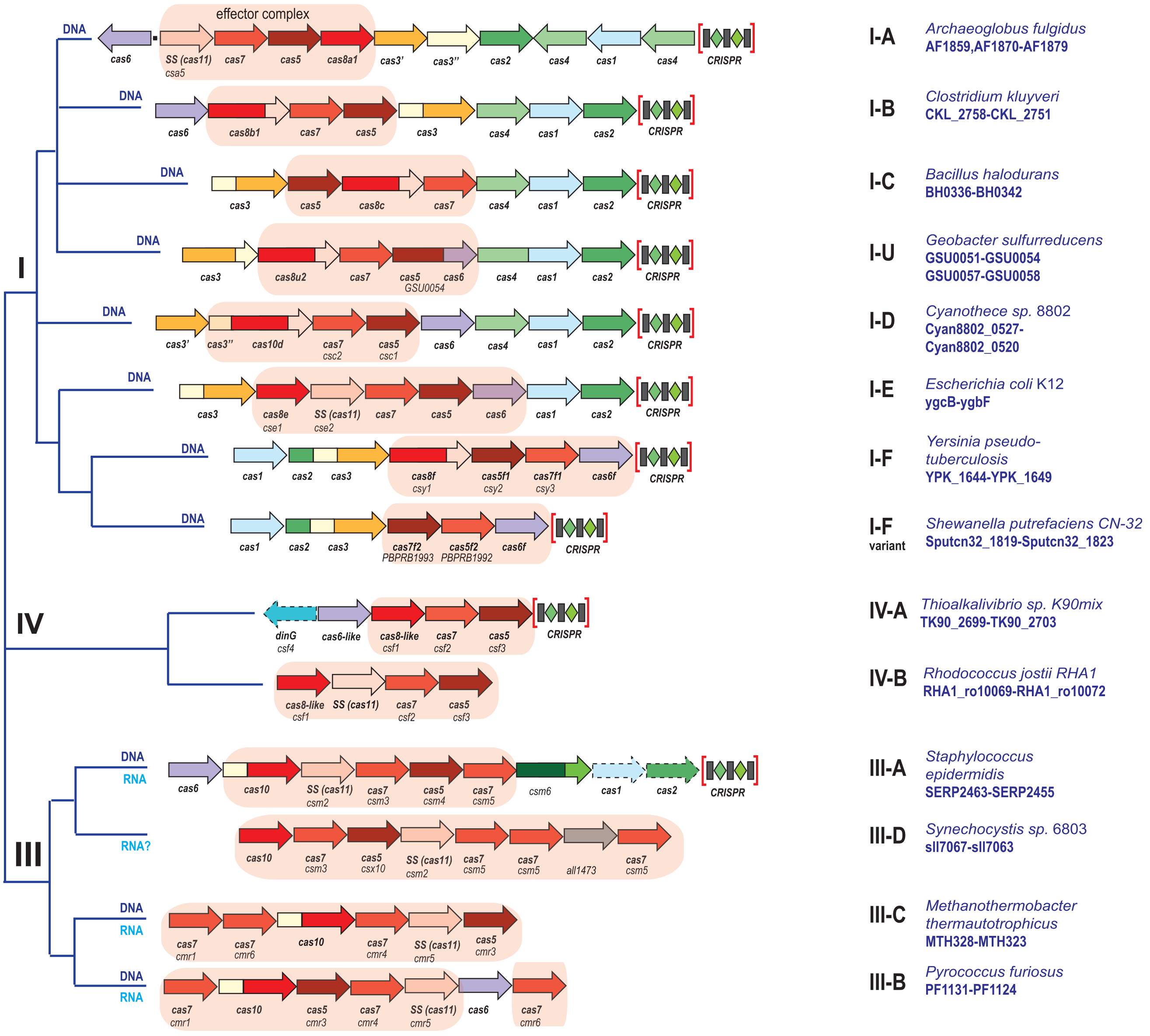


**FIG. 1.** Updated classification of Class 1 CRISPR-Cas systems. Typical operon organization is shown for each CRISPR–Cas system subtype. For each CRISPR–Cas subtype, a representative genome and the respective gene locus tag names are indicated. Homologous genes are color-coded and identified by the respective family name. The gene names follow the classification from Makarova *et al.*[3] Where both a systematic name and a legacy name are commonly used, the legacy name is given under the systematic name. The Cas11 gene name was reserved and is now used for small subunits. Most of them are found to be homologous and predicted to be fused to Cas8 protein in many type I systems. Predicted or known targets (DNA or RNA, or both) are shown for each subtype. The genes for the Class 1 effector module are shaded. The specific strains of bacteria in which these systems were identified and locus tags for the respective protein-coding genes are also indicated. The dashed border line indicates that the respective genes are functionally dispensable. The figure was modified from Koonin *et al.*,[4] with permission.

RNase that is directly responsible for the pre-crRNA processing.[19,20] Type I and type III differ with respect to the relationship between the Cas protein complex that is involved in pre-cRNA processing and the target-cleaving nuclease, that is, whether the effector nuclease is part of the processing complex. In type I systems, the effector enzyme is an HD nuclease that most often comprises a domain of the Cas3 helicase or in some variants is a stand-alone protein.[21–26] Both the helicase and the nuclease are not components of the processing machinery and are recruited to the effector complex after target binding. In contrast, in type III systems, the effector HD nuclease is a domain of Cas10, the large subunit of the processing complex.[27–29] Furthermore, in type III, catalytically active RAMPs of the Cas7 group cleave RNA targets.[28]

Class 2 systems have a much simpler organization, with the effector module consisting of a single, large, multidomain and multifunctional protein (Fig. 2). Class 2 includes the abundant type II (with the effector protein Cas9, the principal tool of the new generation of genome editing methods), and the much rarer types V and VI, each with a unique architecture of the large effector protein (see Fig. 2 and details below).[3,30]

In this brief article, we discuss the uncertainties, limitations, and possible future developments of this classification.

## Evolving Classification and Nomenclature of CRISPR-Cas Systems

### Complexity and complications

In the early years of CRISPR research, the phylogeny of Cas1 protein has been central for classification of CRISPR-Cas systems.[31] However, with the subsequent fast growth of the collection of genomes containing CRISPR-*cas* loci, numerous inconsistences between the Cas1 phylogeny on the one hand and the organization and phylogenies of the effector module genes on the other hand became apparent. The underlying cause of these discrepancies seems to be extensive shuffling of the adaptation and effector modules.[3,8,32]

Figure 3 shows a phylogenetic tree of Cas1 proteins that includes 2,512 representatives covering the entire diversity of the family. Only a few subtypes remain monophyletic in this tree: I-C, I-F, I-E, and V-A. Furthermore, subtypes II-A and II-C are not particularly prone to module shuffling. However, most of the other systems, especially those of type III and subtypes I-A, I-B, and I-D, appear to be able to combine freely with a variety of adaptation modules.[8,33,34] Subtypes V-C and V-D, the only known CRISPR-Cas systems that lack the *cas2* genes[4,30] encoding the structural subunit of the adaptation complex, form a strongly supported long branch, which is indicative of the common origin and distinct evolutionary trajectories of these unusual adaptation modules. In addition, the tree contains two clades that consist of non-CRISPR Cas1 homologs: the casposases, which are the transposases of casposons, the distinct type of transposons thought to be the ancestors of the CRISPR-Cas adaptation modules,[35,36] and archaea-specific uncharacterized ‘‘solo’’ Cas1 (many of these, apparently, inactivated). The central position of the casposases in the tree seems compatible with their ancestral status.

The monophyly of Cas1 proteins from several subtypes implies co-evolution and likely functional connection between the respective adaptation and effector modules. Typically, these systems have a well-defined protospacer-adjacent motif (PAM) sequence and highly conserved repeats.[6,7,37,38] In other systems, especially those of type III, the adaptation and effector modules are virtually independent, and the latter are compatible with a variety of crRNAs and have weaker PAM sequence recognition requirements or require no PAM at all.[38–40] Thus, although Cas1 phylogeny is not a suitable guide for CRISPR-Cas classification due to the module shuffling, it remains useful for inference of evolutionary trends and prediction of functional features.

Because of the extensive module shuffling, distinctive features of the effector modules play a key role in the current classification of CRISPR-Cas systems. At the level of the currently known two classes and six types, the classification criteria are quite straightforward: the fundamental difference in the organization of the effector modules between the classes and the unique signature genes for each of the types.[3] These signatures include *cas3* for type I, *cas10* for type III, *cas9* for type II, *csf1* (large subunit, *cas8*-like) for type IV, *cas12* for type V, and *cas13* for type VI. The signature proteins are either unrelated to each other or, when they share a conserved domain such as the RuvC-like nuclease domain in types II and V, evolutionary reconstructions strongly suggest that the effector genes have been recruited by the respective CRISPR-Cas systems independently.[30,41]

At the subtype level, things become much more complicated and messier. For some of the subtypes, diagnostic genes can be readily defined. For example, the presence of *dinG* immediately indicates subtype IV-A, whereas *csn2* is specific for subtype II-A. In addition, several other subtype-specific variants of *cas* genes are recognizable by sequence similarity to particular sequence profiles (e.g., among the profiles for Cas5, profile cd09649 is specific for subtype I-A, whereas profile cd09645 is specific for subtype I-E).[3] However, in most cases, multiple profiles are required to describe all variants of the signature Cas proteins, even within one subtype, as shown in Figure 4A for the large subunits of type I systems.

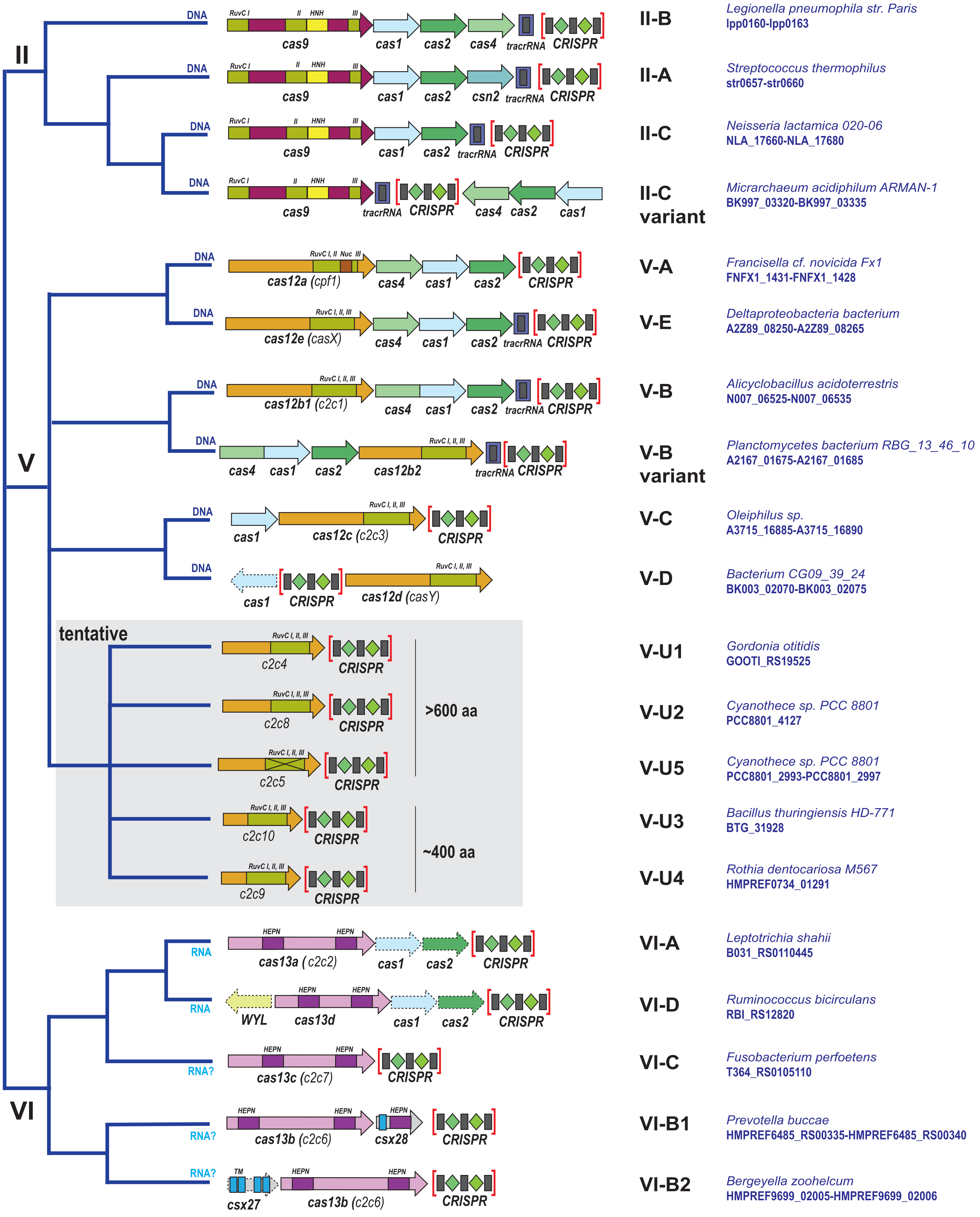


**FIG. 2.** Updated classification of Class 2 CRISPR-Cas systems. RuvC I, RuvC II, and RuvC III are the three distinct motifs that contribute to the nuclease catalytic center. tracrRNA, trans-activating RNA, a helper RNA necessary for pre-crRNA processing and targeting functions; TM, predicted transmembrane segment. The proposed new systematic gene names are shown in red and bold type. Systematic gene names for effector protein candidates are shown below the respective shapes as follows; legacy or old names are also indicated in parentheses. For the V-U5 variant, the inactivation of the RuvC-like nuclease domain is indicated by a cross. The rest of the designations are as in Figure 1. The figure was modified from Koonin *et al.*,[4] with permission.

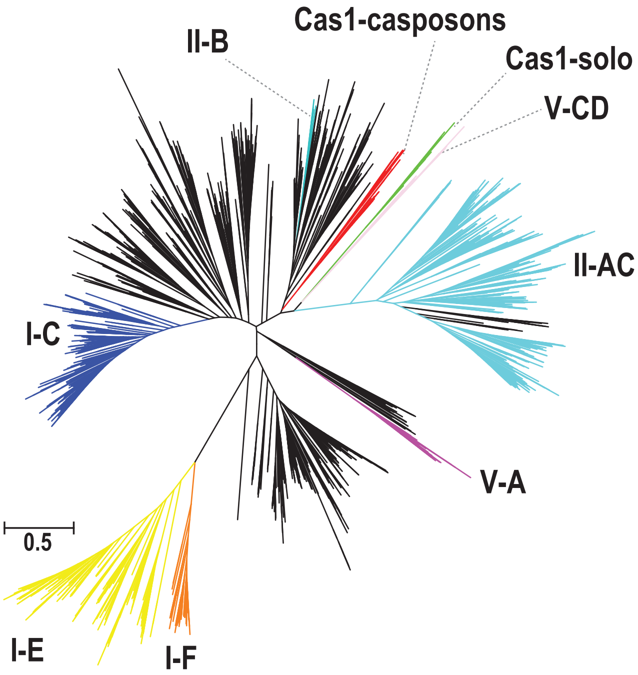


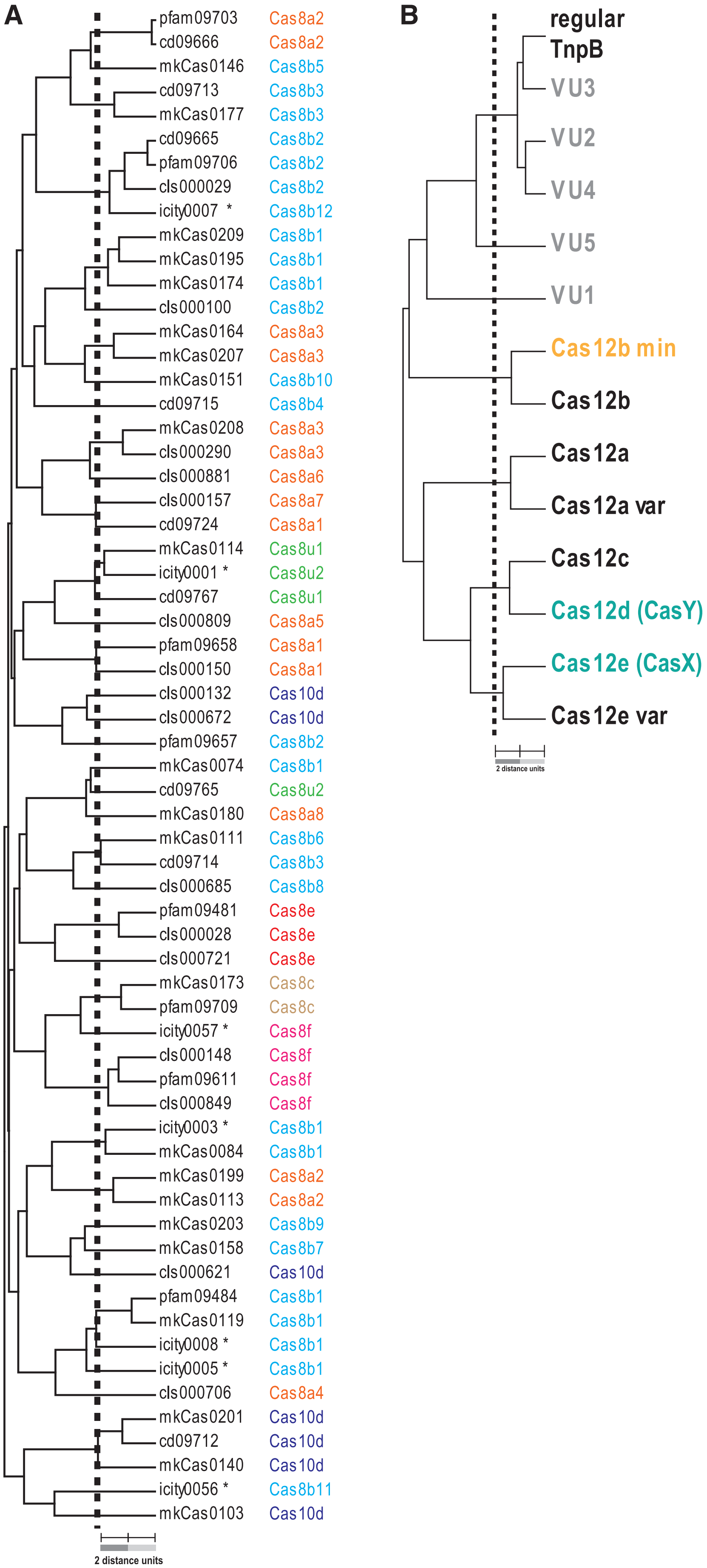
**FIG. 3.** Cas1 phylogeny. The alignment of 2,512 representative Cas1 protein sequences was obtained using iterative clustering and alignment merging of Cas1 sequences (see text). The approximate ML tree was reconstructed using FastTree,[56] with the WAG substitution model and gamma-distributed site rates. Large Cas1 clades that represent (mostly) monophyletic subtype (and other) specific variants are indicated; the other subtypes are scattered across the tree.

There is little or no detectable sequence similarity between different families within the same group, and often several profiles are needed even to identify all instances of describe one variant (see Cas8b1 profiles in Fig. 4A). For example, Cas8 proteins of subtype I-B are currently divided into 13 distinct families that require 27 profiles for complete recognition, and it is expected that this number will grow further with the discovery of additional CRISPR-Cas variants within the same subtype. Obviously, in practice, these details are difficult to follow, which often complicates the identification of subtypes.

Those subtypes for which signature genes are not readily identifiable are defined through comparison of conserved genes and locus organization. This classification is riddled with its fair share of uncertainty and ambiguities. Moreover, the number of CRISPR-Cas variants that could not be classified by the existing approaches is growing. Figure 5 shows examples of such systems. Arguably, the most notable among them are derived variants of I-F and I-B subtypes, which were recruited by Tn7 transposon and apparently have lost the immune function completely.[42] Similarly, subtype I-E variants encoding a STAND superfamily ATPase have lost the *cas3* gene and accordingly the capacity to cleave the target, likely evolving into signal transduction systems.[43] These derived variants challenge the very definition of CRISPR-Cas as adaptive immunity systems so that it becomes questionable whether they should be called CRISPR-Cas systems at all. One possibility is to broaden the definition to include *cas* operons with functions other than adaptive immunity. If the derived variants remain within the CRISPR-Cas classification, it is unclear whether they should stay within the ancestral subtypes or become new ones.

Perhaps the most striking illustrations of classification woes but also some recipes to overcome these come from the recent series of discoveries of new Class 2 CRISPR-Cas systems. In the last few years, the unprecedented success of type II effectors, Cas9 proteins, as genome editing tools has stimulated focused efforts on discovery of new variants of Class 2 systems

**FIG. 4.** Deep relationships between sequence profiles of Cas proteins. **(A)** Relationships between sequence profiles for the type I large subunits. Profile–profile comparisons were performed using HHsearch[48]; scores between two profiles were normalized by the minimum of the self-scores and converted to a distance matrix on the natural log scale. The UPGMA dendrogram was reconstructed from the distance matrix. The dashed line cuts the tree at the depth of 2 ($D=2$ roughly corresponds to the pairwise HHsearch score of $e^{-2D}=0.02$ relative to the self-score). Profile names are colored according to their subtype specificity. According to the current CRISPR-Cas classification and nomenclature,[3] the large subunits are described using the following notation: major type of the large subunit (Cas8, Cas10), a letter that indicates the subtype and a number corresponding to a distinct variant. For example, Cas8b8 is the large subunit of subtype I-B, family 8. **(B)** Relationships between the sequences of the type V effector proteins and the homologous TnpB-like proteins. The dendrogram was constructed using the same procedure as in **(A)**; color highlights the recently discovered variants (minimal Cas12b, CasX and CasY). Proteins from the unclassified type V-U systems are shown in gray. "Cas12a var" includes several sequences typified by KFO67988.1 from Smithella sp. SCADC. "Cas12e var" includes two sequences: GBD34782.1 from bacterium HR35 and A3J58_03210 *Candidatus Sungbacteria* bacterium RIFCSPHIGHO2_02_FULL_52_2.


A
pfam09703 Cas8a2
cd09666 Cas8a2
mkCas0146 Cas8b5
cd09713 Cas8b3
mkCas0177 Cas8b3
cd09665 Cas8b2
pfam09706 Cas8b2
cls000029 Cas8b2
icity0007 * Cas8b12
mkCas0209 Cas8b1
mkCas0195 Cas8b1
mkCas0174 Cas8b1
cls000100 Cas8b2
mkCas0164 Cas8a3
mkCas0207 Cas8a3
mkCas0151 Cas8b10
cd09715 Cas8b4
mkCas0208 Cas8a3
cls000290 Cas8a3
cls000881 Cas8a6
cls000157 Cas8a7
cd09724 Cas8a1
mkCas0114 Cas8u1
icity0001 * Cas8u2
cd09767 Cas8u1
cls000809 Cas8a5
pfam09658 Cas8a1
cls000150 Cas8a1
cls000132 Cas10d
cls000672 Cas10d
pfam09657 Cas8b2
mkCas0074 Cas8b1
cd09765 Cas8u2
mkCas0180 Cas8a8
mkCas0111 Cas8b6
cd09714 Cas8b3
cls000685 Cas8b8
pfam09481 Cas8e
cls000028 Cas8e
cls000721 Cas8e
mkCas0173 Cas8c
pfam09709 Cas8c
icity0057 * Cas8f
cls000148 Cas8f
pfam09611 Cas8f
cls000849 Cas8f
icity0003 * Cas8b1
mkCas0084 Cas8b1
mkCas0199 Cas8a2
mkCas0113 Cas8a2
mkCas0203 Cas8b9
mkCas0158 Cas8b7
cls000621 Cas10d
pfam09484 Cas8b1
mkCas0119 Cas8b1
icity0008 * Cas8b1
icity0005 * Cas8b1
cls000706 Cas8a4
mkCas0201 Cas10d
cd09712 Cas10d
mkCas0140 Cas10d
icity0056 * Cas8b11
mkCas0103 Cas10d
2 distance units
B
regular TnpB
VU3
VU2
VU4
VU5
VU1
Cas12b min
Cas12b
Cas12a
Cas12a var
Cas12c
Cas12d (CasY)
Cas12e (CasX)
Cas12e var
2 distance units

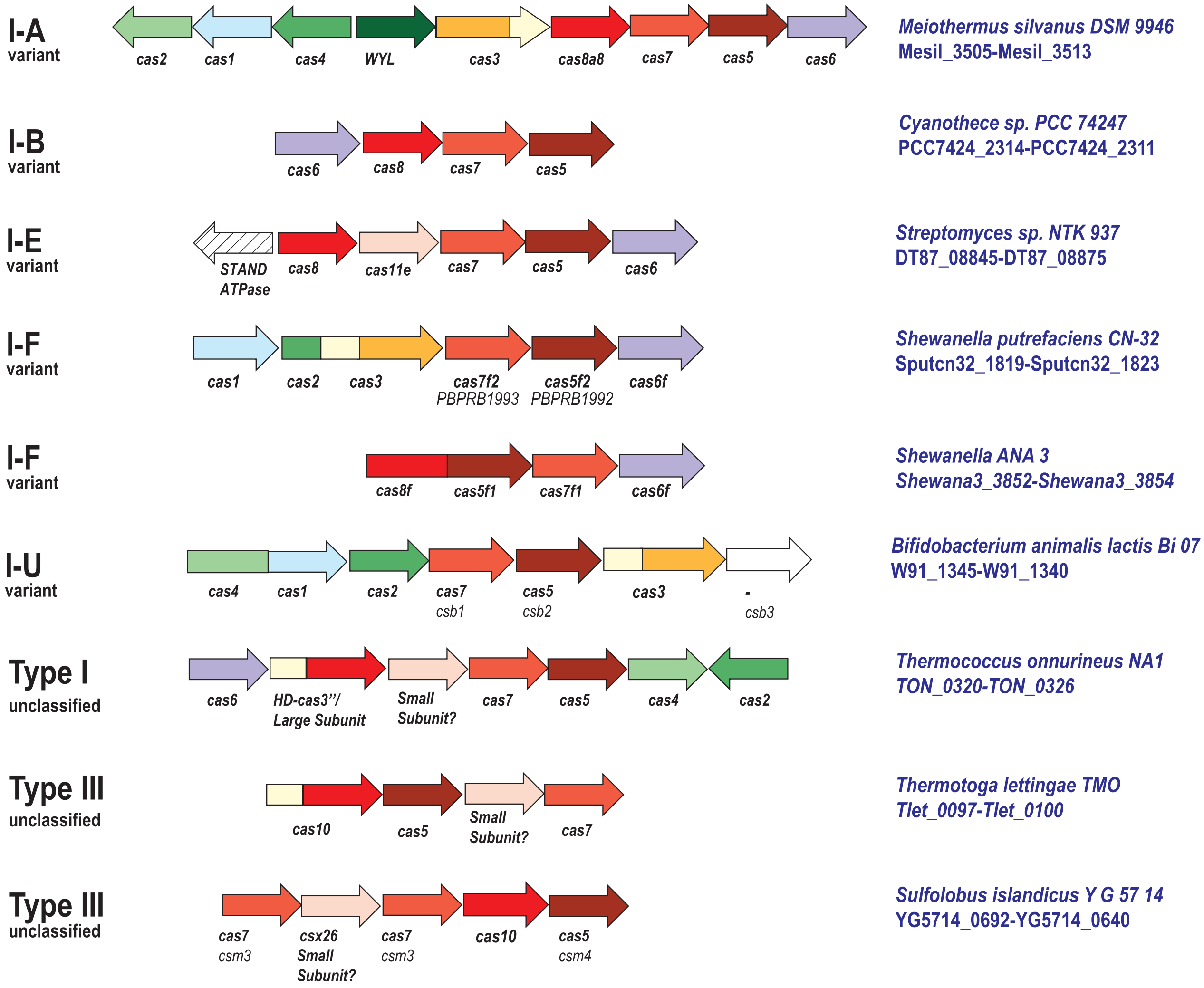


**FIG. 5.** Unusual and derived CRISPR-Cas systems. The depicted unusual and derived CRISPR-Cas systems have the following principal features. (1) Based on the effector genes similarity, this is a subtype I-A variant but with the HD domain (a stand-alone gene in subtype I-A) fused to the C-terminus of Cas3, the effector gene organization typical of subtype I-B. (2) A I-B variant lacking both adaptation genes and *cas3*, carried by a Tn7-like transposon.[42] (3) A I-E variant lacking both the adaptation module and Cas3, and associated with STAND family ATPase.[43] (4) A I-F variant lacking the large subunit and containing atypical, highly diverged Cas5 and Cas7 proteins, a fully functional system[57] that, however, would be considered partial by the current classification schemes. (5) A I-F variant lacking both the adaptation genes and *cas3*, carried by a Tn7-like transposon.[42] (6) A I-U system variant lacking an identifiable large subunit and a *cas6*-like gene but containing the uncharacterized gene *csb3*; also would be considered partial but is widespread in bacteria and likely to be fully functional. (7) A locus identified in *Thermococcus onnurineus* and several other archaea that has been classified as type I based on the general organization of the effector module genes[3]; the HD domain is more similar to that in Cas3 compared to that in Cas10, Cas3 is absent, and Cas7 is most similar to Cas7 protein (Csf2) from type IV systems. (8) A minimal type III system from *Thermotoga* that lacks multiple *cas7*-like genes present in all other type III systems. (9) Distinct type III variant present in several Crenarchaea; csx26, putative small subunit that share no detectable similarity with either *csm2* or *cmr5*, the small subunit genes of subtypes III-A,D, and III-B,C, respectively. The rest of the designations are as in Figure 1.

that might have functionalities and hence applications orthogonal to those of Cas9.[41,43–47] These searches employed dedicated computational pipelines and resulted in the identification of many new varieties of Class 2 systems that neatly fell into two types: V and VI (Fig. 2). The principle of type assignment is simple: each type includes systems with similar domain architectures of the effector protein. Thus, the effectors containing the RuvC-like nuclease domain with an inserted HNH nuclease (Cas9) define type II; effectors in which RuvC is the only recognizable enzymatic domain (Cas12) are the signature of type V; and effectors with two HEPN RNase domains mark type VI (Fig. 2). The criteria for subtype assignment, however, are less obvious and to a considerable extent remain arbitrary. As a general rule of thumb, Class 2 systems are assigned to different subtypes if the respective effectors, despite sharing the domain architecture, do not show significant sequence similarity to each other or at most are ''distantly'' similar. This is, however, a weak and ambiguous criterion and neither has it been applied fully consistently. Sequence similarity criteria are particularly difficult to use for the classification of large, multidomain proteins such as Cas9, Cas12, and Cas13 because even within relatively narrow families, the conserved cores of the catalytic domains are interspersed with long, poorly conserved regions, often containing compositionally biased sequences. Obtaining reliable high-quality alignment is nearly impossible in a fully automated mode, whereas local similarity search methods usually produce a varying number of ''hits'' with widely varying coverage and similarity.

One way to alleviate these problems is to move from comparing individual sequences to profile–profile comparisons (which requires clustering and aligning closely related groups of sequences as the first step) and then cluster sequences using hierarchical similarity dendrograms rather than pairwise similarity thresholds, which is the current standard approach. Both these methodological refinements aim to exploit the increase of the signal-to-noise ratio in aggregated comparisons, first when individual sequences are combined into alignments and then when multiple profile-to-profile comparisons are collectively used to infer the deeper hierarchy.

The use of this approach to classify large multidomain effector genes in Class 2 CRISPR-Cas systems is illustrated in Figure 4B. First, sequences were pre-clustered based on pairwise identity using a conservative similarity threshold and aligned within the clusters. Then, these clusters went through several rounds of a computational procedure, including: (1) the cluster alignments were compared to each other using HHsearch,[48] (2) profile–profile similarity scores were normalized and converted to distances, and (3) unweighted pair group method with arithmetic mean (UPGMA) dendrograms were constructed from the distance matrices. Subtrees below the (user-specified) depth threshold were extracted and used to guide progressive profile–profile alignment with HHalign.[48] After the last round, when the clusters approach known functionally characterized groups and/or reliable full-length alignments become impossible to obtain, the final UPGMA dendrogram (Fig. 4B) is used to assess the relationship between the clusters. In this particular case (type V effector proteins and related TnpB-like proteins), the dendrogram faithfully reproduced the subtype classification such that most of the subtypes radiated into individual lineages above the standard threshold of 1 to 2 (see legend to Fig. 4). The predicted effectors of subtype V-U that are closely related to the transposon-encoded TnpB nucleases and are thought to represent recently evolved, ''baby'' CRISPR-Cas effectors[30] also formed a distinct cluster albeit below the depth threshold. Notably, ''minimal'' variants of Cas12b from two bacterial genomes (*Phycisphaerae bacterium* ST-NAGAB-D1 and *Planctomycetes bacterium* RBG_13_46_10), which are much shorter than the typical Cas12 proteins and can be considered potential intermediates in the effector evolution, display a higher similarity to TnpB in database searches than to other Cas12 proteins. Nevertheless, our procedure confidently clusters them with the rest of the Cas12b. The recently discovered CasX,[49] which has been classified as Cas12e,[4] groups with two new sequences we denote Cas12e var (see the Figure 2 legend for sequence accessions). The CasY effector, discovered in the same work[49] and subsequently classified as Cas12d,[4] confidently groups with Cas12c. Indeed, unification of CasY with Cas12c seems a distinct possibility because in addition to the clustering of the effectors themselves, the respective loci share similar, unusual adaptation modules that lack *cas2* and contain Cas1 proteins that form a clade in the phylogenetic tree of the Cas1 family (Fig. 3). More generally, it should be noted that the comparison of the effector proteins is not the only criterion to assign subtypes. Additional considerations, such as the phylogenetic position of Cas1, the presence of other genes in the *cas* operon(s), fused domains, and more, can and arguably should be taken into account. Judging by the recent pace of discovery of new Cas12 effectors,[4,30] even apart from the different V-U variants, a liberal approach in type V subtype assignment can easily exhaust the Latin alphabet within a few years. A more conservative classification approach similar to that currently applied to Class 1 systems could at least delay such a crisis.

## A Perspective on the Diversity of CRISPR-Cas Systems and Prospects of New Discoveries

The exploration of the CRISPR-Cas system menagerie over the last decade leads to two superficially contradictory conclusions: (1) the known diversity keeps growing, with a slow but steady trickle of new variants being discovered in the expanding sequence databases, but (2) in the overwhelming majority of bacterial and archaeal genomes, most of the CRISPR-*cas* loci belong to already-known types and subtypes.[4,30] This apparent contradiction is resolved by noting that the newly discovered variants are increasingly rare in the genomic and metagenomic databases and have increasingly narrow distributions among the microbial taxa.[30] Thus, the discovery process has already reached far into the tail of the frequency distribution of the CRISPR-Cas systems. This certainly does not rule out the possibility of future discovery of new and potentially technologically promising types and subtypes. However, such discoveries are likely to be feasible only through searching substantially increased volumes of sequence data and might require development of better-performing computational methods for screening such databases. Clearly, this prediction hinges on the incremental model of sequence database expansion. If, for example, major new clades of microbes, especially those with large genomes and occupying complex, highly competitive niches, are discovered, a variety of new CRISPR-Cas systems, perhaps comparable to the currently known diversity, might become available.

In contrast to the decreasing likelihood of discovering new CRISPR-Cas types and subtypes, the number and diversity of the genes that are loosely associated with CRISPR-Cas systems can be expected to grow continuously in the foreseeable future. As suggested by recent efforts on systematic discovery of such genes, CRISPR-Cas molecular machinery has been repeatedly recruited for a variety of cellular processes other than antivirus defense (see examples in Fig. 5).[42,43,50] Each such case reveals new ways Cas proteins interact with their molecular environment and new, specialized variants tuned to the expanded functional repertoire.

## A "Natural" Classification of CRISPR-Cas Systems?

Is a new ''rational'' or ''natural'' CRISPR-Cas classification needed and/or feasible? Obviously, it is preferable for any classification to be based on naturally observed patterns of diversity that reflect evolutionary processes rather than on arbitrarily chosen features. Also, making the classification minimally dependent on expert opinions is highly desirable.

Several features of CRISPR-Cas systems render them recalcitrant to the construction of a fully rational, automated classification. Most fundamentally, the characteristic modular organization, whereby the adaptation and effector modules appear to evolve quasi-independently and are frequently swapped without the loss of functionality,[3,8,51] makes the feasibility of an all-encompassing classification questionable in principle because different parts of the same *cas* locus could and often will be classified differently. This feature either requires recognizing the hybrid status of many loci or making more or less arbitrary decisions on the character precedence. The current system under which CRISPR-Cas systems are classified, primarily by comparing the effector genes, with the adaptation modules and accessory genes assigned secondary roles, illustrates one such choice.

Furthermore, CRISPR-Cas systems are strongly hierarchical in their organization and functional loading of the different parts. The core components of these systems are (nearly) ubiquitous and evolutionarily stable, whereas the wide variety of accessory components are much more fluid.[3,8,43,50,51] Characterization of this hierarchy and the appropriate assignment of diagnostic weights to different parts of CRISPR-Cas systems is likely to remain the domain of expert decision for the foreseeable future. Furthermore, any attempt at automated classification is limited by the rapid decay of sequence similarity with the evolutionary distance for most of the CRISPR-Cas components (Cas1 is a notable exception but, as discussed above, is not particularly useful for classification purposes). The core subunits of the effector complexes in different subtypes are barely recognizable as homologs, even using structural data. Identification and quantification of the similarity between these proteins from sequence alone requires substantial expertise if feasible at all. Until structural data become readily available for a broad diversity of Cas proteins and methods for quantitative structural comparison are adequately refined, classification of CRISPR-Cas systems will require considerable involvement of human experts.

Considering all these issues, we believe it should be accepted that in the foreseeable future, the backbone of the CRISPR-Cas classification (classes and types) will have to remain largely as it is now, that is, based on the expert-assigned hierarchy of features and expert-assisted identification of the system components. In practice, this might not be a severe limitation because, as discussed in the preceding section, it is unlikely that many new types of CRISPR-Cas will be discovered. In contrast, within the types, the similarity among the subtypes can be sufficient to make the classification amenable to more rigorous approaches. Developing a consistent methodology for CRISPR-*cas* loci comparison that is based on both sequence similarity between the components and the gene content, with subsequent clustering, appears to be the most realistic direction.

## The Nomenclatural Morass

If classification of CRISPR-Cas systems is difficult, nomenclature of the *cas* genes is an even more damning problem, as already touched upon above, with regard to the classification of type I large subunits (Fig. 3A). So far, only 13 groups of CRISPR-associated genes that comprise the cores of the adaptation and effector modules but represent a small minority of the entire diversity of CRISPR-linked genes have been assigned systematic *cas* names. Even for these genes, ''legacy'' names, such as *csm*, *cmr*, *cse*, *csy*, or *csa*, continue to be used frequently, making it impossible to infer from the name to which group the given protein belongs.[3,10] Otherwise, though, the approach to the naming of CRISPR-associated genes so far has been quite conservative, such that even some genes that are common in certain types or subtypes and essential for their function have not been assigned *cas* names. A notable example is type III genes encoding proteins consisting of a CARF and HEPN domains, such as *csm6* and *csx1*. Thus, for the great majority of the CRISPR-associated genes, legacy names are all that exists, but even these are used inconsistently, sometimes erroneously, and in particular do not reflect the already-known diversity of the CARF-domain containing genes.[43,52]

The recent systematic efforts on prediction of genes functionally linked to CRISPR-Cas have yielded dozens of new, strong candidates.[43,50] Obviously, for these genes, no CRISPR-related names exist (yet), and often they have no names at all other than the systematic designations specific to the corresponding genomes and the IDs of the respective sequence profiles. So, a major question is whether we need to extend the systematic nomenclature to all genes for which there is evidence of functionally relevant association with CRISPR-Cas, or perhaps only to genes that are found in a reasonably broad range of CRISPR-Cas systems—if so, what are the criteria?—and/or have been linked to CRISPR-Cas functions experimentally. Making no effort on systematic nomenclature at all seems like a losing proposition because even at this stage, published descriptions of CRISPR-associated genes are far from being uniform or transparent, and the situation will only get worse with the further growth of the collection of CRISPR-*cas* loci.

## Concluding Remarks

Classification of defense systems that are typically highly diverse and evolve rapidly in the course of the arms race with parasites is an inherently difficult task as well illustrated, for example, by the long-term efforts on classification of restriction-modification (RM) enzymes.[53–55] The CRISPR-Cas systems do not reach the same level of diversity as RM modules, but the organization of CRISPR-Cas is more complex, which arguably makes the problem no easier. A consistent, ''natural'' classification does not appear feasible, given the complexity of the evolutionary relationships between CRISPR-Cas variants that include extensive module shuffling and are further confounded by the rapid sequence and structural divergence of homologous Cas proteins.[5] Nevertheless, simple and reliable criteria for the delineation of CRISPR-Cas classes and types are available, and moreover it appears unlikely that many new types will be discovered, so that, at the top levels, the current classification could be (nearly) complete. At the level of subtypes, the situation is far more complicated. The current criteria are not fully consistent, and there seems to be room for developing more advanced and rigorous methodologies. Such developments are particularly pertinent because it can be expected that a considerable number of new subtypes will be discovered, at least in Class 2. Furthermore, the present, relatively small number of Class 1 subtypes could be but an illusion caused by the inadequacy of the current classification approaches. A distinct problem is the classification of derived forms of CRISPR-Cas systems and their exaptation for non-defense functions or at least functions distinct from adaptive immunity. Whether such systems even qualify as CRISPR-Cas depends on the definition that perhaps should be expanded to include all *cas* operons. If CRISPR-Cas systems are redefined in this broader manner, it remains to be determined whether the derived variants stay within the respective ancestral subtypes or are given a separate status.

Although (nearly) all the CRISPR-Cas types might already be known and new subtypes are being discovered at a moderate pace, the list of CRISPR-linked genes is growing much faster, so that their classification and especially nomenclature present separate challenges. All in all, although much development, analysis, and expert effort are required, we believe that given the extensive but limited diversity of the CRISPR-Cas systems, the current classification and nomenclature can be substantially improved, resulting in a stable and consistent systematics. This seems to be a realistic task that can be completed within a few years.

## Acknowledgments

The authors would like to thank all members of the CRISPR community who have contributed to the efforts on classification and naming of the CRISPR-Cas systems and *cas* genes over the years. The authors' research is supported by the intramural funds of the U.S. National Institutes of Health.

## Author Disclosure Statement

No competing financial interests exist.

## References

1. Stern A, Sorek R. The phage-host arms race: shaping the evolution of microbes. *Bioessays* 2011;33:43–51. DOI: 10.1002/bies.201000071.
2. Koonin EV, Makarova KS, Wolf YI. Evolutionary genomics of defense systems in archaea and bacteria. *Annu Rev Microbiol* 2017;71:233261. DOI: 10.1146/annurev-micro-090816-093830.
3. Makarova KS, Wolf YI, Alkhnbashi OS, et al. An updated evolutionary classification of CRISPR-Cas systems. *Nat Rev Microbiol* 2015;13:722–736. DOI: 10.1038/nrmicro3569.
4. Koonin EV, Makarova KS, Zhang F. Diversity, classification and evolution of CRISPR-Cas systems. *Curr Opin Microbiol* 2017;37:67–78. DOI: 10.1016/j.mib.2017.05.008.
5. Takeuchi N, Wolf YI, Makarova KS, et al. Nature and intensity of selection pressure on CRISPR-associated genes. *J Bacteriol* 2012;194:1216–1225. DOI: 10.1128/JB.06521-11.
6. Kunin V, Sorek R, Hugenholtz P. Evolutionary conservation of sequence and secondary structures in CRISPR repeats. *Genome Biol* 2007;8:R61. DOI: 10.1186/gb-2007-8-4-r61.
7. Lange SJ, Alkhnbashi OS, Rose D, et al. CRISPRmap: an automated classification of repeat conservation in prokaryotic adaptive immune systems. *Nucleic Acids Res* 2013;41:8034–8044. DOI: 10.1093/nar/gkt606.
8. Silas S, Makarova KS, Shmakov S, et al. On the origin of reverse transcriptase-using CRISPR-Cas systems and their hyperdiverse, enigmatic spacer repertoires. *MBio* 2017;8. DOI: 10.1128/mBio.00897-17.
9. Makarova KS, Aravind L, Wolf YI, et al. Unification of Cas protein families and a simple scenario for the origin and evolution of CRISPR-Cas systems. *Biol Direct* 2011;6:38. DOI: 10.1186/1745-6150-6-38.
10. Makarova KS, Koonin EV. Annotation and Classification of CRISPR-Cas Systems. *Methods Mol Biol* 2015;1311:47–75. DOI: 10.1007/978-1-4939-2687-9_4.
11. Zhao H, Sheng G, Wang J, et al. Crystal structure of the RNA-guided immune surveillance Cascade complex in *Escherichia coli*. *Nature* 2014;515:147–150. DOI: 10.1038/nature13733.
12. van der Oost J, Westra ER, Jackson RN, et al. Unravelling the structural and mechanistic basis of CRISPR-Cas systems. *Nat Rev Microbiol* 2014;12:479–492. DOI: 10.1038/nrmicro3279.
13. Jackson RN, Golden SM, van Erp PB, et al. Structural biology. Crystal structure of the CRISPR RNA-guided surveillance complex from *Escherichia coli*. *Science* 2014;345:1473–1479. DOI: 10.1126/science.1256328.
14. Jackson RN, Wiedenheft B. A conserved structural chassis for mounting versatile CRISPR RNA-guided immune responses. *Mol Cell* 2015;58:722–728. DOI: 10.1016/j.molcel.2015.05.023.
15. Hochstrasser ML, Taylor DW, Bhat P, et al. CasA mediates Cas3-catalyzed target degradation during CRISPR RNA-guided interference. *Proc Natl Acad Sci U S A* 2014;111:6618–6623. DOI: 10.1073/pnas.1405079111.
16. Hochstrasser ML, Taylor DW, Kornfeld JE, et al. DNA targeting by a minimal CRISPR RNA-guided cascade. *Mol Cell* 2016;63:840–851. DOI: 10.1016/j.molcel.2016.07.027.
17. Staals RH, Agari Y, Maki-Yonekura S, et al. Structure and activity of the RNA-targeting type III-B CRISPR-Cas complex of *Thermus thermophilus*. *Mol Cell* 2013;52:135–145. DOI: 10.1016/j.molcel.2013.09.013.
18. Staals RH, Zhu Y, Taylor DW, et al. RNA targeting by the type III-A CRISPR-Cas Csm complex of *Thermus thermophilus*. *Mol Cell* 2014;56:518–530. DOI: 10.1016/j.molcel.2014.10.005.
19. Charpentier E, Richter H, van der Oost J, et al. Biogenesis pathways of RNA guides in archaeal and bacterial CRISPR-Cas adaptive immunity. *FEMS Microbiol Rev* 2015;39:428–441. DOI: 10.1093/femsre/fuv023.
20. Niewoehner O, Jinek M. Structural basis for the endoribonuclease activity of the type III-A CRISPR-associated protein Csm6. *RNA* 2016;22:318–329. DOI: 10.1261/rna.054098.115.
21. Beloglazova N, Petit P, Flick R, et al. Structure and activity of the Cas3 HD nuclease MJ0384, an effector enzyme of the CRISPR interference. *EMBO J* 2011;30:4616–4627. DOI: 10.1038/emboj.2011.377.
22. Sinkunas T, Gasiunas G, Fremaux C, et al. Cas3 is a single-stranded DNA nuclease and ATP-dependent helicase in the CRISPR/Cas immune system. *EMBO J* 2011;30:1335–1342. DOI: 10.1038/emboj.2011.41.
23. Westra ER, van Erp PB, Künne T, et al. CRISPR immunity relies on the consecutive binding and degradation of negatively supercoiled invader DNA by Cascade and Cas3. *Mol Cell* 2012;46:595–605. DOI: 10.1016/j.molcel.2012.03.018.
24. Gong B, Shin M, Sun J, et al. Molecular insights into DNA interference by CRISPR-associated nuclease-helicase Cas3. *Proc Natl Acad Sci U S A* 2014;111:16359–16364. DOI: 10.1073/pnas.1410806111.
25. Huo Y, Nam KH, Ding F, et al. Structures of CRISPR Cas3 offer mechanistic insights into Cascade-activated DNA unwinding and degradation. *Nat Struct Mol Biol* 2014;21:771–777. DOI: 10.1038/nsmb.2875.
26. Xiao Y, Luo M, Dolan AE, et al. Structure basis for RNA-guided DNA degradation by Cascade and Cas3. *Science* 2018;361. DOI: 10.1126/science.aat0839.
27. Jung TY, An Y, Park KH, et al. Crystal structure of the Csm1 subunit of the Csm complex and its single-stranded DNA-specific nuclease activity. *Structure* 2015;23:782–790. DOI: 10.1016/j.str.2015.01.021.
28. Zhang J, Graham S, Tello A, et al. Multiple nucleic acid cleavage modes in divergent type III CRISPR systems. *Nucleic Acids Res* 2016;44:1789–1799. DOI: 10.1093/nar/gkw020.
29. Liu TY, Iavarone AT, Doudna JA. RNA and DNA targeting by a reconstituted *Thermus thermophilus* type III-A CRISPR-Cas system. *PLoS One* 2017;12:e0170552. DOI: 10.1371/journal.pone.0170552.
30. Shmakov S, Smargon A, Scott D, et al. Diversity and evolution of class 2 CRISPR-Cas systems. *Nat Rev Microbiol* 2017;15:169–182. DOI: 10.1038/nrmicro.2016.184.
31. Makarova KS, Grishin NV, Shabalina SA, et al. 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. DOI: 10.1186/1745-6150-1-7.
32. Chylinski K, Makarova KS, Charpentier E, et al. Classification and evolution of type II CRISPR-Cas systems. *Nucleic Acids Res* 2014;42:6091–6105. DOI: 10.1093/nar/gku241.
33. Hudaiberdiev S, Shmakov S, Wolf YI, et al. Phylogenomics of Cas4 family nucleases. *BMC Evol Biol* 2017;17:232. DOI: 10.1186/s12862-017-1081-1.
34. Vestergaard G, Garrett RA, Shah SA. CRISPR adaptive immune systems of Archaea. *RNA Biol* 2014;11:156–167. DOI: 10.4161/rna.27990.
35. Krupovic M, Beguin P, Koonin EV. Casposons: mobile genetic elements that gave rise to the CRISPR-Cas adaptation machinery. *Curr Opin Microbiol* 2017;38:36–43. DOI: 10.1016/j.mib.2017.04.004.
36. Krupovic M, Makarova KS, Forterre P, et al. Casposons: a new superfamily of self-synthesizing DNA transposons at the origin of prokaryotic CRISPR-Cas immunity. *BMC Biology* 2014;12:36. DOI: 10.1186/1741-7007-12-36.
37. Shah SA, Erdmann S, Mojica FJ, et al. Protospacer recognition motifs: mixed identities and functional diversity. *RNA Biol* 2013;10:891–899. DOI: 10.4161/rna.23764.
38. Leenay RT, Maksimchuk KR, Slotkowski RA, et al. Identifying and visualizing functional PAM diversity across CRISPR-Cas systems. *Mol Cell* 2016;62:137–147. DOI: 10.1016/j.molcel.2016.02.031.
39. Pyenson NC, Gayvert K, Varble A, et al. Broad targeting specificity during bacterial type III CRISPR-Cas immunity constrains viral escape. *Cell Host Microbe* 2017;22:343–353.e3. DOI: 10.1016/j.chom.2017.07.016.
40. Pyenson NC, Marraffini LA. Type III CRISPR-Cas systems: when DNA cleavage just isn't enough. *Curr Opin Microbiol* 2017;37:150–154. DOI: 10.1016/j.mib.2017.08.003..
41. Shmakov S, Abudayyeh OO, Makarova KS, *et al.* Discovery and functional characterization of diverse class 2 CRISPR-Cas systems. *Mol Cell* 2015;60:385–397. DOI: 10.1016/j.molcel.2015.10.008.
42. Peters JE, Makarova KS, Shmakov S, et al. Recruitment of CRISPR-Cas systems by Tn7-like transposons. *Proc Natl Acad Sci U S A* 2017;114:E7358–E7366. DOI: 10.1073/pnas.1709035114.
43. Shmakov SA, Makarova KS, Wolf YI, et al. Systematic prediction of genes functionally linked to CRISPR-Cas systems by gene neighborhood analysis. *Proc Natl Acad Sci U S A* 2018;115:E5307–E5316. DOI: 10.1073/pnas.1803440115.

44. Abudayyeh OO, Gootenberg JS, Konermann S, et al. C2c2 is a single-component programmable RNA-guided RNA-targeting CRISPR effector. *Science* 2016;353:aaf5573. DOI: 10.1126/science.aaf5573.
45. Konermann S, Lotfy P, Brideau NJ, et al. Transcriptome engineering with RNA-targeting type VI-D CRISPR effectors. *Cell* 2018;173:665–676. DOI: 10.1016/j.cell.2018.02.033.
46. Smargon AA, Cox DBT, Pyzocha NK, et al. Cas13b is a type VI-B CRISPR-associated RNA-guided RNase differentially regulated by accessory proteins Csx27 and Csx28. *Mol Cell* 2017;65:618–630. DOI: 10.1016/j.molcel.2016.12.023.
47. Yan WX, et al. Cas13d is a compact RNA-targeting type VI CRISPR effector positively modulated by a WYL-domain-containing accessory protein. *Mol Cell* 2018;70:327–339. DOI: 10.1016/j.molcel.2018.02.028.
48. Soding J. Protein homology detection by HMM-HMM comparison. *Bioinformatics* 2005;21:951–960. DOI: 10.1093/bioinformatics/bti125.
49. Burstein D, Harrington LB, Strutt SC, et al. New CRISPR-Cas systems from uncultivated microbes. *Nature* 2017;542:237–241. DOI: 10.1038/nature21059.
50. Shah SA, Alkhnbashi OS, Behler J, et al. Comprehensive search for accessory proteins encoded with archaeal and bacterial type III CRISPR-Cas gene cassettes reveals 39 new cas gene families. *RNA Biol* 2018;1–13. DOI:10.1080/15476286.2018.1483685.
51. Makarova KS, Wolf YI, Koonin EV. The basic building blocks and evolution of CRISPR-Cas systems. *Biochem Soc Trans* 2013;41:1392–1400. DOI: 10.1042/BST20130038.
52. Makarova KS, Anantharaman V, Grishin NV, et al. CARF and WYL domains: ligand-binding regulators of prokaryotic defense systems. *Front Genet* 2014;5:102. DOI: 10.3389/fgene.2014.00102.
53. Roberts RJ, Belfort M, Bestor T, et al. A nomenclature for restriction enzymes, DNA methyltransferases, homing endonucleases and their genes. *Nucleic Acids Res* 2003;31:1805–1812.
54. Roberts RJ, Vincze T, Posfai J, et al. REBASE—enzymes and genes for DNA restriction and modification. *Nucleic Acids Res* 2007;35:D269–270. DOI: 10.1093/nar/gkl891.
55. Pingoud A, Wilson GG, Wende W. Type II restriction endonucleases—a historical perspective and more. *Nucleic Acids Res* 2014;42:7489–7527. DOI: 10.1093/nar/gku447.
56. Price MN, Dehal PS, Arkin AP. FastTree 2—approximately maximum-likelihood trees for large alignments. *PLoS One* 2010;5:e9490. DOI: 10.1371/journal.pone.0009490.
57. Pausch P, et al. Structural variation of type I-F CRISPR RNA guided DNA surveillance. *Mol Cell* 2017;67:622–632.e4. DOI: 10.1016/j.molcel.2017.06.036.