Evolution and classification of the CRISPR/Cas systems

Kira S. Makarova[1], Daniel H. Haft[2], Rodolphe Barrangou[3], Stan J. J. Brouns[4], Emmanuelle Charpentier[5], Philippe Horvath[6], Sylvain Moineau[7], Francisco J. M. Mojica[8], Yuri I. Wolf[1], Alexander F. Yakunin[9], John van der Oost[4], and Eugene V. Koonin[1,*]

[1]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, 8600 Rockville Pike, Bethesda, MD 20894

[2]The J Craig Venter Institute, 9704 Medical Center Drive, Rockville, MD 20850

[3]Danisco USA Inc., 3329 Agriculture Drive, Madison, WI 53716

[4]Laboratory of Microbiology, Wageningen University, Dreijenplein 10, 6703 HB Wageningen, The Netherlands

[5]Laboratory for Molecular Infection Medicine Sweden, Umeå Centre for Microbial Research, Department of Molecular Biology, Umeå University, S-90187 Umeå, Sweden

[6]Danisco France SAS, BP10, 86220 Dangé-Saint-Romain, France

[7]Département de Biochimie, Microbiologie et Bio-informatique, Faculté des Sciences et de Génie, Groupe de Recherche en Ecologie Buccale, Université Laval, Quebec City, Quebec, G1V 0A6, Canada

[8]Departamento de Fisiología, Genética y Microbiología, Universidad de Alicante, 03080-Alicante, Spain

[9]Banting and Best Department of Medical Research, University of Toronto, Toronto, Ontario, Canada

*for correspondence: koonin@ncbi.nlm.nih.gov

Abstract

The CRISPR/Cas modules are adaptive immunity systems that are widespread in Archaea and Bacteria. These defense systems show extraordinary diversity in operon architecture along with high rates of *cas* gene evolution. Several classifications of CRISPR/Cas systems and Cas proteins have been proposed but it appears difficult to capture the full complexity of these systems in a coherent scheme. Here, we provide an updated analysis of the evolutionary relationships between CRISPR/Cas systems and Cas proteins. Three major types of CRISPR/Cas systems are delineated, with a further division into several subtypes and a few chimeric variants. Given the complexity of genomic architectures and the extremely dynamic evolution of the CRISPR/Cas systems, a unified classification of these systems should be based on multiple criteria. Accordingly, a "polythetic" classification is proposed that integrates the phylogenies of the most common *cas* genes, the sequence and arrangements of the CRISPR repeats, and the architecture of the CRISPR/Cas loci.

The CRISPR/Cas (clustered regularly interspaced short palindromic repeats / CRISPR-associated proteins) modules are adaptive immunity systems that act against invading genetic elements and are encoded by most Archaea and many Bacteria [1-6] (Supplementary Table 1). Distinct arrays of short repeats interspersed with unique spacers (CRISPR) have been recognized in bacterial and archaeal genomes for years, and although it has been proposed that these repeat arrays could have an important common function [7], the nature of that function has been elucidated only much later. Independently, Cas protein sequences encoded by putative operons adjacent to CRISPR were analyzed in detail with computational methods and found to contain domains characteristic of several nucleases, a helicase, a polymerase and RNA-binding proteins[8]. Initially, it was speculated that these proteins might jointly constitute a novel DNA repair system [9]. However, the observation that some of the unique CRISPR spacers were (nearly) identical to fragments of virus and plasmid genes led to the hypothesis that CRISPR/Cas might be involved in defense against selfish elements [10-12]. On the basis of these findings combined with a comprehensive computational re-analysis of the Cas proteins [13,14], a model was proposed [14] that drew an analogy between the putative novel prokaryotic defense system and the eukaryotic RNA interference (RNAi) mechanisms[15]. However, unlike the eukaryotic RNAi systems, the CRISPR/Cas system integrates a small piece of DNA derived from foreign nucleic acid into the CRISPR locus of the host genome as the first step in the series of events that leads to immunity to the given agent [14]. The hypothesis that the CRISPR/Cas system plays a role in defense against invading DNA has been validated by the demonstration that integration of a short bacteriophage-specific sequence into the CRISPR locus of the lactic acid bacterium *Streptococcus*

thermophilus conferred resistance to the cognate phage[16]. The phage resistance was abrogated by even a single mismatch between the CRISPR insert (referred to as spacer) and the target phage sequence (referred to as proto-spacer) [16].

The CRISPR/Cas systems mediate immunity to invading genetic elements via three stages (Figure 1)[1-3,6]. The first stage is adaptation which involves the integration of short pieces of DNA homologous to virus or plasmid sequences into the CRISPR loci[16-18]. Viral challenge typically triggers insertion of a single virus-derived, resistance-conferring spacer with a characteristic length of approximately 30 bp; acquisition of multiple spacers from the same phage is less frequent. The insertion of new spacers has been reported to depend on short (few nucleotides) proto-spacer adjacent motifs (PAMs) which differ between variants of the CRISPR/Cas system and appear to determine the selection of the inserted spacer [19,20].

The second stage is the expression and processing, during which the long primary transcript of a CRISPR locus (pre-crRNA) is processed into short crRNAs . The latter processing step is catalyzed by endoribonucleases that either operate as a subunit of a larger complex (e.g. Cascade, CRISPR-associated complex for antiviral defense in *Escherichia coli*) (Figure 1) or as a stand alone enzyme (e.g., Cas6 in the archaeon *Pyrococcus furiosus*). In the case of the Cascade complex[21], the mature crRNA remains associated with the complex after initial endonuclease cleavage (Figure 1), whereas in *P. furiosus* the crRNA, processed by Cas6, is passed on to a distinct Cas protein complex (Cmr-type, see below) where it is processed further [22,23].

94 The third and final step is interference, during which the alien nucleic acid
95 (foreign DNA or RNA) is targeted and cleaved within the proto-spacer sequence[6,17,18].
96 The crRNAs guide the respective complexes of Cas proteins to the complementary target
97 sequences of invading viruses or plasmids that match the spacer. In *S. thermophilus* and
98 *E. coli* targeting either strand of the phage DNA confers immunity to the cognate phage,
99 an observation that is best compatible with DNA being the target[16,24]. Furthermore,
100 insertion of a self-splicing intron into the proto-spacer sequence of the target gene
101 rendered the respective plasmid resistant to the CRISPR-mediated immunity in
102 *Staphylococcus epidermidis*, indicating that the invading DNA rather than mRNA is
103 targeted [25]. However, *in vitro* experiments with the CRISPR/Cas system from *P. furiosus*
104 showed that in this case the crRNA rather targets the viral mRNA [23]. These findings
105 emphasize the remarkable mechanistic and functional diversity of the CRISPR/Cas
106 systems, although the full range of their activities remains to be determined. Various Cas
107 proteins might participate in either one or multiple stages of the CRISPR/Cas system
108 action, most likely, as protein complexes [6].
109 In agreement with the bioinformatic predictions, nuclease activities, RNAse
110 and/or DNAse, have been demonstrated for several Cas proteins, including the two
111 universal core Cas proteins: Cas1 (a metal-dependent DNAse with no sequence
112 specificity that has been proposed to be involved in the integration of the alien DNA
113 (spacer) into the CRISPR cassettes [26]) and Cas2 (a metal-dependent endoribonuclease
114 whose role in the CRISPR/Cas mechanism remains unclear [27]). The Cas proteins known
115 as RAMPs (Repeat-Associated Mysterious Proteins) contain a double ferredoxin-fold

domain; some of the RAMPs have been shown to possess sequence- or structure-specific RNAse activity that is involved in the processing of pre-crRNA transcripts [21,22,24].

The CRISPR/Cas systems can be divided into two distinct, quasi-independent subsystems. The highly conserved "information processing" subsystem includes the Cas1 and Cas2 proteins in its core, and is thought to be involved in the maintenance and replenishment of the spacer-repeat library. The "executive" subsystem, which is highly variable in content, typically includes multiple RAMP proteins, and is involved in processing of the CRISPR transcript (Cascade complex and its analogs) and the crRNA-directed interference of invading genetic elements.

Extensive bioinformatic analyses have shown that the genomes of various CRISPR-containing organisms encode approximately 65 distinct Cas proteins which can be classified into 23 to 45 families depending on the classification criteria (granularity of clustering)[13,14]. Furthermore, 8 distinct subtypes of the CRISPR/Cas systems (CASS1 to CASS8) have been delineated on the basis of the composition and architecture of the *cas* operons and Cas1 phylogeny [13,14].

The diversity of CRISPR systems identified in newly sequenced genomes (in a representative set of 703 archaeal and bacterial genomes, 310 (44%) encode one or more CRISPR/Cas modules; Table 2 and Supplementary Table 1) is rapidly increasing [1,4], hence an urgent need exists for a rational and unified classification and nomenclature of the *cas* genes. In this article, we summarize the shortcomings of the existing classifications and nomenclatures of the CRISPR/Cas systems, and propose a new,

139 "polythetic" classification which combines information from comparative-genomic and
140 phylogenetic analyses.
141
142 **Problesm with the existing CRISPR/Cas classification**
143 The original, widely used classification proposed by Haft et al. was based on the
144 topology of the Cas1 phylogenetic tree and *cas* operon organization in eight organisms [13].
145 The names of four core *cas* genes were kept as they were originally proposed by Jansen
146 et al. in 2002 [8]. Two other core genes, *cas5* and *cas6*, were then added using the same
147 principle [13]. In addition, gene names for proteins specific to each of the eight CRISPR
148 systems were proposed. For example, the unique genes found in the *E. coli* system were
149 denoted *cse1* (CRISPR system of *E. coli* gene number 1), *cse2*, *cse3*, *cse4*, and *cse5*
150 (elsewhere, these *E. coli* genes were also labeled as *casA, casB, casC, casD, and casE*)[21].
151 Although the original approach [13] offers attractive simplicity, its major
152 shortcoming is the failure to identify distant relationships between many proteins. For
153 example, the proteins that belong to the COG1857 family, that are present in the majority
154 of the CRISPR/Cas systems and are obviously orthologous, have been given at least 5
155 different names: Cse3, Csd2, Csh2, Cst2, and Csa2 (Table 1). The other shortcoming of
156 this classification is that it does not take into account the complexity of the relationships
157 between the CRISPR/Cas systems and the respective Bacteria and Archaea. In particular,
158 the classification ignores the apparent relatedness between several CRISPR/Cas systems:
159 for example, the Ecoli and Ypest systems are definitely related, and so are the Apern,
160 Tneap/Hmari and Dvulg systems which share a common signature gene of the BH0338
161 family. Conversely, extensive recombination within CRISPR operons has resulted in

hybrid CRISPR/Cas systems that cannot be assigned to any of the proposed groups although they contain typical *cas* genes. The linkage between CRISPR/Cas groups and particular organisms can be misleading due to the presence of multiple CRISPR/Cas systems in the same organism, the presence of different CRISPR/Cas systems in different strains of same species, and again, the rather wide spread of hybrid systems. The inconsistencies between the nomenclatures of the CRISPR/Cas systems and the names of Cas proteins are rapidly growing. In particular, many of these proteins are currently classified into families that do not have systematic names pointing to their involvement with CRISPR/Cas (e.g., CXXC_CXXC family protein, GSU0053 family protein, etc.).

Evolutionary relationships as a basis for a new classification of CRISPR/Cas systems

Two Cas proteins (Cas1 and Cas2) are present in all CRISPR/Cas systems that are predicted to be functionally active and are thought to be involved in spacer integration (the adaptation stage) as the "information processing" subsystem. The *cas1* and *cas2* genes comprise the cores of three distinct types (I, II and III) of CRISPR/Cas systems that form the basis of a new, polythetic (based on multiple criteria) classification we propose here (Figure 2 and Table 1).

Type I CRISPR/Cas system. Typical Type I loci contain a gene for a predicted helicase/nuclease (Cas3) as well as several other proteins that probably form Cascade-like complexes with different compositions [21,24]. These complexes include multiple proteins of the RAMP superfamily, in particular, the widespread COG1857 (Cas7) family

185 recently proposed to adopt the RAMP fold (KSM, unpublished observations), and
186 BH0338-like families; in addition, the complexes may contain other, less conserved
187 subunits. In the Cascade complex, a RAMP protein has been demonstrated to be the
188 major enzyme (RNA endonuclease) that catalyzes the processing of the long
189 spacer/repeat-containing transcript into a mature crRNA[21,24]. In most cases, the catalytic
190 RAMP protein (Cas6, Cas6e and Cas6f; see Table 1) does not belong to the most
191 prevalent Cas5 or Cas7 family of RAMPs and is often encoded in the periphery of the
192 respective operon. However, an exception might be the subtype I-C system (also known
193 as Dvulg/CASS1, Table 1 and Figure 2) in which either Cas5 or Cas7 are likely
194 candidates to possess RNAse activity. The Type I CRISPR systems appear to target
195 DNA, and the cleavage might be catalyzed by the HD nuclease domains of Cas3 and/or
196 by the RecB-family nucleases (Cas4). However, the fact that in several Type I
197 CRISPR/Cas systems the RecB nuclease domain is fused to Cas1, may suggest a role for
198 Cas4 in spacer acquisition.
199
200 ***Type II CRISPR/Cas system.*** The Type II systems include the "HNH" type
201 system (*Streptococcus*-like, also known as Nmeni subtype or CASS4, Table 3) in which
202 (in addition to the ubiquitous Cas1 and Cas2) a single, very large protein, Cas9
203 (COG3513), appears to be sufficient for both generating crRNA and cleaving the target
204 DNA. The Cas9 protein (~1,000 amino acids) contains at least two nuclease domains,
205 namely, the N-terminal RuvC-like nuclease (RNAse H fold) and the HNH (McrA-like)
206 nuclease domain that is located in the middle of the protein. The functional specialization
207 of these nuclease domains remains to be elucidated. However, the HNH nuclease domain

208 is abundant in restriction enzymes and possesses endonuclease activity[28,29], so it is likely
209 to be responsible for target cleavage.
210
211 ***Type III CRISPR/Cas system.*** The Type III systems contain polymerase-RAMP
212 modules in which at least some of the RAMPs appear to be involved in the processing of
213 the CRISPR-spacer transcripts analogously to the Cascade complex. Targeting of plasmid
214 DNA by this system (subtype III-B) has been demonstrated *in vivo* in *S. epidermidis* [25],
215 and it seems plausible that the HD domain of the polymerase-like protein (COG1353) is
216 involved in the target DNA cleavage. There is also strong evidence that at least *in vitro*
217 the Type III-A CRISPR/Cas system from *P. furiosus* can target RNA [23]. Apart from the
218 universal Cas2 protein, the only identified ribonucleases in the Type III CRISPR/Cas
219 systems are RAMP proteins. Beside Cas6 that is involved in CRISPR transcript
220 processing, Type III systems contain at least two additional RAMPs. These RAMPs
221 appear to be the most likely candidate enzymes for the subsequent trimming of the
222 crRNA. In many organisms, Type III CRISPR/Cas operons lack the *cas1-cas*2 gene pair;
223 in all these cases, an additional CRISPR locus (Type I or II) is present in the respective
224 genome, so the polymerase-RAMP module probably interacts with Cas1-Cas2 *in trans*.
225 In other organisms, the polymerase-RAMP modules are present in a single operon with
226 *cas1* and *cas2*, e.g., a module with the typical architecture in *Staphylococcus epidermidis*
227 and *Mycobacterium tuberculosis* (Type III-A) and a distinct version in *Halorhodospira*
228 *halophila* (Type III-B). In these organisms the Type III operon is the only CRISPR/Cas
229 locus, suggesting that, combined with Cas1-Cas2, the polymerase-RAMP module forms a
230 fully functional, autonomous Type III system.

The three types of CRISPR systems show a distinctly non-uniform distribution among the major lineages of Archaea and Bacteria (Table 2). In particular, the Type II systems so far have been found exclusively in bacteria whereas Type III systems are more common in Archaea. The previously observed trend of overrepresentation of CRISPR in Archaea compared to Bacteria[30] still holds (Table 2). Moreover, the majority of archaeal genomes carry more than one CRISPR/Cas system; typically, different modules within the same genome are unrelated.

Subtypes of the CRISPR/Cas systems and their evolution

Based on the gene composition and architecture of the respective *cas* operons, the three basic types of CRISPR/Cas systems can be further classified into subtypes that largely agree with the previously delineated variants [13, 14]. Each of the subtypes contains a distinct signature gene (Figure 2 and Table 1). The ubiquitous, highly conserved Cas1 protein can be used as a scaffold to investigate the evolution of the CRISPR/Cas system (the other universal protein, Cas2, is too small to yield a well-resolved tree). The phylogenetic tree of Cas1 includes several well-resolved branches that generally agree with the classification of CRISPR/Cas systems into subtypes (I-A, I-B, I-C, I-E, I-F, II and III-A) [14], with a few notable exceptions (Figure 3). In particular, Cas1 proteins associated with the polymerase-RAMP module (type III) appear in several unrelated positions in the tree (Figure 3), suggesting that this module can operate with a variety of *cas1-cas2* genes both *in cis* and *in trans*.

The CRISPR repeats can be classified into at least 12 groups based on sequence similarity [31]. Four groups of CRISPR repeats clearly correspond to distinct CRISPR/Cas subtypes (all of Type I): group 2 – I-E (Ecoli/CASS2), group 3 – I-C (Dvulg/CASS1), group 4 – I-F (Ypest/CASS3) and group 10 – type II (Nmeni/CASS4). These four variants of CRISPR/Cas systems have the most stable operon organizations; on the other hand, subtypes I-A, I-B, I-D, and Type III appear to be prone to recombination between different subtypes. Structural characteristics of CRISPR repeats of these four groups can be potentially employed for classification, in addition to phylogenetic data and signature genes. Other groups of repeats cannot be unequivocally associated with particular CRISPR/Cas system subtypes.

Integration of all the above considerations in a dendrogram, reflects our present understanding of the evolutionary history of CRISPR/Cas systems (Figure 2). Subtypes of the Type I system are grouped according to their operon organizations and the phylogeny of their Cas1 proteins.

Proposals for the CRISPR/Cas system nomenclature

Most of the CRISPR/Cas loci can be readily classified into the proposed types and subtypes based on the presence of type- and subtype-specific signature genes (Table 1 and Table 3). In addition, we introduce the catch-all subtypes I-U, II-U and III-U (U for unclassified) for systems that lack currently defined subtype-specific signature genes but might fit one of the established subtypes based on further structure and sequence analysis or potentially could become founders of new subtypes. In the same vein, we propose

Type U for the loci that cannot be classified even at the type level (e.g., the CRISPR/Cas system in *Acidithiobacillus ferrooxidans* ATCC 23270 discussed below).

We propose to retain the well-established names for core genes of the information processing (ubiquitous *cas1* and *cas2*) and executive (*cas3-6*, characteristic for Type I system) components of the CRISPR/Cas systems. In several cases for which orthology could be confidently traced, we extend the usage of these six *cas* gene names (thus, *cmx5* of I-C is renamed *cas5* and *cmx6* is renamed *cas6*). In cases when significant sequence similarity between Cas proteins is observed but orthologous relationships cannot be definitively assigned, we use an additional letter derived from the subtype label (hence *cas6e* and *cas6f*, former *cse3* and *csy4,* respectively, which are likely to be extremely divergent derivatives of *cas6*; Table 1).

In Type I systems, there are two additional genes for which orthology is readily detectable between different subtypes. We refer to these genes as *cas7* and *cas8(abc);* both encode subunits of the Cascade complex (Table 1). The *cas8a*, *cas8b* and *cas8c* genes are the signature genes for subtypes I-A, I-B and I-C, respectively. In type II and type III systems, the respective signature genes are designated *cas9* (formerly *csn1* and *csx12*) and *cas10* (*cmr2*, *csm1* and *csx11*).

When a gene clearly is a fusion or fission of established genes, we propose an *ad hoc* nomenclature indicating the relationship of this variant to the "canonical" forms:

thus, *cas2/cas3* in I-F systems denotes a fusion of *cas2* and *cas3,* whereas *cas3'* and *cas3''* denote the solo helicase domain and the solo HD domain, respectively.

For other, less common genes that have been named previously[13], the "legacy" nomenclature can be used along with the family classification given in Table 1 (eg. *csx9*). Due to the fact that many of the Cas protein sequences are highly diverged, it is expected that with the increasing representation of sequences and structures, many of these genes eventually will be included into existing families. We propose to continue assigning further "numerical" names to newly merged orhologous families in the future (*cas11*, *cas12* etc.).

For the remaining CRISPR-associated genes, we propose to assign interim gene names (*csx1*, where "x" stands for unclassified family), with an indication of the (super)family where known (e.g. *csx1*, COG1517 family or *csx10*, RAMP superfamily).

Outstanding problems

Subtype assignment. As pointed out above, the phylogenetic tree of Cas1 reproduces most of the previously established groups fairly well, with the exception of the Type III systems (Figure 3). However, for the deep branches, assigning a subtype can be problematic. In many cases, detailed analysis of the gene orders reveals a more complicated picture, with different arrangements of *cas* genes in the operons, apparently, due to frequent horizontal gene transfer (HGT) and recombination that involve the CRISPR loci. In particular, a notable recombinant CRISPR/Cas system is present at least

in certain cyanobacteria (e.g., *Synechocystis* sp. PCC 6803: slr7010-ssr7072) and Archaea. In this case, the Type I-C system combines with a distinct polymerase/RAMP module genes in the following arrangement: *cas3*, *cas10* (predicted inactivated polymerase with a HD-domain), *csc2* (COG1337 family, RAMP superfamily), *csc2* (RAMP subfamily), *cas6*, *cas4*, *cas1*, *cas2*. This peculiar hybrid system containing signature genes for both type I and type III systems is represented in a variety of genomes (and thus likely functional), so we introduce it as a new subtype I-D (Figure1).

Another interesting CRISPR/Cas system typified by *Acidithiobacillus ferrooxidans* ATCC 23270 (AFE_1037-AFE_1040) was found so far only in four genomes. This locus appears to possess a distinct gene content and potentially could contribute to our understanding of the functions and evolution of CRISPR/Cas systems in general. This system contains neither of the two ubiquitous core genes (*cas1* or *cas2*) nor any other signature genes of the three CRISPR/Cas types or the 10 subtypes. The *A. ferrooxidans* system consists of four genes denoted *csf1*, *csf2*, *csf3* and *csf4* (TIGRFAMS: TIGR03114, TIGR03115, TIGR03116, TIGR03117, respectively). These genes encode, respectively, a Zn-finger domain containing protein, a protein containing two RAMP domains, another distinct RAMP protein and a DinG-like helicase of the XPD family [30]. According to the CRISPRdb database [32], a CRISPR array is present in the vicinity of the above four genes in all of the respective genomes; the architecture of these arrays is unique in each genome. Thus, this system might function in conjunction with different CRISPR arrays and does not require a distinct repeat signature. Indeed, three of the four genomes containing this system additionally possess *cas1* and *cas2* genes that

are located in other parts of these genomes and are associated with Type I CRISPR/Cas systems. It remains unclear whether this is a self-sufficient system or rather a defective system that captures and utilizes preexisting CRISPR arrays generated by other, Cas1-containing CRISPR/Cas systems. More data are needed to classify this novel system as a separate CRISPR/Cas type, but this finding illustrates the diversity of CRISPR systems and the challenges associated with their classification.

Gene name assignments. Many *cas* genes, in particular genes that encode RAMP proteins, evolve at exceptionally high rates. Gene (protein) family assignment becomes increasingly complicated with the appearance of CRISPR/Cas systems containing genes that encode highly divergent proteins that, after the structure is solved, might (or might not) fall into a known Cas protein family. For example, a CRISPR system very similar to that of subtype I-F (Ypest/CASS3 as determined by Cas1 similarity) is present in *Photobacterium profundum* and several other bacteria. This system includes two proteins, PBPRB1992 and PBPRB1993, that show no significant sequence similarity to any Cas proteins. However, analysis of the sequence motifs that are conserved in these proteins, the predicted secondary structure, as well as the length and position of the corresponding genes in the operon, strongly suggest that they belong to the Cas7 and Cas5 families of RAMPs, respectively. Another example includes the CRISPR/Cas system of *Geobacter sulfurreducens* that, according to the phylogeny of Cas1, should be assigned to the I-C subtype. This operon encodes three uncharacterized proteins GSU0052, GSU0053, GSU0054; the last two proteins contain several motifs similar to the characteristic motifs of the RAMP superfamily and thus might be RAMP homologs (Table 1). However, none

of these proteins could be linked to known Cas families, even using the most sensitive of the available methods for remote sequence similarity detection [33-35]. Thus, only comparison of solved structures will shed light on the relationships between these and other highly diverged Cas proteins and known Cas families. In such cases, assignment of new gene names appears to be premature because these proteins are likely to eventually assume already existing names. Therefore, it is proposed that these genes are given temporary "*csx*" names.

Many CRISPR loci belong to "islands" that contain various "high-mobility" genes such as components of other defense systems, toxin-antitoxins, and transposases [36]. Some of these genes can be erroneously linked to CRISPR/Cas systems, so caution should be exercised in the classification and naming of genes as "cas" or even "csx" before functional connections with CRISPR/Cas systems are convincingly established.

An additional challenge to the nomenclature is presented by the variable domain architectures of some of the Cas proteins including domain fusions and fissions as discussed above for the Cas3 protein. Other notable fusions include *cas2-cas3* in the Ypest/CASS3 system, *cas1-cas4* (eg. GSU0057 from *Geobacter sulfurreducens*), *cas1* and DEDDh family exonuclease (eg. LBUL_0800 from *Lactobacillus delbrueckii subsp. bulgaricus*), *cas1* and reverse transcriptase fusion (eg. VVA1544 from *Vibrio vulnificus*), and more.

In several genomes, homologs of some *cas* genes also appear in contexts different from CRISPR/Cas systems; these proteins might either represent (components of) distinct antivirus defense systems, or they could be involved in other functions such

as DNA repair. These proteins include RAMPs of the COG5551 subfamily, COG1517, COG1468 and COG3513 families. In cases like this, classification and labeling of such genes as "*cas*" should be avoided.

The CRISPR arrays contain few stop codons and accordingly are often erroneously translated into bogus "hypothetical proteins". Unfortunately, these artifacts then enter the databases and tend to be amplified during the analysis of new genomes, so currently there are at least two Pfam entries each of which consists of non-existent "pseudo-Cas proteins" (pfam11194; pfam11664). Care should be taken during the annotation of new genome sequences to avoid further proliferation of these errors.

Conclusion

Given the complexity and the highly dynamic mode of evolution of the CRISPR/Cas systems, it would be counter-productive to attempt classification on the basis of any single criterion, for instance, the phylogeny of Cas1. Thus, we propose here a "polythetic" classification that integrates the phylogenies of the conserved *cas* genes, sequences and structural similarities between other Cas proteins, and the composition and organization of the (putative) operons. It should be emphasized that a robust family classification of the Cas proteins, many of which diverge rapidly, is not only a matter of convenient description, but also a basis for experimental validation of the respective functional predictions. Therefore it is important that this classification is continuously updated and revised when necessary, using new sequence and structure information combined with state of the art computational methods. The classification described in this

414 article is available at the NCBI website along with tools for identification of Cas proteins
415 (ftp://ftp.ncbi.nih.gov/pub/wolf/_suppl/CRISPRclass/index.html). In the future, a fine-
416 grain classification of the CRISPR/Cas systems should become feasible on the basis of
417 phylogenies and structures of Cas proteins, *cas* operon organizations of *cas* genes, and
418 CRISPR repeat architectures.
419
420 **Acknowledgements**
421 The authors thank Michael Terns for critical reading of the manuscript and useful
422 discussions. KSM, YIW and EVK are supported by the intramural funds of the US
423 Department of Health and Human Services (National Library of Medicine); E.C.
424 acknowledges funding from Umeå University and the Swedish Research CouncilS.M.
425 acknowledges funding from NSERC of Canada (Discovery program); FJMM
426 acknowledges support from the University of Alicante (Vicerrectorado de Investigacion,
427 Desarrollo e Innovacion) for the use of its research technical services; AFY is supported
428 by Genome Canada (through the Ontario Genomics Institute; 2009-OGI-ABC-1405).
429
430

References

1 Deveau, H., Garneau, J. E. & Moineau, S. CRISPR/Cas System and Its Role in Phage-Bacteria Interactions. *Annu Rev Microbiol* **64**, 475-493 (2010).

2 Horvath, P. & Barrangou, R. CRISPR/Cas, the immune system of bacteria and archaea. *Science* **327**, 167-170 (2010).

3 Karginov, F. V. & Hannon, G. J. The CRISPR system: small RNA-guided defense in bacteria and archaea. *Mol Cell* **37**, 7-19 (2010).

4 Koonin, E. V. & Makarova, K. S. CRISPR-Cas: an adaptive immunity system in prokaryotes. *F1000 Biol Rep* **1**, 95 (2009).

5 Sorek, R., Kunin, V. & Hugenholtz, P. CRISPR--a widespread system that provides acquired resistance against phages in bacteria and archaea. *Nat Rev Microbiol* **6**, 181-186 (2008).

6 van der Oost, J., Jore, M. M., Westra, E. R., Lundgren, M. & Brouns, S. J. CRISPR-based adaptive and heritable immunity in prokaryotes. *Trends Biochem Sci* **34**, 401-407 (2009).

7 Mojica, F. J., Diez-Villasenor, C., Soria, E. & Juez, G. Biological significance of a family of regularly spaced repeats in the genomes of Archaea, Bacteria and mitochondria. *Mol Microbiol* **36**, 244-246 (2000).

8 Jansen, R., Embden, J. D., Gaastra, W. & Schouls, L. M. Identification of genes that are associated with DNA repeats in prokaryotes. *Mol Microbiol* **43**, 1565-1575 (2002).

9 Makarova, K. S., Aravind, L., Grishin, N. V., Rogozin, I. B. & Koonin, E. V. A DNA repair system specific for thermophilic Archaea and bacteria predicted by genomic context analysis. *Nucleic Acids Res* **30**, 482-496 (2002).

10 Mojica, F. J., Diez-Villasenor, C., Garcia-Martinez, J. & Soria, E. Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. *J Mol Evol* **60**, 174-182 (2005).

11 Bolotin, A., Quinquis, B., Sorokin, A. & Ehrlich, S. D. Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. *Microbiology* **151**, 2551-2561 (2005).

12 Pourcel, C., Salvignol, G. & Vergnaud, G. CRISPR elements in Yersinia pestis acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies. *Microbiology* **151**, 653-663 (2005).

13 Haft, D. H., Selengut, J., Mongodin, E. F. & Nelson, K. E. A guild of 45 CRISPR-associated (Cas) protein families and multiple CRISPR/Cas subtypes exist in prokaryotic genomes. *PLoS Comput Biol* **1**, e60 (2005).

14 Makarova, K. S., Grishin, N. V., Shabalina, S. A., Wolf, Y. I. & Koonin, E. V. 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* **1**, 7 (2006).

15 Carthew, R. W. & Sontheimer, E. J. Origins and Mechanisms of miRNAs and siRNAs. *Cell* **136**, 642-655 (2009).

16 Barrangou, R. *et al.* CRISPR provides acquired resistance against viruses in prokaryotes. *Science* **315**, 1709-1712 (2007).
17 Garneau, J. E. *et al.* The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. *Nature* **468**, 67-71 (2010).
18 Sontheimer, E. J. & Marraffini, L. A. Microbiology: slicer for DNA. *Nature* **468**, 45-46 (2010).
19 Mojica, F. J., Diez-Villasenor, C., Garcia-Martinez, J. & Almendros, C. Short motif sequences determine the targets of the prokaryotic CRISPR defence system. *Microbiology* **155**, 733-740 (2009).
20 Marraffini, L. A. & Sontheimer, E. J. Self versus non-self discrimination during CRISPR RNA-directed immunity. *Nature* **463**, 568-571 (2010).
21 Brouns, S. J. *et al.* Small CRISPR RNAs guide antiviral defense in prokaryotes. *Science* **321**, 960-964 (2008).
22 Carte, J., Wang, R., Li, H., Terns, R. M. & Terns, M. P. Cas6 is an endoribonuclease that generates guide RNAs for invader defense in prokaryotes. *Genes Dev* **22**, 3489-3496 (2008).
23 Hale, C. R. *et al.* RNA-guided RNA cleavage by a CRISPR RNA-Cas protein complex. *Cell* **139**, 945-956 (2009).
24 Haurwitz, R. E., Jinek, M., Wiedenheft, B., Zhou, K. & Doudna, J. A. Sequence- and structure-specific RNA processing by a CRISPR endonuclease. *Science* **329**, 1355-1358 (2010).
25 Marraffini, L. A. & Sontheimer, E. J. CRISPR interference limits horizontal gene transfer in staphylococci by targeting DNA. *Science* **322**, 1843-1845 (2008).
26 Wiedenheft, B. *et al.* Structural basis for DNase activity of a conserved protein implicated in CRISPR-mediated genome defense. *Structure* **17**, 904-912 (2009).
27 Beloglazova, N. *et al.* A novel family of sequence-specific endoribonucleases associated with the clustered regularly interspaced short palindromic repeats. *J Biol Chem* **283**, 20361-20371 (2008).
28 Kleanthous, C. *et al.* Structural and mechanistic basis of immunity toward endonuclease colicins. *Nat Struct Biol* **6**, 243-252 (1999).
29 Jakubauskas, A., Giedriene, J., Bujnicki, J. M. & Janulaitis, A. Identification of a single HNH active site in type IIS restriction endonuclease Eco31I. *J Mol Biol* **370**, 157-169 (2007).
30 White, M. F. Structure, function and evolution of the XPD family of iron-sulfur-containing 5'-->3' DNA helicases. *Biochem Soc Trans* **37**, 547-551 (2009).
31 Kunin, V., Sorek, R. & Hugenholtz, P. Evolutionary conservation of sequence and secondary structures in CRISPR repeats. *Genome Biol* **8**, R61 (2007).
32 Grissa, I., Vergnaud, G. & Pourcel, C. The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats. *BMC Bioinformatics* **8**, 172 (2007).
33 Altschul, S. F. & Koonin, E. V. PSI-BLAST - a tool for making discoveries in sequence databases. *Trends Biochem Sci.* **23**, 444-447 (1998).
34 Marchler-Bauer, A. & Bryant, S. H. CD-Search: protein domain annotations on the fly. *Nucleic Acids Res* **32**, W327-331 (2004).

35 Soding, J., Remmert, M., Biegert, A. & Lupas, A. N. HHsenser: exhaustive transitive profile search using HMM-HMM comparison. *Nucleic Acids Res* **34**, W374-378 (2006).
36 Makarova, K. S., Wolf, Y. I., van der Oost, J. & Koonin, E. V. Prokaryotic homologs of Argonaute proteins are predicted to function as key components of a novel system of defense against mobile genetic elements. *Biol Direct* **4**, 29, (2009).
37 Deveau, H. *et al.* Phage response to CRISPR-encoded resistance in *Streptococcus thermophilus*. *J Bacteriol* **190**, 1390-1400 (2008).
38 Guindon, S. & Gascuel, O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol* **52**, 696-704 (2003).
39 Han, D., Lehmann, K. & Krauss, G. SSO1450--a CAS1 protein from Sulfolobus solfataricus P2 with high affinity for RNA and DNA. *FEBS Lett* **583**, 1928-1932 (2009).
40 Han, D. & Krauss, G. Characterization of the endonuclease SSO2001 from Sulfolobus solfataricus P2. *FEBS Lett* **583**, 771-776 (2009).
41 Guy, C. P., Majernik, A. I., Chong, J. P. & Bolt, E. L. A novel nuclease-ATPase (Nar71) from archaea is part of a proposed thermophilic DNA repair system. *Nucleic Acids Res* **32**, 6176-6186 (2004).
42 Selengut, J. D. *et al.* TIGRFAMs and Genome Properties: tools for the assignment of molecular function and biological process in prokaryotic genomes. *Nucleic Acids Res* **35**, D260-264 (2007).

Figure legends

Figure 1. **The current model of CRISPR/Cas mechanism**.

The figure shows a schematic representation of the three stages of the CRISPR/Cas action based on a summary of biochemical and genetic data for the CRISPR/Cas systems from different organisms and some specific details for the *Escherichia coli* system (Type IE, see Fig.2).

(1) Adaptation. DNA fragments of an invading virus/plasmid are integrated as a new unit at the leader side of a CRISPR in the host chromosome, with the simultaneous duplication of a repeat. Selection of sequences to be integrated (proto-spacers) is likely to depend on a proto-spacer adjacent motif (* PAM)[19]. Although no direct evidence is presently available on the mechanism of spacer acquisition, the core of the CRISPR/Cas system (Cas1/Cas2), possibly with additional proteins, are prime candidates[16,21]. In Type II and Type III systems, substantial variations have been demonstrated (see text).

(2) Expression and Processing. In the Type IE system, a Cascade complex is responsible for processing of pre-crRNA to crRNA via a single cut in the repeat part[21].

(3) Interference. In *E. coli*, the Cascade complex with the crRNA guide target the complementary DNA of an invading virus/plasmid, and Cas3 is required, most likely, to cleave the alien DNA through the endonuclease activity of the HD domain[21]. The PAM (*) appears to play an important role in the interference process[20,37]. In Type II and Type III systems, no Cas3 ortholog is involved.

Figure 2. **The three major types and 10 subtypes of CRISPR systems and their relationships**

The operon organization cartoons and color scheme were generally adopted from Ref. 14, with the addition of the identification of Cas7 (COG1857) as a member of the RAMP superfamily. Orthologous genes are color-coded and identified by a family name as in Table 1: bold for definitive proposals and regular font for "legacy names". The signature genes for CRISPR/cas types are shown within gree*n boxes*, and for subtypes within red boxes. The letters above the genes show major categories of Cas proteins: L, large CASCADE subunit; S, small CASCADE subunit; R, RAMP CASCADE subunit; RE, RAMP family RNase involved in crRNA processing (experimentally characterized nucleases shown be asterisks); T, transcriptional regulator. Genes coding for inactivated (putative) polymerases are shown by crosses. Question marks denote tentative predictions based on weak sequence similarity. For subtype I-A, the *cas8a1* and *cas8a2* genes are typically mutually exclusive but both can be considered signature genes for the subtype. For type III systems, the *cas1* and *cas2* genes are shown in a dashed box that indicates the loose association of these genes with the type III polymerase/RAMP modules.

Figure 3. **A schematic representation of the phylogenetic tree for Cas1/COG1518 proteins.**

The maximum likelihood tree was constructed using the PHYML program [38] from 182 informative positions in the multiple alignment of a representative set of 228 Cas1 proteins from 442 complete genomes. Six major CRISPR/Cas system subtypes of Type I systems as well as type II and type III systems are color-coded. Dashed lines show *cas*1 genes found in "hybrid" CRISPR loci containing genes from both Type I and Type III CRISPR/Cas systems (see text). The subtypes of CRISPR/Cas systems are denoted as in Figure 2. Subtypes I-U, II-U and III-U (U for unclassified) denote CRISPR/Cas systems that lack currently defined subtype-specific signature genes (see text).

Table 1. Classification and nomenclature of CRISPR genes

Proposed gene name	Type or subtype	[a]Name from[13]	Name from[21]	Structure (PDB code)	[b]Family (superfamily)	Representatives	Comment and references
cas1	Type I, Type II, Type III	*cas1*	*cas1*	3GOD, 3LFX, 2YZS	COG1518	ygbT, SPy1047, SERP2463	metal-dependent deoxyribonuclease; a unique fold consisting of a N-terminal β strand domain and a C-terminal α-helical domain[26,39]; also binds RNA[26,39]
cas2	Type I, Type II, Type III	*cas2*	*cas2*	2IVY, 2I8E, 3EXC	COG1343, COG3512	ygbF, SPy1047, SERP2462, y1723 (N-term. domain)	small protein related to VapD; shown to be a RNAse specific to U-rich regions[27]
cas3'	**Type I**	*cas3*	*cas3*	-	COG1203	ygcB, APE1232	DNA helicase; most proteins have fusion to HD nuclease [21] (*cas3'*)
cas3''	I-A, I-B	-	-	-	COG2254	APE1231, BH0336	HD-like nuclease, specifically digesting double-stranded oligonucleotides and preferably cleaving at G:C pairs [40]
cas4	I-A, I-B, I-C, I-D, II-B	*cas4, csa1*	-	-	COG1468	APE1239, BH0340	RecB-like nuclease with three-cysteine C-terminal cluster
cas5	I-A, I-B, I-C, I-E	*cas5e,d,a,t,h ,p, cmx5*	*casD*	3KG4	COG1688 (RAMP)	APE1234, BH0337, DevS, ygcI	predicted subunit of the Cascade complex [21]; in subtype I-C this protein might be the endoribonuclease that generates crRNAs
cas6	I-A, I-B, I-D, III-A, III-B	*cas6, cmx6*	-	3I4H	COG1583 COG5551 (RAMP)	PF1131, slr7014	Cas6 is an endoribonuclease that generates crRNAs[22,23], predicted subunit of Cascade complex
cas6e	I-E	*cse3*	*casE*	1WJ9	(RAMP)	ygcH	homologous to Cas6, but distinct family
cas6f	I-F	*csy4*	-	2XLJ	(RAMP)	y1727	homologous to Cas6, but distinct family; shown to be is an endoribonuclease that generates crRNAs [21]
cas7	I-A, I-B, I-C, I-E	*cse4, csd2, csh2, cst2, csa2, csp1*	*casC*	-	COG1857, COG3649 (RAMP)	YgcJ, DevR	α/β protein; subunit of Cascade complex [21]
cas8a1	**I-A**	*csx8, cmx1, csp2, cst1, CxxC-CxxC, csx13*	-	-	BH0338-like	BH0338, MTH1090, TM1802, LA3191[c]	large proteins, some contain Zn-finger domain; nuclease activity has been reported for MTH1090 [41]

Cas8a2	I-A	*csa4, csx9*	-	-	PH0918	MJ0385, PF0637, AF0070, PH0918, SSO1401, AF1873	see cas8a
cas8b	**I-B**	*csh1, TM1802*,	-	-	-		see cas8a
cas8c	**I-C**	*csd1*	-	-	-		see cas8a
cas9	**Type II**	*csn1, csx12*	-	-	COG3513	SPy1046, FTN_0757	very large protein containing McrA/HNH-nuclease related domain and a RuvC-like nuclease domain;
cas10	**Type III**	*cmr2, csm1, csx11*	-	-	COG1353	MTH326, alr1562[c], slr7011[c]	multidomain protein with permuted HD nuclease domain, palm domain, polymerase-thumb-like domain and Zn-ribbon; MTH326-like has inactivated polymerase catalytic domain; alr1562 and slr7011 – predicted only on the basis of size, presence of HD domain, and location with RAMPs in one operon; subunit of Cmr complex[23]
cas10d	**I-D**	*csc3*	-	-	COG1353	slr7011	inactivated homolog of Cas10, contains N-terminal HD domain
csy1	**I-F**	*csy1*	-	-	y1724-like	y1724	~450 aa protein, predicted to be a subunit of Cascade complex
csy2	I-F	*csy2*	-	-	(RAMP)	y1725	predicted Cas7 ortholog
csy3	I-F	*csy3*	-	-	(RAMP)	y1726	predicted Cas5 ortholog
cse1	**I-E**	*cse1*	*casA*	-	ygcL-like	ygcL	large Zn-finger containing proteins; a subunit of the Cascade complex [21]; signature gene for Ecoli/CASS2 subtype
cse2	I-E	*cse2*	*casB*	2ZCA	ygcK-like	ygcK	~180 aa protein; a subunit of the Cascade complex[21]
csc1	I-D	*csc1*	-	-	alr1563-like (RAMP)	alr1563	
csc2	I-D	*csc2, csc1*	-	-	COG1337 (RAMP)	slr7012	
csa5	I-A	*csa5*	-	-	AF1870	AF1870, SSO1398, PF0643, MJ0380	~150 aa protein; alpha-helical

csn2	II-A	*csn2*	-	-	SPy1049-like	SPy1049	~220 aa protein; predicted to be a functional analog of Cas4[6] based on anti-correlated phyletic patterns
csm2	**III-A**	*csm2*	-	-	COG1421	MTH1081, SERP2460	~150 aa protein; mostly α-helical protein; part of distinct polymerase cassette
csm3	III-A	*csm3, csc2*	-	-	COG1337 (RAMP)	SERP2459, MTH1080	
csm4	III-A	*csm4*	-	-	COG1567 (RAMP)	SERP2458, MTH1079	
csm5	III-A	*csm5*	-	-	COG1332 (RAMP)	SERP2457, MTH1078	
csm6	III-A	*csm6, APE2256*	-	2WTE	COG1517	SSO1445, APE2256	HTH-type transcriptional regulator; often fused to COG1517-like domain
cmr1	III-B	*cmr1*	-	-	COG1367, (RAMP)	PF1130	subunit of Cmr complex[23]
cmr3	III-B	*cmr3*	-	-	COG1769 (RAMP)	PF1128	subunit of Cmr complex[23]
cmr4	III-B	*cmr4*	-	-	COG1336 (RAMP)	PF1126	subunit of Cmr complex[23]
cmr5	**III-B**	*cmr5*	-	2ZOP, 2OEB	COG3337	PF1125, MTH324	subunit of Cmr complex[23]
cmr6	III-B	*cmr6*	-	-	COG1604 (RAMP)	PF1124	subunit of Cmr complex[23]
csb2	I-U [d]	-	-	-	(RAMP?)	GSU0054, Balac_1305	Contains RAMP superfamily motif (G-rich loop) and conserved histidine at the C-terminus
csb1	I-U	*GSU0053*	-	-	(RAMP?)	GSU0053, Balac_1306	Contains several motifs similar to Cas7 family
csb3	I-U	-	-	-	(RAMP?)	Balac_1303	Contains RAMP superfamily motif (G-rich loop) and conserved histidine at the C-terminus
csx17	I-U	-	-	-	-	Btus_2683	
csx14	I-U	-	-	-	-	GSU0052	
csx10	I-U	*csx10*	-	-	(RAMP)	Caur_2274	
csp2	I-U	*csp2*	-	-	PG2018	PG2018	predicted Cas8 ortholog
csx16	III-U	*VVA1548*	-	-	-	VVA1548	~100 aa protein; often seen in proximity to COG1517
csaX	III-U	*csaX*	-	-	-	SSO1438	~350 aa protein, no prediction
csx3	III-U	*csx3*	-	-	-	AF1864	~ 100 aa domain, in some cases fused to COG1517 family domains

csx1	III-U	*csa3, csx1* (DxTHG motif)*, csx2, NE0113, TIGR02710*	-	1XMX, 2I71	COG1517, COG4006	PF1127, MJ1666, TM1812, NE0113	some are fused to HTH domain (see COG1517/HTH); some proteins have the domain duplication; some have a fusion with HTH and RecB-family nuclease domain; domain appears to have a Rossmann-like fold.
csx15	???	-	-	-	TTE2665	TTE2665	~130 aa protein, no prediction; some are fused to AAA ATPase domain
csf4	Type U	*csf4*	-	-	-	AFE_1037	DinG family helicase
csf3	Type U	*csf3*	-	-	(RAMP)	AFE_1040	
csf2	Type U	*csf2*	-	-	(RAMP)	AFE_1039	
csf1	Type U	*csf1*	-	-	-	AFE_1038	Zn-finger domain

Note:

[a] Subsequent to the original publication[13], Haft et al. introduced a number of new types of the CRISPR system and gene names that are included in the TIGRFAM database [42] but mostly fit into previously described gene/protein families. Most of these new names are included in this column.

[b] Most of the families correspond to those proposed by Makarova et al.[14] with a few changes and additions.

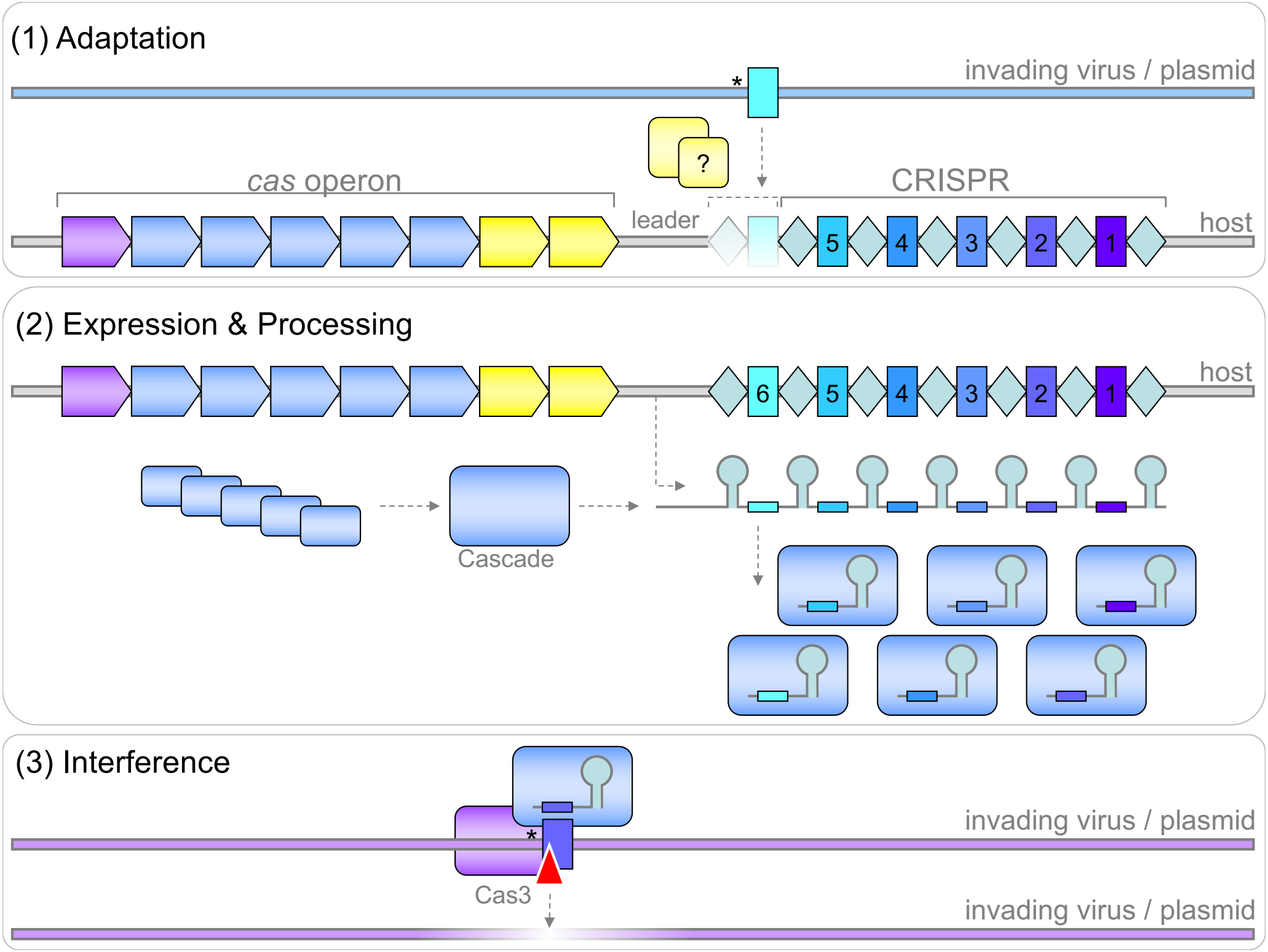
[c] Tentative predictions based on weak sequence similarity, sequence length and gene order in an operon.

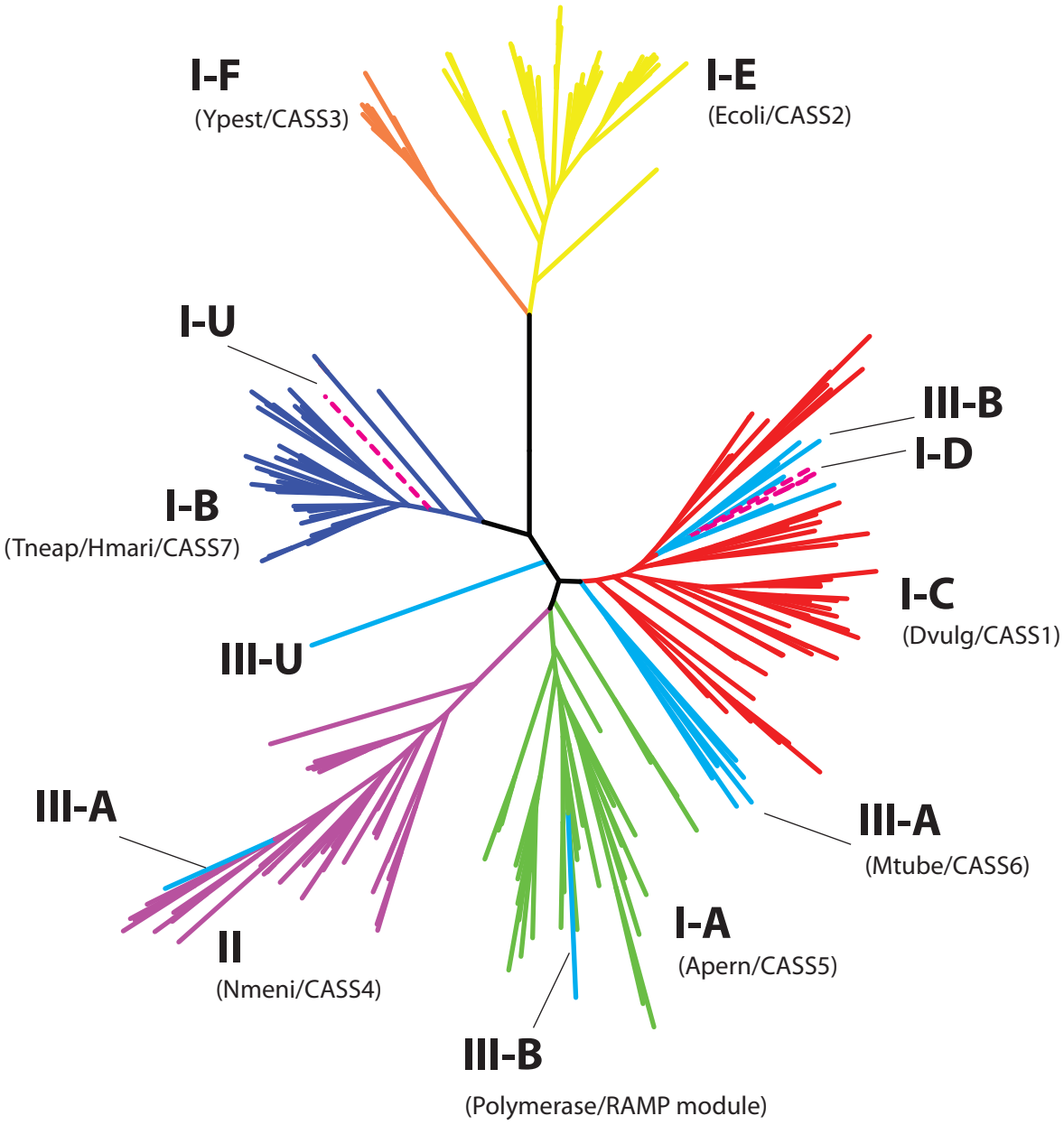
[d] Unclassified.

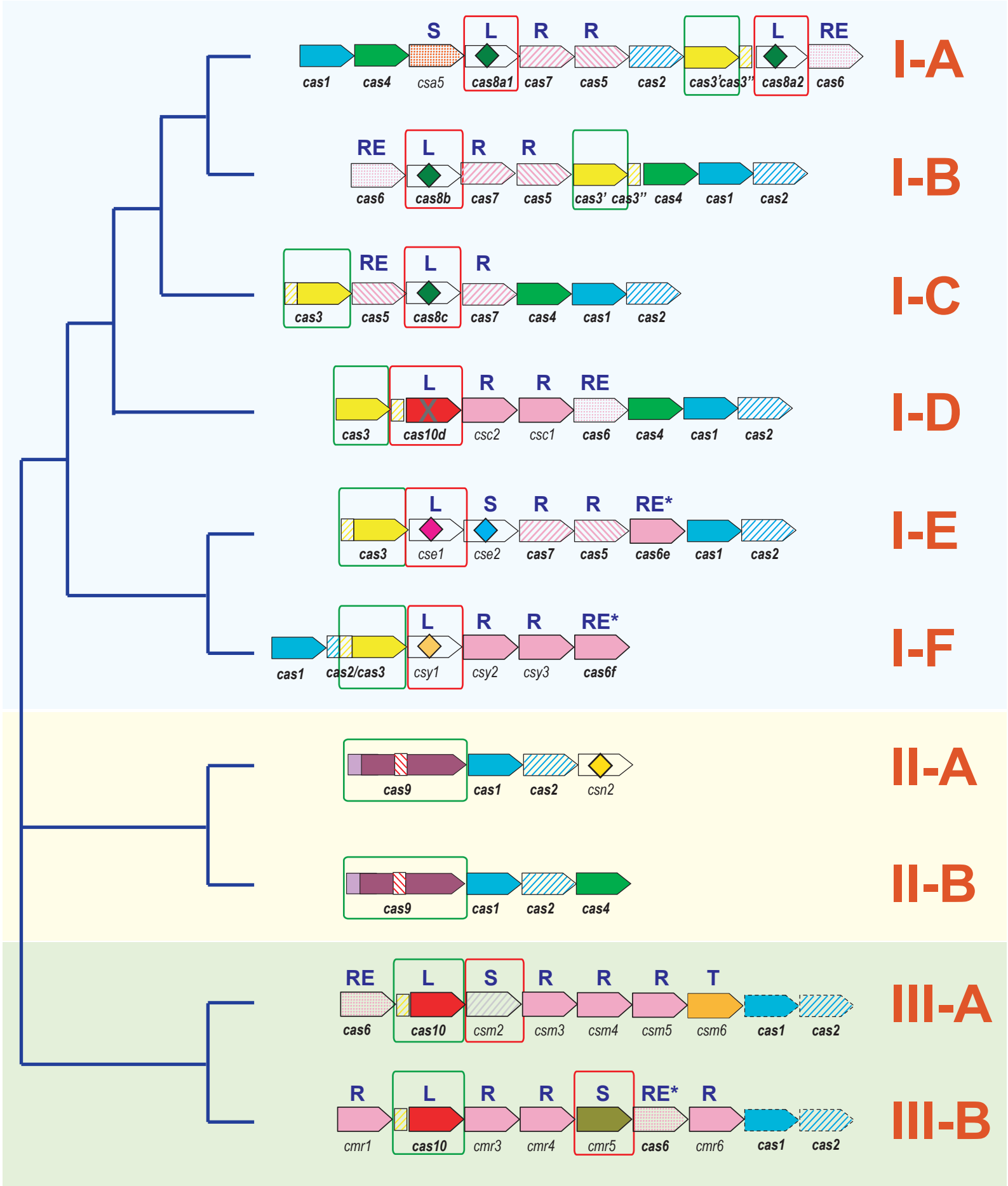
Signature genes for CRISPR/Cas system types and subtypes are shown in the second column in bold.

Table 2. **Taxonomic distribution of three CRISPR system types**

	Genomes analyzed	**Cas1**	**Fraction of genomes with cas1**	**Type I (Cas7 and Cas3)**	**Type II (Cas9)**	**Type III (Cas10)**
Archaea	67	54	0.81	50	0	40
Bacteria	639	256	0.40	245	65	99
Crenarchaeota	17	15	0.88	15	0	16
Euryarchaeota	47	37	0.79	33	0	23
Actinobacteria	72	26	0.36	28	15	8
Aquificae	7	5	0.71	7	1	4
Bacteroidetes/ Chlorobi group	32	16	0.50	14	2	6
Chlamydiae/ Verrucomicrobia group	10	2	0.20	0	1	1
Chloroflexi	10	9	0.90	9	2	7
Cyanobacteria	14	7	0.50	7	1	7
Firmicutes	126	56	0.44	40	17	23
Proteobacteria	318	107	0.34	117	20	22
Spirochaetes	13	3	0.23	2	1	0
Thermotogae	11	10	0.91	10	0	9


(1) Adaptation
invading virus / plasmid
*
?
cas operon
CRISPR
leader
host
5
4
3
2
1
(2) Expression & Processing
host
6
5
4
3
2
1
Cascade
(3) Interference
invading virus / plasmid
*
Cas3
invading virus / plasmid


I-F
(Ypest/CASS3)
I-E
(Ecoli/CASS2)
I-U
III-B
I-D
I-B
(Tneap/Hmari/CASS7)
I-C
(Dvulg/CASS1)
III-U
III-A
III-A
(Mtube/CASS6)
I-A
(Apern/CASS5)
II
(Nmeni/CASS4)
III-B
(Polymerase/RAMP module)


S L R R L RE
cas1 cas4 csa5 cas8a1 cas7 cas5 cas2 cas3'cas3'' cas8a2 cas6
I-A
RE L R R
cas6 cas8b cas7 cas5 cas3' cas3'' cas4 cas1 cas2
I-B
RE L R
cas3 cas5 cas8c cas7 cas4 cas1 cas2
I-C
L R R RE
cas3 cas10d csc2 csc1 cas6 cas4 cas1 cas2
I-D
L S R R RE*
cas3 cse1 cse2 cas7 cas5 cas6e cas1 cas2
I-E
L R R RE*
cas1 cas2/cas3 csy1 csy2 csy3 cas6f
I-F
cas9 cas1 cas2 csn2
II-A
cas9 cas1 cas2 cas4
II-B
RE L S R R R T
cas6 cas10 csm2 csm3 csm4 csm5 csm6 cas1 cas2
III-A
R L R R S RE* R
cmr1 cas10 cmr3 cmr4 cmr5 cas6 cmr6 cas1 cas2
III-B