/631/326/4041	Biological sciences	Microbiology	CRISPR-Cas systems
/631/181/2474	Biological sciences	Evolution	Evolutionary genetics
/631/181/735	Biological sciences	Evolution	Molecular evolution
/631/326/325	Biological sciences	Microbiology	Microbial genetics
/631/250/2152	Biological sciences	Immunology	Adaptive immunity
/631/326/1321	Biological sciences	Microbiology	Bacteriophages
/631/114	Biological sciences	Computational biology and bioinformatics	
/631/326/26/2526	Biological sciences	Microbiology	Archaea Archaeal genomics
/631/326/41/2530	Biological sciences	Microbiology	Bacteria Bacterial genomics

Evolutionary classification of CRISPR–Cas systems: a burst of class 2 and derived variants

Kira S. Makarova[1], Yuri I. Wolf[1], Jamie Iranzo[1], Sergey A. Shmakov[1], Omer S. Alkhnbashi[2], Stan J. J. Brouns[3], Emmanuelle Charpentier[4], David Cheng[5], Daniel H. Haft[1], Philippe Horvath[6], Sylvain Moineau[7], Francisco J. M. Mojica[8], David Scott[5], Shiraz A. Shah[9], Virginijus Siksnys[10], Michael P. Terns[11], Česlovas Venclovas[10], Malcolm F. White[12], Alexander F. Yakunin[13,14], Winston Yan[5], Feng Zhang[15,16,17,18], Roger A. Garrett[19], Rolf Backofen[2,21], John van der Oost[22], Rodolphe Barrangou[23] and Eugene V. Koonin[1]*

1. National Center for Biotechnology Information, National Library of Medicine, Bethesda, MD, USA.
2. Bioinformatics group, Department of Computer Science, University of Freiberg, Freiberg, Germany.
3. Kavli Institute of Nanoscience, Department of Bionanoscience, Delft University of Technology, Delft, The Netherlands.
4. Max Planck Unit for the Science of Pathogens, Humboldt University, Berlin, Germany.

5. Arbor Biotechnologies, Cambridge, MA, USA.
6. DuPont Nutrition and Health, Dangé-Saint-Romain, France.
7. Département de biochimie, de microbiologie et de bio-informatique, Faculté des sciences et de génie, Groupe de recherche en écologie buccale, Félix d'Hérelle Reference Center for Bacterial Viruses, Faculté de médecine dentaire, Université Laval, Québec City, Québec, Canada.
8. Departamento de Fisiología, Genética y Microbiología. Universidad de Alicante, Alicante, Spain.
9. COPSAC, Copenhagen Prospective Studies on Asthma in Childhood, Herlev and Gentofte Hospital, University of Copenhagen, Gentofte, Denmark.
10. Institute of Biotechnology, Life Sciences Center, Vilnius University, Vilnius, Lithuania.
11. Biochemistry and Molecular Biology, Genetics and Microbiology, University of Georgia,, Athens, GA, USA.
12. Biomedical Sciences Research Complex, University of St. Andrews, St. Andrews, UK.
13. Department of Chemical Engineering and Applied Chemistry, University of Toronto, Toronto, Canada.
14. Centre for Environmental Biotechnology, School of Natural Sciences, Bangor University, Bangor, Gwynedd LL57 2UW, UK.
15. Broad Institute of MIT and Harvard, Cambridge, MA, USA.
16. McGovern Institute for Brain Research, Massachusetts Institute of Technology, Cambridge, MA, USA.
17. Howard Hughes Medical Institute, Cambridge, MA, USA.
18. Department of Brain and Cognitive Sciences and Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA.
19. Archaea Centre, Department of Biology, Copenhagen University, Copenhagen, Denmark.
20. BIOSS Centre for Biological Signaling Studies, Cluster of Excellence, University of Freiburg, Freiburg, Germany.
21. Laboratory of Microbiology, Wageningen University, Wageningen, The Netherlands.
22. Department of Food, Bioprocessing, and Nutrition Sciences, North Carolina State University, Raleigh, NC, USA.

*e-mail: koonin@ncbi.nlm.nih.gov

Abstract

The number and diversity of CRISPR–Cas systems has substantially increased in recent years. Here, we provide an updated evolutionary classification of CRISPR–Cas systems and *cas* genes, with an emphasis on the major developments that occurred since the publication of the latest classification in 2015. The new classification includes 2 classes, 6 types and 33 subtypes compared to 5 types and 16 subtypes in 2015. A key development is the ongoing discovery of multiple, novel class 2 CRISPR–Cas systems that now include 3 types and 17 subtypes. A second major novelty is the discovery of numerous derived CRISPR–Cas variants, often associated with mobile genetic elements that lack the nucleases required for interference. Some of these variants are involved in RNA-guided transposition whereas others are predicted to perform functions distinct from adaptive immunity that remain to be characterized experimentally. The third highlight is the discovery of numerous families of ancillary CRISPR-linked genes, often implicated in signal transduction. Together, these findings substantially clarify the functional diversity and evolutionary history of CRISPR–Cas.

[H1] Introduction

CRISPR–Cas systems, that are best known as key components of a new generation of genome engineering tools[1,2], naturally function as adaptive immunity mechanisms in bacteria and archaea. The CRISPR–Cas immune response consists of three main stages: adaptation, expression and interference. At the adaptation stage, a distinct complex of Cas proteins binds to a target DNA, often after recognizing a distinct, short motif known as PAM (protospacer-adjacent motif), and cleaves out a portion of the target DNA, the protospacer. After duplication of the repeat at the 5′end of the CRISPR array, the adaptation complex inserts the protospacer DNA into the array, so that it becomes a spacer. Some CRISPR–Cas systems employ an alternative mechanism of adaptation, namely spacer acquisition from RNA via reverse transcription by a reverse transcriptase encoded in the CRISPR–*cas* locus.

At the expression stage, the CRISPR array is typically transcribed as a single transcript — the pre-CRISPR(cr)RNA — that is processed into mature crRNAs, each containing the spacer sequence and parts of the flanking repeats. In different CRISPR–Cas variants, the pre-crRNA processing is mediated by a distinct subunit of a multi-protein Cas complex, by a single, multidomain Cas protein, or by non-Cas, host RNases.

At the interference stage, the crRNA that typically remains bound to the processing complex (protein) serves as the guide to recognize the protospacer (or a closely similar sequence) in an invading genome of a virus or plasmid that is then cleaved and inactivated by a Cas nuclease (or nucleases) that is either part of the effector or is recruited at the interference stage. The above is a brief, over-simplified description of CRISPR–Cas functionality that inevitably omits many details. These can be found in recent reviews on different aspects of CRISPR–Cas biology[3-9].

Similarly to other biological defence mechanisms, archaeal and bacterial CRISPR–Cas systems show remarkable diversity of Cas protein sequences, gene composition, and architecture of the genomic loci[3,5,10-15]. Our knowledge of this diversity is continuously expanding through screening of the ever growing genomic and metagenomic databases. To keep pace with such expansion, a robust classification of CRISPR–Cas systems based on evolutionary relationships is essential for the progress of CRISPR research, but presents formidable challenges due to the lack of universal markers and the fast evolution of the CRISPR–*cas* loci[16]. Therefore, the two

previous versions of CRISPR–Cas classification published in *Nature Reviews Microbiology* in 2011 and 2015 employed a multipronged approach that combined comparisons of gene composition of CRISPR–Cas systems and loci architectures with sequence similarity-base clustering and phylogenetic analysis of conserved Cas proteins, such as Cas1[17,18]. The 2015 classification included 5 types and 16 subtypes, as well as introduced the major division of CRISPR–Cas systems into two classes that radically differ with respect to the architectures of their effector modules involved in CRISPR RNA (crRNA) processing and interference. The class 1 systems have effector modules composed of multiple Cas proteins, some of which form crRNA-binding complexes (such as the Cascade complex in type I systems) that, with contributions from additional Cas proteins, mediate pre-crRNA processing and interference. By contrast, class 2 systems encompass a single, multidomain crRNA-binding protein (such as Cas9 in type II systems) that combines all activities required for interference and, in some variants, also those involved in pre-crRNA processing (Box 1).

Since the publication of the 2015 classification, there have been at least three major developments in the study of the diversity of CRISPR–Cas systems. First, driven partly by the interest in new tools for genome engineering, dedicated efforts have been undertaken to predict and experimentally validate additional class 2 systems[19-28]. As a result, the RNA-targeting type VI and multiple previously unknown subtypes of type V CRISPR–Cas systems have been discovered. Moreover, it has been shown that type V systems repeatedly evolved from transposon-encoded TnpB nucleases, yielding a large pool of type V variants, many of which can be expected to eventually become separate subtypes[20,29,30]. The second key development was the discovery of several class 1 and class 2 CRISPR–Cas variants that appear to lack the targeted cleavage activity and thus likely perform functions distinct from adaptive immunity[8,31,32]. Such derived CRISPR–Cas systems include type IV, several variants of type I, and at least one type V system variant, and are often encoded within mobile genetic elements (MGEs)[29,30,33]. Recently, the involvement of two of these derived CRISPR–Cas variants, encoded by Tn7-like transposons, in crRNA-dependent DNA transposition has been demonstrated experimentally[34,35]. Although the origin of some of these derived forms from particular class 1 subtypes is readily identifiable, their placement in the CRISPR–Cas classification scheme remain problematic. The third important finding involves the identification of numerous gene families that are associated

with specific variants of CRISPR–Cas systems, particularly, those of type III systems, and are implicated in signal transduction and regulatory roles[8,31,32,36,37].

In this article, we re-assess and update the classification of CRISPR–Cas systems using the previously developed strategies along with the analysis of the modular structure of bipartite networks of gene sharing. Special emphasis is put on the classification of quickly proliferating class 2 variants. The new class 2 classification now includes 3 types and 18 subtypes, compared to 2 types and 4 subtypes in the 2015 version, and opens the door for many more subtypes of type V systems to be identified. Although experimental study of the recently discovered class 2 variants is only in its initial phase, it is already clear that their properties are highly diverse and are difficult to predict from Cas protein sequences alone. Therefore, robust classification and systematic study of class 2 variants are essential for understanding their functionality in microorganisms and for the development of versatile genome editing tools. In addition, several distinct class 1 variants, including 3 new subtypes were identified, bringing the total number of CRISPR–Cas subtypes to 33. We describe the current state and prospects of the classification and nomenclature of CRISPR–Cas systems and *cas* genes, and additionally, outline the emerging scenario of CRISPR–Cas evolution.

[H1] The classification approach

No genes are shared by all CRISPR–Cas systems, ruling out the possibility of a straightforward, comprehensive phylogenetic classification analogous to that employed for cellular life forms. Instead, a multi-pronged computational strategy has been adopted that includes identification of signature genes for CRISPR–Cas types and subtypes, comparison of gene repertoires and genomic loci organization, as well as sequence similarity-based clustering and phylogenetic analysis of genes that are conserved in different subsets of CRISPR–Cas systems. Experimental data were also taken into consideration when available[16-18,38] (Box 2).

Briefly, in this work, 566 amino acid sequence profiles (see Supplementary Methods) representing all variants of the 13 core *cas* genes, several still uncharacterized components of effector complexes as well as reliably identified known ancillary genes were compared to the protein sequences that are annotated in 13,116 complete archaeal and bacterial genomes

available at the NCBI as of March 1, 2019, using position-specific iterated BLAST[39]. This search, followed by extensive manual curation, resulted in the identification of 7,915 CRISPR–*cas* loci (Supplementary Dataset 1) that were fit into the previously developed classification on the basis of presence of the respective signature Cas proteins; sequence similarity between Cas proteins; phylogenies of the most highly conserved Cas proteins (including Cas1 as well as effector proteins for individual types and subtypes); and conservation of the locus organization. The loci that did not meet the criteria for inclusion in any of the previously identified subtypes were assigned to new subtypes (Supplementary Tables 1 and 2, Supplementary Dataset 1). The updated collection of Cas protein family profiles (Supplementary Dataset 2) is a resource for identification of CRISPR–Cas systems in sequenced genomes and metagenomes.

Here, we additionally employed bipartite network analysis[40,41] (Supplementary Dataset 3) for the identification of cohesive modules in the network that reflect shared gene content together with Cas protein sequence conservation, and for helping to identify distinct CRISPR–Cas subgroups that might have sub- or neofunctionalized.

[H1] Functional modules and core genes

All *cas* genes can be subdivided into four distinct, although partially overlapping, functional modules (Box 1)[18,42].The adaptation module includes the gene encoding the key enzyme involved in spacer insertion (the Cas1 integrase) and the structural subunit of the adaptation complex Cas2, as well as the Cas4 nuclease in several CRISPR–Cas subtypes, the Csn2 protein in subtype II-A, and reverse transcriptase in many type III systems. The expression-processing module is responsible for pre-crRNA processing. In most class 1 systems, Cas6 is the enzyme that is directly responsible for processing. In type II systems, processing is catalyzed by the bacterial RNase III (a non-Cas protein) whereas, in many type V and apparently all type VI systems, the large effector Cas protein contains a distinct catalytic center responsible for processing. The interference or effector module is involved in target recognition and nucleic acid cleavage. In class 1 CRISPR–Cas systems, the effector module consists of multiple Cas proteins, namely, Cas3 (sometimes fused to Cas2), Cas5-Cas8 and Cas10-Cas11, in different combinations. By contrast, in class 2 systems, the effector module is represented by a single,

large protein, Cas9, Cas12 or Cas13. The signal transduction or ancillary module is a diffuse collection of CRISPR-linked genes for most of which their roles in CRISPR–Cas functions are, at best, tentatively predicted. However, for type III systems, a signal transduction pathway has been characterized, which involves activation of the Csm6 (Csx1) HEPN (higher eukaryotes and prokaryotes nucleotide-binding) RNase by cyclic oligoA synthesized by the Cas10 polymerase, showing that signal transduction can be essential for CRISPR-Cas function[36,37,43]. Comparative genomic analyses revealed partial independence of the adaptation and effector modules of CRISPR–Cas that, especially in the case of type III systems, appeared to have recombined on many independent occasions[44-46]. As a result, the topology of the phylogenetic tree of Cas1 shows only limited agreement with the CRISPR–Cas classification (Supplementary Dataset 4).

The classification of CRISPR–Cas systems is based primarily on Cas protein composition differences and sequence divergence between the effector modules[16,18]. The class 1 effector complexes involved in pre-crRNA processing and target recognition have similar organization between types I, III and IV although the sequence conservation among the three types is minimal[47,48]. The backbone of the effector complexes in all three class 1 types is formed by the distantly related RNA recognition motif (RRM) domain-containing proteins of the repeat-associated mysterious proteins (RAMPs) Cas5 and Cas7, the latter typically present in multiple copies. In most class 1 CRISPR–Cas systems, the third RAMP, Cas6, is the dedicated RNase responsible for the pre-crRNA processing that may or may not be physically associated with the effector complex. The RAMPs are characterized by extreme sequence divergence so that the sequences of Cas5, Cas6 and Cas7 from different subtypes could be linked only by using the most sensitive methods for profile against profile sequence comparison, or by direct comparison of protein structures. The large subunits of the effector complexes of type I and III systems, Cas8 and Cas10, respectively, occupy analogous positions in the complexes, but show no sequence similarity and, at best, only remote structural similarity. Whether or not Cas8 and Cas10 are homologous, remains an open question; if they are, the divergence is extreme, effectively, beyond detection[49]. Moreover, the Cas8 sequences show no detectable similarity even between some of the class 1 subtypes, such that the respective variants can serve as subtype signatures (Supplementary Table 2). The ‘small subunits’ of the class 1 effector complexes (Cas11), shows no statistically significant sequence similarity between type I and III systems, but the structural

similarity between the Cas11 proteins, as well as between Cas11 and the carboxyl-terminal α-helical domain of Cas10, strongly suggest that these are highly diverged homologs[47,50]. In type I systems, a key, stand-alone (although, in some variants, fused with Cas2) component of the effector module is Cas3, a large protein that typically consists of fused helicase and HD nuclease domains, and is directly responsible for the target DNA cleavage. Type III systems differ fundamentally, with the HD nuclease fused to Cas10, the large subunit of the complex involved in transcription-dependent cleavage of the target DNA.

Type IV CRISPR–Cas systems are highly derived variants that typically lack adaptation modules as well as the nucleases required for interference. Moreover, only Cas5 and Cas7 proteins are readily identifiable in type IV loci by sequence similarity with the counterparts in other types. Recent comparisons of the structures of the effector complexes of type IV and I systems have identified the type IV counterpart of the large subunit of the effector complex[48], suggesting that type IV systems could be highly diverged type I or type III derivatives.

The class 2 effector modules are single large proteins, with domain architectures clearly differentiating type II, V and VI systems (Box 1; and see discussion below)[20]. The types and subtypes within class 2 substantially differ with respect to the mechanisms of pre-crRNA processing[51-54]. In type VI and subtype V-A systems, the large effector protein also encompasses the pre-crRNA processing RNase activity[55-57], whereas in type II and several type V subtypes, the processing activity is typically relegated to a non-Cas enzyme, RNase III. In the latter cases, the effector module includes an additional RNA molecule, the trans-activating CRISPR (tracr) RNA that forms stable duplexes with the partially complementary direct repeat of the pre-crRNA. After cleavage of the RNA duplex by RNase III, the mature guide RNA, that is, the crRNA–tracrRNA complex, remains stably bound to the effectors, allowing for specific DNA interference[51-54].

The set of Cas1 to Cas13 proteins that comprise the adaptation and effector modules define the types and subtypes, and thus represent the core of the CRISPR–Cas systems. This core is accompanied by numerous ancillary proteins that are more loosely associated with CRISPR–Cas. The repertoire of the ancillary genes has recently drastically expanded, in large part, through the use of dedicated computational protocols for systematic detection of CRISPR-linked genes[31,32].

We discuss the ancillary genes in a later section, after describing the current state of the CRISPR–Cas classification.

In addition to the distinctions between the *cas* gene composition and the sequences and structures of Cas proteins, the types and subtypes of CRISPR–Cas systems can be, to some extent, differentiated by the distinct sequence and structural features of the repeats themselves[58,59]. However, the correspondence is incomplete such that several branches in the cluster dendrogram of CRISPR–Cas collate multiple subtypes[59].

[H1] CRISPR–Cas classification

[H2] Class 1 and its derivatives

The classification of class 1 CRISPR–Cas systems, which include types I, III and IV, has remained relatively stable compared to the 2015 version[18] (Figure 1). The 2015 class 1 classification scheme included 12 subtypes that can be distinguished by sequence similarity clustering of effector proteins, as well as by comparison of loci organizations and the sequences of the repeats. In the updated scheme, four subtypes were added— subtypes III-E, III-F, IV-B and IV-C. In addition, given the experimental demonstration of new spacer incorporation by subtype I-U systems[60], this subtype was reclassified as subtype I-G.

Subtype III-E, identified in 14 contigs from the NCBI non-redundant nucleotide sequence database that appear to come from 8 bacterial species (Supplementary Figure 1, Supplementary Table 2, Supplementary Dataset 5), is characterized by a unique fusion of several Cas7 proteins and a putative Csm2-like small subunit (Cas11), such that the crRNA-binding part of the effector module is compressed within a single, large, multidomain protein. In this respect, subtype III-E resembles class 2 CRISPR–Cas systems although the domain composition and sequence analysis unequivocally place it within type III of class 1, and moreover, a specific relationship with subtype III-D could be traced (Supplementary Figure 1). The multidomain subtype III-E effector is predicted to cleave pre-crRNA given the conservation of aspartate residues that are known to be involved in RNA cleavage in the homologous Csm4 protein, and might also contribute to target RNA cleavage (Supplementary Figure 2). The subtype III-E loci often include a putative ancillary gene encoding a large protein that contains a CHAT domain, a caspase family protease that is, typically, involved in programmed cell death[61], fused to tetratricopeptide (TPR) repeats

(Figure 1). The presence of this ancillary protein suggests sub- or neofunctionalization of subtype III-E systems and their potential involvement in complex defence pathways (Figure 1, Supplementary Figure 1).

The subtype III-F systems have been identified previously[62] and included in the 2015 CRISPR–Cas census[18] but were not classified as a distinct subtype because their number was too small. Now that this variant was found in 12 additional genomes, it became apparent that they qualify as a separate subtype (Supplementary Figure 3, Supplementary Table 2). The Cas7 and Cas5 subunits as well as the large subunit of the subtype III-F effector complex show a distant but substantial imilarity with the corresponding components of other type III subtypes, whereas the putative small subunit does not show any similarity to Cas11. Unlike all other type III systems, subtype III-F contains only one Cas7-like protein. The HD domain fused to the Cas10-like large subunit retains all catalytic residues and therefore is predicted to cleave the target DNA. However, the cyclase or polymerase domain of the Cas10-like subunit is inactivated as indicated by amino acid substitutions in the catalytic site, and furthermore, the subtype III-F loci lack any genes encoding CARF (CRISPR-associated Rossmann fold) domain proteins. Thus, this type III subtype clearly does not function via cyclic oligoA signaling as shown for subtype III-A and implied for the rest of type III systems containing an active Cas10 polymerase[36,37,43].

In the 2015 classification, subtype IV-B was reported as a variant that, unlike subtype IV-A systems, lacks the *dinG* gene but contains a distinct version of the predicted small subunit of the effector complex; furthermore, most of the subtype IV-B loci encompass the ancillary gene *cysH*[32]. These systems have been discovered on plasmids from numerous, diverse bacteria[30] and, accordingly, the variant was upgraded to a subtype. Subtype IV-C loci were also detected in the 2015 census but were not formally classified. Now, this type IV variant was identified in 9 contigs, mostly, from thermophilic microorganisms (Supplementary Figure 4, Supplementary Table 2), and its classification as a distinct subtype also appears justified. The Cas7 and Cas5 homologs of the subtype IV-C systems show statistically significant similarity to the corresponding proteins of subtype IV-A and IV-B systems, whereas the putative large and small subunits of the effector complex lack any detectable similarity with the counterparts from any other CRISPR–Cas system. Notably, unlike in the other two type IV subtypes, the putative large subunit of subtype IV-C contains an HD nuclease domain, suggesting that it cleaves the target

DNA. The order of the HD nuclease motifs is the same as in Cas3 in type I systems but different from that in the HD nuclease domains fused to Cas10 in most of the type III systems, apparently, as a result of a circular permutation occurring during the evolution of the CRISPR-associated HD nucleases.

Several additional, distinct variants of class 1 could become subtypes when more taxonomic diversity and/or more structural and experimental data become available. Among such cases, there is a distinct type III system variant found in archaea of the order *Sulfolobales* represented by the locus YN1551_RS11700..YN1551_RS11720 from *Sulfolobus islandicus* (Supplementary dataset 1). This variant features extremely diverged Cas10 and Cas5 homologs, and a unique, uncharacterized predicted component of the effector complex, Csx26. Another distinct type III system, so far found only in the archaeon *Ignisphaera aggregans* (locus Igag_0607..Igag_0623), includes several proteins that are not similar to any known Cas or ancillary proteins.

A variety of derived, apparently defective variants of type I systems have been discovered, such as, for example, 'minimal' subtype I-F and subtype I-B systems that are encoded by distinct families of Tn7-like transposons[30,33]. These variants lack the helicase-nuclease Cas3 that is required for interference[63] and therefore are predicted to perform functions distinct from adaptive immunity. A hypothesis has been proposed that these minimal type I variants mediate guide-RNA-dependent transposition[30,33], and recently, such an activity has been demonstrated experimentally[35]. Defective CRISPR–Cas systems also have been reported in preliminary studies to be encoded by some of the recently discovered giant phages, where their roles remain to be deciphered[64]. An analogous interference-deficient derivative of subtype I-E CRISPR–Cas systems was detected in the genomes of many bacteria of the genus *Streptomyces*[32]. This variant is not associated with any detectable mobile genetic elements but is tightly linked to a gene encoding a STAND superfamily NTPase[65], suggesting involvement of these interference-deficient CRISPR–Cas systems in signal transduction, and possibly, dormancy induction or programmed cell death. The differences in the Cas protein compositions between these minimal CRISPR–Cas variants and fully functional type I systems potentially could be used as an argument for classification of the defective variants into separate subtypes. However, Cas protein sequence comparison and phylogenetic analysis unequivocally demonstrate the origin of these variants from subtypes I-F, I-E, and I-B, respectively[30,32,33]. Therefore, we propose to keep them

within the respective subtypes as distinct variants, denoted, for example, I-F1, I-F2, etc. (Figure 1).

Apart from the newly identified subtype IV-C, most of the type IV systems also are defective CRISPR–Cas forms that lack the nucleases involved in the target cleavage and thus resemble the transposon-encoded variants with respect to organization and, perhaps, functionality. Indeed, the distinctive biological feature of type IV systems is their apparent (nearly) exclusive localization on plasmids, integrated conjugating elements and prophages [30]. Furthermore, preliminary data suggests multiple spacers targeting heterologous plasmids have been detected in type IV CRISPR arrays, suggesting that one of the functions of type IV systems is inter-plasmid competition[66].

Some derived variants are so distant from the canonical organization that their status as CRISPR–Cas systems appears questionable. A case in point is a recently described locus found in many Haloarchaea that only retain highly divergent forms of Cas5 and Cas7 (Haloarchaeal RAMPs; HRAMPs) along with an uncharacterized conserved protein and various nucleases[67] (Supplementary Figure 5A). The search of Asgard archaea genomes[68] performed in the course of this work also revealed highly derived CRISPR–Cas variants that resemble HRAMPs in terms of Cas protein composition and encompass an unusual large protein containing a diverged Cas1 domain, along with distinct variants of Cas5 (a fusion with an HD nuclease) and Cas7 as well as additional nucleases (Supplementary Figure 5B). The functions of these extremely derived systems are unknown, and given the lack of adjacent CRISPR arrays, it is not even clear whether their activity is guide RNA-dependent. If these systems are shown to function via a CRISPR–Cas-like mechanism, they might qualify as distinct types, given the drastic reduction of the Cas protein repertoire.

Thus, the formation of derived variants that lack the interference capacity and are likely to perform functions distinct from adaptive immunity is a pervasive trend in the evolution of CRISPR–Cas. Additional highly divergent CRISPR–Cas derivatives are likely to be discovered, and their experimental characterization is likely to become a major research direction.

[H2] The expanding class 2

Class 2 CRISPR–Cas systems include types II, V and VI. The distinguishing feature of these types is that their effector complexes consist of a single, large, multidomain protein, such as Cas9 in type II. Thanks to the focused efforts on computational discovery of new class 2 systems, partly, as a quest for potential new genome editing tools, this class underwent a drastic expansion since the 2015 classification[11,20-23]. From 2 types and 4 subtypes in 2015, class 2 expanded to 3 types and 18 subtypes (Figure 2). The new discoveries include multiple, diverse variants of type V as well as type VI systems, the first and so far the only variety of CRISPR–Cas systems that exclusively cleaves RNA.

Type V systems fundamentally differ from type II by the domain architecture of the effector proteins. The type II effectors (Cas9) contain two nuclease domains that are each responsible for the cleavage of one strand of the target DNA, with the HNH nuclease inserted inside the RuvC-like nuclease domain sequence[51]. By contrast, the type V effectors (Cas12) only contain the RuvC-like domain that cleaves both strands[69,70]. Type VI effectors (Cas13) are unrelated to the effectors of type II and V systems, contain two HEPN domains and apparently target transcripts of invading DNA genomes. Cas13 proteins also display collateral, non-specific RNase activity that is triggered by target recognition and induces dormancy in virus-infected bacteria[71].

Assignment of subtypes within type II, V and VI systems is a challenge because of the uniform domain architecture of the respective effector proteins. The current practice (which, admittedly, involves a degree of arbitrariness) is to call a new subtype for variants that do not show statistically significant sequence similarity to any of the already established subtypes in BLAST searches[39]; the presence of additional accessory genes is also taken into consideration. This approach has so far resulted in the identification of 3 subtypes of type II systems, 10 subtypes of type V systems and 4 subtypes of type VI systems with typical, large effector proteins (Figure 2).

In addition, a heterogeneous assemblage of putative type V variants with smaller RuvC-like domain containing proteins, provisionally classified as subtype V-U, has been discovered[20] (Supplementary Figure 6). The putative subtype V-U effectors show high sequence similarity to TnpB proteins (predicted RuvC-like nucleases) encoded by IS605-like transposons and are thought to be intermediates on the evolutionary path from TnpB to fully-fledged type V effectors. CRISPR–Cas systems that evolved from different groups of TnpB on multiple, independent occasions as shown by phylogenetic analysis of the TnpB family[20]. Recently, the

interference activity of four subtype V-U effectors was validated experimentally, and as a result, one of these variants has been upgraded to a separate subtype, V-F [22,23]. Notably, these newly characterized CRISPR–Cas variants show major differences in the interference specificity compared to the previously characterized type V effector and to one another. The subtype V-F effector, Cas12f (originally denoted Cas14), has been shown to cleave single-stranded DNA(ssDNA)[22] although, subsequently, a double-stranded DNA cleavage activity has been reported in a preliminary study as well[72], whereas Cas12g is an RNA-guided RNase that also possesses collateral RNase and ssDNase activities[23]. These findings emphasize the remarkable functional diversity of CRISPR–Cas systems that remains to be fully characterized through the discovery and study of new subtypes. Different variants within subtype V-F (currently, variant V-F1-3) appear to originate from different groups of *tnpB* genes as indicated by the phylogenetic analysis of the TnpB family[20] (Supplementary Dataset 4). Nevertheless, given the highly significant sequence similarity between these effector proteins, they are all currently classified within a single subtype.

One of the former V-U variants, V-U5, contains an apparently inactivated RuvC-like nuclease domain as indicated by the replacement of essential catalytic residues, and is encoded by cyanobacterial Tn7-like transposons[30]. The prediction that this variant evolved to function in transposons analogously to the defective type I systems, that is, by mediating guide RNA-dependent transposition, has been recently experimentally validated (and the subtype has been accordingly upgraded to subtype V-K)[34].

It is expected that the remaining subtype V-U variants will be classified into the already created or additional subtypes as they are experimentally characterized. Furthermore, in all likelihood, multiple subtypes of type V systems that independently originated from TnpB nucleases remain to be discovered, and consequently, the number of recognized subtypes will grow further.

The origin of type VI systems is much less clear than the derivation of type V systems from TnpB. The HEPN RNase domain is widespread in various defence systems, in particular, as the toxin components of numerous toxin-antitoxin modules, which are likely to be ultimate ancestors of CRISPR-associated HEPN domains[7,73]. Given that the presence of two HEPN domains is a unique signature of type VI effectors (Cas13), it is appealing to surmise that these effectors evolved from a common ancestor after duplication of the HEPN domain. However, the two

HEPN domains in each of the Cas13 proteins are only distantly related to each other, and phylogenetic analysis results appear not to be compatible with the duplication scenario (Supplementary Figure 7). In the phylogenetic tree of the HEPN family, the amino-terminal and C-terminal HEPN domains form distinct branches, pointing to a common ancestor with two HEPN domains. This ancestral *cas13* gene might have evolved by recombination between two genes encoding distinct HEPN-containing proteins, and possibly, a distinct family of toxin components of abortive infection modules[73]. Type VI systems appear to be far less diverse than type V systems, but discovery of new subtypes remains possible. For example, we identified a distinct variant of type VI system in *Brachyspira* species with a two-HEPN effector that shows no significant similarity to the Cas13 sequences from the four current subtypes (Supplementary Figure 8). Presently, we refrain from calling it a new subtype because of its narrow spread in bacteria but, as the genomic database grows, this will be a strong candidate.

[H1] A bipartite gene-sharing network

In addition to the classification approaches outlined above, we performed a quantitative analysis of a bipartite network in which CRISPR–*cas* loci are connected through shared genes (Supplementary Figure 9). To identify clusters of tightly connected loci that share overlapping gene sets, we applied a previously described consensus clustering approach that combines bipartite modularity maximization and hierarchical clustering, followed by significance-based filtering of the results[40]. By highlighting distinct sets of genes and loci that are mutually associated, identification of modules in the gene-sharing network could contribute both to CRISPR–Cas classification and to functional prediction.

Altogether, 126 modules were identified in the CRISPR–Cas network that can be roughly assigned to four categories: modules sharing distinct ancillary gene sets (category 1); derived variants characteristic of specific bacterial or archaeal lineages (category 2); mixed modules that apparently result from recombinational shuffling among CRISPR–*cas* loci that typically share closely related adaptation genes but have distinct effector genes (category 3); and modules that lack any of the above distinctive features but include highly diverged Cas proteins (category 4) (Supplementary Figure 9, Supplementary Dataset 3). The recently characterized minimal variant of subtype I-F (I-F3) associated with Tn7-like transposons, a remarkable case of CRISPR–Cas neofunctionalization (module 16), is an example from category 1. Cyanobacteria-specific

modules 65 and 98, that consist of distinct variants of subtype III-B exemplify category 2. A case of previously described gene shuffling in *Methanosarcina* species [62,74] is captured in module 84 which belongs in category 3. Most of the identified modules include CRISPR–*cas* loci that belong to the same subtype. The exceptions are modules that combine two or three subtypes of type I (modules 10 and 101) or type V systems (module 126) that share overlapping gene compositions. More notably, three modules (46, 93 and 108) join loci of types I and III systems, apparently, reflecting recombinational events. Only a few relatively rare, low-abundance subtypes are represented by a single module. Most of the subtypes are divided into multiple modules, with subtypes I-E and I-B showing the highest heterogeneity (14 and 13 modules, respectively). This reflects the functional and evolutionary plasticity of these subtypes that, conceivably, underlie their high abundance in current genomic databases (see below).

The fine-grained modules produced by bipartite network analysis could be useful for identification of distinct functional variants of CRISPR–Cas systems that might be obscured by the conservative assignment of subtypes and variants. Moreover, this approach could provide a fast and straightforward way to assign new CRISPR–*cas* loci to pre-defined types and subtypes for which related loci have been already identified. In support of this possibility, the present bipartite network analysis was able to correctly assign most of the incomplete CRISPR–*cas* loci to the types and subtypes where they belong. To delineate coarse-grained modules that would facilitate classification of novel CRISPR–Cas systems in an unsupervised way, more sophisticated multi-resolution approaches will be required.

[H1] Distribution of CRISPR–Cas systems

The CRISPR–Cas systems are non-uniformly distributed among bacterial and archaeal phyla. We present a census of CRISPR–*cas* loci in the current collection of complete bacterial and archaeal genomes. Analysis of 13,116 complete genomes showed that CRISPR–*cas* loci are represented in a substantial majority of archaea (276 of 324 genomes (85.2%)), including almost all hyperthermophiles (89 of 92 genomes (96.7%)), but only in ~40% of bacteria (5,412 of 12,792 genomes (42.3%)) (Figure 3and Supplementary Dataset 6). Clear trends are observed in the distribution of specific CRISPR–Cas classes, types and subtypes. In particular, class 2 is still exclusive to bacteria. The absence of class 2 in archaea, at least, in part, can be explained by the absence of RNase III, the pan-bacterial enzyme that is responsible for pre-crRNA processing in

type II and some subtypes of type V systems, that is, most of the class 2 systems[46,75]. By contrast, the genomes of Crenarchaeota are substantially enriched for type III systems of class 1. Overall, and in most groups of bacteria and archaea, class 1 is far more abundant than class 2. However, there are notable exceptions, for example, Tenericutes bacteria, in which only class 2 systems have been identified so far (Figure 3). Some groups of bacteria, such as *Chlamydia* species (Figure 3) or the recently discovered Candidate Phyla Radiation, that appears to consist, mostly, of symbiotic microorganisms are nearly devoid of CRISPR–Cas systems[76-78]. Conversely, the majority of type VI systems, and in particular, all instances of the most abundant subtype VI-B, have been identified in bacterial genomes of the phyla Bacteroidetes and Fusobacteria (Figure 3).

The biological underpinnings of the non-uniform phyletic spread of CRISPR–Cas systems remain to be elucidated. Considering the high horizontal mobility of CRISPR–*cas* loci, it appears likely that their loss or retention in prokaryotic genomes depends on the trade-off between the fitness cost that is determined, mostly, by auto-immunity and curtailment of horizontal gene transfer, and the benefits of defence conferred by adaptive immunity[79-84]. These benefits, most likely, depend on the abundance and diversity of viruses in specific habitats as well as the biology of host–parasite interactions in specific groups of microorganisms[85,86]. The evolutionary dynamics that determines the distribution of CRISPR–Cas among bacteria and archaea can be expected to become one of the major directions in CRISPR–Cas research in the next few years. In particular, these dynamics might depend, to a large extent, on the interactions between CRISPR–Cas and DNA repair mechanisms, such as the double-strand break repair systems [87].

[H1] Core and ancillary *cas* genes

The components of the adaptation and effector modules comprise the suite of core Cas proteins. The core Cas proteins in the widespread CRISPR–Cas types and subtypes are well characterized, although the discovery of novel class 2 effector proteins continues to gradually expand the core gene repertoire. Furthermore, in the course of the systematic search for new CRISPR-linked proteins, many highly diverged variants of the core proteins have been identified [32].

By contrast, the list of the (predicted) ancillary CRISPR-linked proteins has greatly expanded as a result of dedicated searches of CRISPR–Cas genomic neighborhoods[31,32]. For the great majority of these proteins, no experimental data are available yet, but computational analysis of

their domain architectures points to multiple connections to various signal transduction pathways as well as membrane association or functional links to membrane transport processes for many CRISPR–Cas systems, particularly, those of type III systems that drastically stand out in the complexity of the gene repertoire among all CRISPR–Cas forms (Figure 4 A,B). Several accessory proteins, for example, those in subtypes VI-B and VI-D, have been directly shown to modulate the activity of the respective effectors[25,26]. Furthermore, some of the genes that are currently classified as ancillary are actually represented in numerous CRISPR–Cas systems and could perform major roles in the immune response. The most obvious example is Csm6, a HEPN-domain RNase that is a component of the majority of subtype III-A CRISPR–Cas systems and is activated by the signal transduction pathway initiated by cyclic oligoA produced by the Cas10 polymerase[30,31]. Systematic experimental characterization of the roles of accessory proteins in CRISPR–Cas functions, undoubtedly, will be another key research area in the study of CRISPR–Cas biology for years to come.

The discovery of new class 2 subtypes and numerous accessory proteins poses obvious problems for the systematic nomenclature of CRISPR-linked genes. So far, a conservative approach has been adopted under which the *cas* designation is reserved for core genes, or more precisely, families of homologous core genes (Supplementary Table 1). The numbered *cas* gene names were originally assigned to the 11 most common genes among diverse CRISPR–Cas systems, and subsequently, *cas12* and *cas13*, the effectors of type V and type VI systems, respectively, have been added. Currently, the *cas* names are reserved for type-specific effector genes, whereas subtypes are specified by suffixes, for example, *cas12a*, *cas12b*, *cas12c*, etc. The recent designation of small type V effector proteins related to those in subtype V-U systems as Cas14 (REF. [22]) does not conform with this criterion. We believe that the appropriate name for these proteins should be Cas12f (REF. [23]), given that Cas12 is supposed to apply to all type V system effectors. Obviously, under this approach, the number of *cas* genes cannot be expected to substantially increase because both discovery of new types and identification of new core genes for already established types are rare. The ancillary genes remain to be known under their legacy names or as *csx* followed by a number, although a systematic nomenclature might be considered in the future.

[H1] Origins and evolution of CRISPR–Cas

Comparative analysis of CRISPR–Cas systems, in particular, the newly discovered class 2 subtypes, provides for the reconstruction, at least, in outline, of a nearly complete scenario of CRISPR–Cas evolution (Figure 5). A striking feature of the evolutionary history of CRISPR–Cas is the repeated recruitment of genes from different mobile genetic elements for various functions in adaptive immunity[7,29]. Thus, the adaptation module, along with the CRISPR repeats themselves, appears to originate from an immobilized transposon of the casposon family, so named because these elements employ a Cas1 homolog as the transposase[88-90]. The casposon could have contributed not only *cas1* but also the *cas4* gene encoding another nuclease that is involved in PAM selection during adaptation in many CRISPR–Cas systems[60,91-93], given that Cas4 homologs are among the cargo genes in some casposons.

The effector module of type III systems appear to be the best candidate for the ancestral state, given their widespread (especially in archaea), complex gene composition and the fact that, in most of the type III variants, the large subunit of the effector complex (Cas10) is an active enzyme, a cyclic oligoA polymerase[7]. The effector moiety of CRISPR–Cas could have started as a putative signaling system that has been identified in several bacteria and consists of a small-sized, 'minimal' Cas10 homolog and a homolog of Csm6 with fused CARF and HEPN domains[7,94] (Figure 5). This system is predicted to function analogously to the signal transduction pathway in type III CRISPR–Cas systems, namely, by synthesizing cyclic oligoA (most likely, in response to stress) that is then bound by the CARF domain and allosterically activates the RNase activity of the HEPN domain[36,37]. The indiscriminate RNA cleavage by the HEPN domain would induce dormancy or programmed cell death. The putative ancestral system remains to be studied experimentally but, even without such validation, it resembles an abortive infection (Abi) module. Indeed, recently, the HEPN-containing Csm6 protein of subtype III-A systems has been shown to act as a toxin causing growth arrest of the host cell[95], which is compatible with the origin of the type III effector module from an Abi system. Similarly to the known Abis[10,96], the ancestor of the effector module is likely to be subject to extensive horizontal gene transfer and might, effectively, possess features of a mobile genetic element.

Thus, different types of mobile genetic elements seem to have given rise to both the adaptation and the effector parts of class 1 CRISPR–Cas systems. The subsequent evolution of the effector module would have involved serial duplication of the RRM domain of the Cas10 homolog and

capture of additional proteins, in particular, the target-cleaving HD nuclease[7]. The key event in the evolution of type I systems was the capture of the helicase-nuclease Cas3 and the replacement of the oligoA polymerase Cas10 with the enzymatically inactive Cas8 as the large subunit of the effector complex. Whether the latter event involved extreme divergence following inactivation of Cas10 or capture of an unrelated protein, remains uncertain.

The origin of type IV systems remains uncertain but the recent discovery of subtype IV-C systems, with the large subunit fused to an HD domain, together with the observation that both Cas5 and Cas7 components of type IV systems share a greater sequence similarity with the counterparts from type III than with those from type I systems, suggest that type IV could have evolved from type III. These observations are compatible with the lack of association of the IV-C systems with any known mobile genetic elements. Similar lines of evidence could point to subtype I-D systems as a potential evolutionary intermediate between type III and type I systems. The structure of both the subtype I-D effector complex and the Cas10d protein should shed more light on the origin of type I systems. The origin of the HRAMP system, a highly derived CRISPR-less class 1 variant is unclear as well, but both Cas5 and Cas7 components are more similar to the respective proteins of type III than to those of type I systems, suggesting a route of evolution parallel to that of type IV systems[67].

In class 2, the effectors of different types and subtypes of types V and, possibly, type II systems appear to have evolved, on multiple, independent occasions, from TnpB nucleases encoded by yet another class of mobile genetic element, the IS605-like transposons [20]. Type II systems apparently evolved from a distinct variety of TnpB (denoted IscB) that contains an HNH nuclease domain inserted into the RuvC-like domain[97]. The type VI system effectors (Cas13) seem to originate from HEPN-containing components of an Abi module[7,20] (Figure 5). The functional analogy between Cas13a and Abi has been recently validated by experiments that demonstrated growth arrest of phage-infected bacteria dependent on Cas13a activity[71]. A recurrent trend in the evolution of CRISPR–Cas effectors is the accretion of additional proteins (in class 1) or domains (in class 2), on top of the core nuclease domains, providing for the flexibility required to accommodate the crRNA and the target DNA or RNA[7].

Another general trend in CRISPR–Cas evolution is the spawning of defective variants, many of which are appropriated by mobile genetic elements[30,33]. The defective forms of CRISPR–Cas

systems are predicted to perform various functions that require target recognition but not cleavage. A striking case of such functionality is the crRNA-dependent, site-specific transposition that has been recently demonstrated for the transposon-encoded derived variants of subtype I-F and subtype V-K systems[34,35].

[H1] Concluding remarks

Because the most abundant types and subtypes of CRISPR–Cas systems are now known, the overall structure of the current classification is likely to stand the test of time. However, the discovery of comparatively rare but functionally and evolutionarily interesting and informative variants has not stopped and, in all likelihood, will continue, especially, as diverse environments are explored by methods of metagenomics and single cell genomics. Some of these variants are distinct enough to become new subtypes but so far, no new types have been identified after the discovery of type VI. According to the currently adopted criteria, to qualify as a new type, a CRISPR–Cas variant has to encompass an effector module unrelated (or extremely distantly related) to those of the known types. Other types might remain to be discovered, but it is becoming increasingly clear that, if such additional types exist, they are rare and/or highly specialized. Investigation of the numerous ancillary components of CRISPR–Cas is starting to uncover multiple connections between CRISPR–Cas and various functionally distinct systems of bacterial and archaeal cells, particularly, those involved in different forms of signal transduction.

In summary, the diversity of the identified CRISPR–Cas systems has substantially increased over the last four years thanks to a combination of computational and experimental approaches. Notably, the new varieties could be classified into distinct types and subtypes by using several criteria. Arguably, this granularity stems from punctuated evolution whereby diversification of emerging subtypes slows down after an initial period of rapid innovation. Notwithstanding the apparent distinctness of the subtypes, the increasing diversity of CRISPR–Cas creates further challenges to classification and nomenclature, and calls for the development of robust classification criteria. Delineation of types and, to a large extent, subtypes will likely remain qualitative, given the paucity of shared components. However, classification of variants within subtypes, some of which might qualify as separate subtypes, can be quantified, for example, by using bipartite network analysis as shown here. On the whole, we believe that the classification

of CRISPR–Cas systems has entered the era of consolidation and refinement. Experimental characterization of CRISPR–Cas functions still lags behind predictions produced by computational analysis. It is our hope that the updated classification will facilitate experimental studies and promote new directions.

References

1 Komor, A. C., Badran, A. H. & Liu, D. R. CRISPR-Based Technologies for the Manipulation of Eukaryotic Genomes. *Cell* **168**, 20-36, doi:S0092-8674(17)30417-8 [pii]10.1016/j.cell.2017.04.005 (2017).

2 Pickar-Oliver, A. & Gersbach, C. A. The next generation of CRISPR-Cas technologies and applications. *Nat Rev Mol Cell Biol* **20**, 490-507, doi:10.1038/s41580-019-0131-5 (2019).

3 Mohanraju, P. *et al.* Diverse evolutionary roots and mechanistic variations of the CRISPR-Cas systems. *Science* **353**, aad5147, doi:10.1126/science.aad5147 aad5147 [pii] 353/6299/aad5147 [pii] (2016).

4 Jackson, S. A. *et al.* CRISPR-Cas: Adapting to change. *Science* **356**, doi:eaal5056 [pii] 10.1126/science.aal5056 356/6333/eaal5056 [pii] (2017).

5 Barrangou, R. & Horvath, P. A decade of discovery: CRISPR functions and applications. *Nat Microbiol* **2**, 17092, doi:10.1038/nmicrobiol.2017.92 nmicrobiol201792 [pii] (2017).

6 Jiang, F. & Doudna, J. A. CRISPR-Cas9 Structures and Mechanisms. *Annu Rev Biophys* **46**, 505-529, doi:10.1146/annurev-biophys-062215-010822 (2017).

7 Koonin, E. V. & Makarova, K. S. Origins and evolution of CRISPR-Cas systems. *Philos Trans R Soc Lond B Biol Sci* **374**, 20180087, doi:10.1098/rstb.2018.0087 (2019).

8 Faure, G., Makarova, K. S. & Koonin, E. V. CRISPR-Cas: complex functional networks and multiple roles beyond adaptive immunity. *J. Mol. Biol.* 431, 3-20. doi: 10.1016/j.jmb.2018.08.030(2019).

9 McGinn, J. & Marraffini, L. A. Molecular mechanisms of CRISPR-Cas spacer acquisition. *Nat Rev Microbiol* **17**, 7-12, doi:10.1038/s41579-018-0071-7 (2019).

10 Koonin, E. V., Makarova, K. S. & Wolf, Y. I. Evolutionary Genomics of Defense Systems in Archaea and Bacteria. *Annu Rev Microbiol* **71**, 233-261, doi:10.1146/annurev-micro-090816-093830 (2017).

11 Koonin, E. V., Makarova, K. S. & Zhang, F. Diversity, classification and evolution of CRISPR-Cas systems. *Curr Opin Microbiol* **37**, 67-78, doi:S1369-5274(17)30023-1 [pii] 10.1016/j.mib.2017.05.008 (2017).

12 Ishino, Y., Krupovic, M. & Forterre, P. History of CRISPR-Cas from Encounter with a Mysterious Repeated Sequence to Genome Editing Technology. *J Bacteriol* **200**, doi:10.1128/JB.00580-17 (2018).

13 Hille, F. & Charpentier, E. CRISPR-Cas: biology, mechanisms and relevance. *Philos Trans R Soc Lond B Biol Sci* **371**, doi:10.1098/rstb.2015.0496 (2016).

14 Wright, A. V., Nunez, J. K. & Doudna, J. A. Biology and Applications of CRISPR Systems: Harnessing Nature's Toolbox for Genome Engineering. *Cell* **164**, 29-44, doi:10.1016/j.cell.2015.12.035 S0092-8674(15)01699-2 [pii] (2016).

15 Klompe, S. E. & Sternberg, S. H. Harnessing "A Billion Years of Experimentation": The Ongoing Exploration and Exploitation of CRISPR-Cas Immune Systems. *CRISPR J* **1**, 141-158, doi:10.1089/crispr.2018.0012 (2018).

16 Makarova, K. S., Wolf, Y. I. & Koonin, E. V. Classification and Nomenclature of CRISPR-Cas Systems: Where from Here? *CRISPR J* **1,** 325-336, doi: 10.1089/crispr.2018.0033. (2018).

17 Makarova, K. S. *et al.* Evolution and classification of the CRISPR-Cas systems. *Nat Rev Microbiol* **9**, 467-477, doi:10.1038/nrmicro2577 nrmicro2577 [pii] (2011).

18 Makarova, K. S. *et al.* An updated evolutionary classification of CRISPR-Cas systems. *Nat Rev Microbiol* **13**, 722-736, doi:10.1038/nrmicro3569 nrmicro3569 [pii] (2015).

19 Shmakov, S. *et al.* Discovery and Functional Characterization of Diverse Class 2 CRISPR-Cas Systems. *Mol Cell* **60**, 385-397, doi:10.1016/j.molcel.2015.10.008 S1097-2765(15)00775-3 [pii] (2015).

20 Shmakov, S. *et al.* Diversity and evolution of class 2 CRISPR-Cas systems. *Nat Rev Microbiol*, doi:10.1038/nrmicro.2016.184 nrmicro.2016.184 [pii] (2017).

This work demonstrates the relationships between the effectors of different types and subtypes of class 2 CRISPR–Cas systems and nucleases encoded by mobile genetic elements. On the basis of sequence comparison and phylogenetic analysis of Cas12 (type V effectors) and TnpB nucleases encoded by transposons, a scenario of independent recruitment of distinct TnpB variants giving rise to different type V subtypes is proposed.

21 Burstein, D. *et al.* New CRISPR-Cas systems from uncultivated microbes. *Nature* **542**, 237-241, doi:10.1038/nature21059 nature21059 [pii] (2017).

This work describes the metagenomic discovery of two new subtypes of type V CRISPR–Cas systems and experimental validation of their activity.

22 Harrington, L. B. *et al.* Programmed DNA destruction by miniature CRISPR-Cas14 enzymes. *Science* **362**, 839-842, doi:10.1126/science.aav4294 (2018).

This work experimentally validates the enzymatic activity of small predicted effectors that have been assigned to subtype V-U in Ref. 20 and are here re-classified as subtype V-F. It is shown that these enzymes substantially differ from the previously characterized large type II and type V effectors, and catalyze both crRNA-specific and non-specific cleavage of single-stranded DNA.

23 Yan, W. X. *et al.* Functionally diverse type V CRISPR-Cas systems. *Science* **eeav7271** (2018).

This paper reports the experimental characterization of CRISPR–Cas subtypes V-C, V-G, V-H and V-I. Whereas Cas12c, Cas12h, and Cas12i proteins all demonstrate RNA-guided double-stranded DNA interference similarly to previously described CRISPR–Cas effectors, Cas12g is shown to function as an RNase with collateral RNase and single-strand DNase activities.

24 Abudayyeh, O. O. *et al.* C2c2 is a single-component programmable RNA-guided RNA-targeting CRISPR effector. *Science* **353**, aaf5573, doi:10.1126/science.aaf5573 aaf5573 [pii] science.aaf5573 [pii] (2016).

25 Smargon, A. A. *et al.* Cas13b Is a Type VI-B CRISPR-Associated RNA-Guided RNase Differentially Regulated by Accessory Proteins Csx27 and Csx28. *Mol Cell*, doi:S1097-2765(16)30866-8 [pii] 10.1016/j.molcel.2016.12.023 (2017).

26 Yan, W. X. *et al.* Cas13d Is a Compact RNA-Targeting Type VI CRISPR Effector Positively Modulated by a WYL-Domain-Containing Accessory Protein. *Mol Cell*, doi:S1097-2765(18)30173-4 [pii] 10.1016/j.molcel.2018.02.028 (2018).

This paper demonstrates RNA-targeting by the smallest known type VI effector Cas13d and shows that the accessory WYL-domain-containing protein stimulates this activity.

27 Murugan, K., Babu, K., Sundaresan, R., Rajan, R. & Sashital, D. G. The Revolution Continues: Newly Discovered Systems Expand the CRISPR-Cas Toolkit. *Mol Cell* **68**, 15-25, doi:10.1016/j.molcel.2017.09.007 (2017).

28 Stella, S., Alcon, P. & Montoya, G. Class 2 CRISPR-Cas RNA-guided endonucleases: Swiss Army knives of genome editing. *Nat Struct Mol Biol* **24**, 882-892, doi:10.1038/nsmb.3486 (2017).

29 Koonin, E. V. & Makarova, K. S. Mobile Genetic Elements and Evolution of CRISPR-Cas Systems: All the Way There and Back. *Genome Biol Evol* **9**, 2812-2825, doi:10.1093/gbe/evx192 4161385 [pii] (2017).

30 Faure, G. *et al.* CRISPR–Cas in mobile genetic elements: counter-defense and beyond *Nature Rev Microbiol* 17, 513-525, doi: 10.1038/s41579-019-0204-7 (2019).

31 Shah, S. A. *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*, 1-13, doi:10.1080/15476286.2018.1483685 (2018).

Along with Ref. 32, this paper describes a computational approach to predict proteins that are functionally linked to CRISPR–Cas systems and applies this approach to type III systems.

32 Shmakov, S. A., Makarova, K. S., Wolf, Y. I., Severinov, K. V. & Koonin, E. V. Systematic prediction of genes functionally linked to CRISPR-Cas systems by gene neighborhood analysis. *Proc Natl Acad Sci U S A* **115**, E5307-E5316, doi:10.1073/pnas.1803440115 1803440115 [pii] (2018).

Along with Ref. 31, this paper describes a computational approach for systematic prediction of proteins that are functionally linked to CRISPR–Cas systems ('CRISPRicity' protocol) and applies it to all CRISPR–Cas types and subtypes.

33 Peters, J. E., Makarova, K. S., Shmakov, S. & Koonin, E. V. Recruitment of CRISPR-Cas systems by Tn7-like transposons. *Proc Natl Acad Sci U S A* **114**, E7358-E7366, doi:10.1073/pnas.1709035114 1709035114 [pii] (2017).

This paper describes, for the first time, defective CRISPR–Cas systems encoded in Tn7-like transposons and predicts their function in RNA-guided transposition.

34 Strecker, J. *et al.* RNA-guided DNA insertion with CRISPR-associated transposases. *Science*, doi:10.1126/science.aax9181 (2019).

This work validates the prediction made in Ref. 20 by showing that V-U5 variant effector proteins that are inactivated TnpB homologs encoded in Tn7-like transposons form a complex with the transposase subunit and enable crRNA-dependent transposition.

35 Klompe, S. E., Vo, P. L. H., Halpin-Healy, T. S. & Sternberg, S. H. Transposon-encoded CRISPR-Cas systems direct RNA-guided DNA integration. *Nature*, doi:10.1038/s41586-019-1323-z (2019).

This work complements Ref. 34 by experimentally validating the prediction made in Ref. 33, that interference-deficient subtype I-F CRISPR–Cas systems encoded in Tn7-like transposons enable crRNA-dependent transposition.

36 Kazlauskiene, M., Kostiuk, G., Venclovas, C., Tamulaitis, G. & Siksnys, V. A cyclic oligonucleotide signaling pathway in type III CRISPR-Cas systems. *Science* **357**, 605-609, doi:10.1126/science.aao0100 science.aao0100 [pii] (2017).

Along with Ref. 37, this article describes the signaling pathway that is involved in the function of type III CRISPR–Cas systems and involves the synthesis of cyclic oligoA molecules by Cas10, binding of these signaling molecules to the CARF domain of Csm6 and activation of the second domain of Casm6, the HEPN nuclease that catalyzes promiscuous RNA cleavage.

37 Niewoehner, O. *et al.* Type III CRISPR-Cas systems produce cyclic oligoadenylate second messengers. *Nature* **548**, 543-548, doi:10.1038/nature23467 (2017).

Along with Ref. 36, this article describes the signaling pathway that is involved in the function of type III CRISPR–Cas systems and involves the synthesis of cyclic oligoA molecules by Cas10, binding of these signaling molecules to the CARF domain of Csm6 and activation of the second domain of Casm6, the HEPN nuclease that catalyzes promiscuous RNA cleavage.

38 Makarova, K. S. & Koonin, E. V. Annotation and Classification of CRISPR-Cas Systems. *Methods Mol Biol* **1311**, 47-75, doi:10.1007/978-1-4939-2687-9_4 (2015).

39 Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389-3402 (1997).

40 Iranzo, J., Krupovic, M. & Koonin, E. V. The Double-Stranded DNA Virosphere as a Modular Hierarchical Network of Gene Sharing. *MBio* **7**, doi:10.1128/mBio.00978-16 (2016).

41 Iranzo, J., Martincorena, I. & Koonin, E. V. Cancer-mutation network and the number and specificity of driver mutations. *Proc Natl Acad Sci U S A* **115**, E6010-E6019, doi:10.1073/pnas.1803155115 (2018).

42 Makarova, K. S., Wolf, Y. I. & Koonin, E. V. The basic building blocks and evolution of CRISPR-cas systems. *Biochem Soc Trans* **41**, 1392-1400, doi:10.1042/BST20130038 BST20130038 [pii] (2013).

43 Koonin, E. V. & Makarova, K. S. Discovery of Oligonucleotide Signaling Mediated by CRISPR-Associated Polymerases Solves Two Puzzles but Leaves an Enigma. *ACS Chem Biol* **13**, 309-312, doi:10.1021/acschembio.7b00713 (2018).

44 Silas, S. *et al.* On the Origin of Reverse Transcriptase-Using CRISPR-Cas Systems and Their Hyperdiverse, Enigmatic Spacer Repertoires. *MBio* **8**, doi:e00897-17 [pii] 10.1128/mBio.00897-17 mBio.00897-17 [pii] (2017).

45 Puigbo, P., Makarova, K. S., Kristensen, D. M., Wolf, Y. I. & Koonin, E. V. Reconstruction of the evolution of microbial defense systems. *BMC Evol Biol* **17**, 94, doi:10.1186/s12862-017-0942-y 10.1186/s12862-017-0942-y [pii] (2017).

46 Garrett, R. A., Vestergaard, G. & Shah, S. A. Archaeal CRISPR-based immune systems: exchangeable functional modules. *Trends Microbiol* **19**, 549-556, doi:10.1016/j.tim.2011.08.002 S0966-842X(11)00149-1 [pii] (2011).

47 Reeks, J., Naismith, J. H. & White, M. F. CRISPR interference: a structural perspective. *Biochem J* **453**, 155-166, doi:10.1042/BJ20130316 BJ20130316 [pii] (2013).

48 Ozcan, A. *et al.* Type IV CRISPR RNA processing and effector complex formation in Aromatoleum aromaticum. *Nat Microbiol*, doi:10.1038/s41564-018-0274-8 (2018).

49 Makarova, K. S., Aravind, L., Wolf, Y. I. & Koonin, E. V. Unification of Cas protein families and a simple scenario for the origin and evolution of CRISPR-Cas systems. *Biol Direct* **6**, 38, doi:10.1186/1745-6150-6-38 1745-6150-6-38 [pii] (2011).

50 Venclovas, C. Structure of Csm2 elucidates the relationship between small subunits of CRISPR-Cas effector complexes. *FEBS Lett* **590**, 1521-1529, doi:10.1002/1873-3468.12179 (2016).

51 Chylinski, K., Makarova, K. S., Charpentier, E. & Koonin, E. V. Classification and evolution of type II CRISPR-Cas systems. *Nucleic Acids Res* **42**, 6091-6105, doi:10.1093/nar/gku241 gku241 [pii] (2014).

52 Briner, A. E. & Barrangou, R. Guide RNAs: A Glimpse at the Sequences that Drive CRISPR-Cas Systems. *Cold Spring Harb Protoc* **2016**, pdb top090902, doi:10.1101/pdb.top090902 2016/7/pdb.top090902 [pii] (2016).

53 Faure, G. *et al.* Comparative genomics and evolution of trans-activating RNAs in Class 2 CRISPR-Cas systems. *RNA Biol*, 1-14, doi:10.1080/15476286.2018.1493331 (2018).

54 Chyou, T. Y. & Brown, C. M. Prediction and diversity of tracrRNAs from type II CRISPR-Cas systems. *RNA Biol* **16**, 423-434, doi:10.1080/15476286.2018.1498281 (2019).

55 Fonfara, I., Richter, H., Bratovic, M., Le Rhun, A. & Charpentier, E. The CRISPR-associated DNA-cleaving enzyme Cpf1 also processes precursor CRISPR RNA. *Nature* **532**, 517-521, doi:10.1038/nature17945 nature17945 [pii] (2016).

56 East-Seletsky, A. *et al.* Two distinct RNase activities of CRISPR-C2c2 enable guide-RNA processing and RNA detection. *Nature*, 538, 270-273, doi: 10.1038/nature19802 [pii] (2016).

57 Liu, L. *et al.* Two Distant Catalytic Sites Are Responsible for C2c2 RNase Activities. *Cell* **168**, 121-134 (2017).

58 Kunin, V., Sorek, R. & Hugenholtz, P. Evolutionary conservation of sequence and secondary structures in CRISPR repeats. *Genome Biol* **8**, R61, doi:gb-2007-8-4-r61 [pii] 10.1186/gb-2007-8-4-r61 (2007).

59 Lange, S. J., Alkhnbashi, O. S., Rose, D., Will, S. & Backofen, R. CRISPRmap: an automated classification of repeat conservation in prokaryotic adaptive immune systems. *Nucleic Acids Res* **41**, 8034-8044, doi:10.1093/nar/gkt606 gkt606 [pii] (2013).

60 Almendros, C., Nobrega, F. L., McKenzie, R. E. & Brouns, S. J. J. Cas4-Cas1 fusions drive efficient PAM selection and control CRISPR adaptation. *Nucleic Acids Res* **47**, 5223-5230, doi:10.1093/nar/gkz217 (2019).

61 Koonin, E. V. & Aravind, L. Origin and evolution of eukaryotic apoptosis: the bacterial connection. *Cell Death Differ* **9**, 394-404, doi:10.1038/sj/cdd/4400991 (2002).

62 Vestergaard, G., Garrett, R. A. & Shah, S. A. CRISPR adaptive immune systems of Archaea. *RNA Biol* **11**, 156-167, doi:10.4161/rna.27990 27990 [pii] (2014).

63 Sinkunas, T. *et al.* Cas3 is a single-stranded DNA nuclease and ATP-dependent helicase in the CRISPR/Cas immune system. *EMBO J,* **30**, 1335-1342, doi:emboj201141 [pii] 10.1038/emboj.2011.41 (2011).

64 Al-Shayeb , B. *et al.* Clades of huge phage from across Earth's ecosystems. *bioRxiv*, doi: https://doi.org/10.1101/572362 (2019).

65 Leipe, D. D., Koonin, E. V. & Aravind, L. STAND, a class of P-loop NTPases including animal and plant regulators of programmed cell death: multiple, complex domain architectures, unusual phyletic patterns, and evolution by horizontal gene transfer. *J Mol Biol* **343**, 1-28, doi:10.1016/j.jmb.2004.08.023 S0022-2836(04)01001-0 [pii] (2004).

66 Newire, E., Aydin, A., Juma, S., Enne, V. & Roberts, A. P. Identification of a Type IV CRISPR-Cas system located exclusively on IncHI1B/ IncFIB plasmids in Enterobacteriaceae. *bioRxiv*, doi:doi: https://doi.org/10.1101/536375 (2019).

67 Makarova, K. S. *et al.* Predicted highly derived class 1 CRISPR-Cas system in Haloarchaea containing diverged Cas5 and Cas7 homologs but no CRISPR array. *FEMS Microbiol Lett* **366**, doi:10.1093/femsle/fnz079 (2019).

68 Zaremba-Niedzwiedzka, K. *et al.* Asgard archaea illuminate the origin of eukaryotic cellular complexity. *Nature* **541**, 353-358, doi:10.1038/nature21031 (2017).

69 Strecker, J. *et al.* Engineering of CRISPR-Cas12b for human genome editing. *Nat Commun* **10**, 212, doi:10.1038/s41467-018-08224-4 (2019).

70 Swarts, D. C. & Jinek, M. Mechanistic Insights into the cis- and trans-Acting DNase Activities of Cas12a. *Mol Cell* **73**, 589-600 e584, doi:10.1016/j.molcel.2018.11.021 (2019).

71 Meeske, A. J., Nakandakari-Higa, S. & Marraffini, L. A. Cas13-induced cellular dormancy prevents the rise of CRISPR-resistant bacteriophage. *Nature* **570**, 241-245, doi:10.1038/s41586-019-1257-5 (2019).

72 Karvelis, T. *et al.* PAM recognition by miniature CRISPR-Cas14 triggers programmable double-stranded DNA cleavage. *bioRxiv* **654897**, doi:https://doi.org/10.1101/654897. (2019).

73 Anantharaman, V., Makarova, K. S., Burroughs, A. M., Koonin, E. V. & Aravind, L. Comprehensive analysis of the HEPN superfamily: identification of novel roles in intra-genomic conflicts, defense, pathogenesis and RNA processing. *Biol Direct* **8**, 15, doi:10.1186/1745-6150-8-15 1745-6150-8-15 [pii] (2013).

74 Hudaiberdiev, S. *et al.* Phylogenomics of Cas4 family nucleases. *BMC Evol Biol* **17**, 232, doi:10.1186/s12862-017-1081-1 10.1186/s12862-017-1081-1 [pii] (2017).

75 Charpentier, E., Richter, H., van der Oost, J. & White, M. F. Biogenesis pathways of RNA guides in archaeal and bacterial CRISPR-Cas adaptive immunity. *FEMS Microbiol Rev* **39**, 428-441, doi:10.1093/femsre/fuv023 fuv023 [pii] (2015).

76 Dudek, N. K. *et al.* Novel Microbial Diversity and Functional Potential in the Marine Mammal Oral Microbiome. *Curr Biol* **27**, 3752-3762 e3756, doi:10.1016/j.cub.2017.10.040 (2017).

77 Castelle, C. J. *et al.* Biosynthetic capacity, metabolic variety and unusual biology in the CPR and DPANN radiations. *Nat Rev Microbiol* **16**, 629-645, doi:10.1038/s41579-018-0076-2 (2018).

78 Burstein, D. *et al.* Major bacterial lineages are essentially devoid of CRISPR-Cas viral defence systems. *Nat Commun* **7**, 10613, doi:10.1038/ncomms10613 ncomms10613 [pii] (2016).

79 Levin, B. R. Nasty viruses, costly plasmids, population dynamics, and the conditions for establishing and maintaining CRISPR-mediated adaptive immunity in bacteria. *PLoS Genet* **6**, e1001171, doi:10.1371/journal.pgen.1001171 (2010).

80 Iranzo, J., Lobkovsky, A. E., Wolf, Y. I. & Koonin, E. V. Evolutionary dynamics of the prokaryotic adaptive immunity system CRISPR-Cas in an explicit ecological context. *J Bacteriol* **195**, 3834-3844, doi:10.1128/JB.00412-13 JB.00412-13 [pii] (2013).

81 Iranzo, J., Lobkovsky, A. E., Wolf, Y. I. & Koonin, E. V. Immunity, suicide or both? Ecological determinants for the combined evolution of anti-pathogen defense systems. *BMC Evol Biol* **15**, 43, doi:10.1186/s12862-015-0324-2 10.1186/s12862-015-0324-2 [pii] (2015).

82 Gurney, J., Pleska, M. & Levin, B. R. Why put up with immunity when there is resistance: an excursion into the population and evolutionary dynamics of restriction-modification and CRISPR-Cas. *Philos Trans R Soc Lond B Biol Sci* **374**, 20180096, doi:10.1098/rstb.2018.0096 (2019).

83 Garcia-Martinez, J., Maldonado, R. D., Guzman, N. M. & Mojica, F. J. M. The CRISPR conundrum: evolve and maybe die, or survive and risk stagnation. *Microb Cell* **5**, 262-268, doi:10.15698/mic2018.06.634 (2018).

84 van Houte, S. *et al.* The diversity-generating benefits of a prokaryotic adaptive immune system. *Nature* **532**, 385-388, doi:10.1038/nature17436 nature17436 [pii] (2016).

85 Weinberger, A. D., Wolf, Y. I., Lobkovsky, A. E., Gilmore, M. S. & Koonin, E. V. Viral diversity threshold for adaptive immunity in prokaryotes. *MBio* **3**, e00456-00412, doi:10.1128/mBio.00456-12 e00456-12 [pii] mBio.00456-12 [pii] (2012).

86 Westra, E. R. *et al.* Parasite Exposure Drives Selective Evolution of Constitutive versus Inducible Defense. *Curr Biol* **25**, 1043-1049, doi:10.1016/j.cub.2015.01.065 S0960-9822(15)00129-3 [pii] (2015).

87 Bernheim, A., Bikard, D., Touchon, M. & Rocha, E. P. C. A matter of background: DNA repair pathways as a possible cause for the sparse distribution of CRISPR-Cas systems in bacteria. *Philos Trans R Soc Lond B Biol Sci* **374**, 20180088, doi:10.1098/rstb.2018.0088 (2019).

88 Koonin, E. V. & Krupovic, M. Evolution of adaptive immunity from transposable elements combined with innate immune systems. *Nat Rev Genet* **16**, 184-192, doi:10.1038/nrg3859 nrg3859 [pii] (2015).

89 Krupovic, M., Beguin, P. & Koonin, E. V. Casposons: mobile genetic elements that gave rise to the CRISPR-Cas adaptation machinery. *Curr Opin Microbiol* **38**, 36-43, doi:S1369-5274(16)30171-0 [pii] 10.1016/j.mib.2017.04.004 (2017).

90 Krupovic, M., Makarova, K. S., Forterre, P., Prangishvili, D. & Koonin, E. V. Casposons: a new superfamily of self-synthesizing DNA transposons at the origin of prokaryotic CRISPR-Cas immunity. *BMC Biology* **12**, 36 (2014).

91 Kieper, S. N. *et al.* Cas4 Facilitates PAM-Compatible Spacer Selection during CRISPR Adaptation. *Cell Rep* **22**, 3377-3384, doi:10.1016/j.celrep.2018.02.103 (2018).

92 Lee, H., Zhou, Y., Taylor, D. W. & Sashital, D. G. Cas4-Dependent Prespacer Processing Ensures High-Fidelity Programming of CRISPR Arrays. *Mol Cell* **70**, 48-59 e45, doi:10.1016/j.molcel.2018.03.003 (2018).

93 Shiimori, M., Garrett, S. C., Graveley, B. R. & Terns, M. P. Cas4 Nucleases Define the PAM, Length, and Orientation of DNA Fragments Integrated at CRISPR Loci. *Mol Cell* **70**, 814-824 e816, doi:10.1016/j.molcel.2018.05.002 (2018).

This work reveals the molecular details of the involvement of Cas4, an ancillary protein that cooperates with Cas1 and Cas2 in several CRISPR–Cas subtypes, in the process of adaptation.

94 Burroughs, A. M., Zhang, D., Schaffer, D. E., Iyer, L. M. & Aravind, L. Comparative genomic analyses reveal a vast, novel network of nucleotide-centric systems in biological conflicts, immunity and signaling. *Nucleic Acids Res* **43**, 10633-10654, doi:10.1093/nar/gkv1267 gkv1267 [pii] (2015).

95 Rostol, J. T. & Marraffini, L. A. Non-specific degradation of transcripts promotes plasmid clearance during type III-A CRISPR-Cas immunity. *Nat Microbiol* **4**, 656-662, doi:10.1038/s41564-018-0353-x (2019).

This work demonstrates that the indiscriminate RNA cleavage by the HEPN RNase domain of the Csm6 protein of type III CRISPR–Cas systems induces growth arrest in the host bacteria, providing a back-up defence mechanism.

96 Makarova, K. S., Wolf, Y. I., Snir, S. & Koonin, E. V. Defense islands in bacterial and archaeal genomes and prediction of novel defense systems. *J Bacteriol* **193**, 6039-6056, doi:JB.05535-11 [pii] 10.1128/JB.05535-11 (2011).

97 Kapitonov, V. V., Makarova, K. S. & Koonin, E. V. ISC, a Novel Group of Bacterial and Archaeal DNA Transposons That Encode Cas9 Homologs. *J Bacteriol* **198**, 797-807, doi:10.1128/JB.00783-15 JB.00783-15 [pii] (2015).

98 Shmakov, S. A. *et al.* Systematic prediction of functionally linked genes in bacterial and archaeal genomes. *Nat Protoc* **14**, 3013-3031, doi:10.1038/s41596-019-0211-1 (2019).

99 Athukoralage, J. S., Rouillon, C., Graham, S., Gruschow, S. & White, M. F. Ring nucleases deactivate type III CRISPR ribonucleases by degrading cyclic oligoadenylate. *Nature*, doi:10.1038/s41586-018-0557-5 (2018).

This works expands the characterization of the signaling pathway in type III CRISPR–Cas sequence by showing that a distinct variety of CARF domains cleave the cyclic oligoA molecules produced by Cas10 and thus regulate the pathway.

Acknowledgements

K.S.M., Y.I.W., J.I., S.A.S., and E.V.K. are supported through the Intramural Research Program of the National Institutes of Health of the USA; S.A.S. was also supported by RFBR (research project 18-34-00012) and a Systems Biology Fellowship of Philip Morris Sales and Marketing; S.M. was funded by funding from Natural Sciences and Engineering Research Council of Canada (Discovery program) and holds a Tier 1 Canada Research Chair in Bacteriophages.

Author contributions

K.S.M., Y.I.W., J.I., S.A.S., D.C., Č.V., and E.V.K. researched data for the article.

K.S.M., Y.I.W., S.J.J.B., O.S.A., E.C., D.C., D.H.H., P.H., S.M., F.J.M.M., D.S., S.A.A., V.S., M.P.T., Č.V., M.F.W., A.F.Y., W.Y., F.Z., R.A.G., R.B., J.v.d.O., R.B. and E.V.K. substantially contributed to discussion of content.

K.S.M., J.I., Y.I.W. and E.V.K. wrote the article.

K.S.M.,Y.I.W, S.M., R.A.G, J.v.d.O., R.B. and E.V.K. reviewed/edited the manuscript before submission.

Competing interests

The authors declare no competing interests.

Publisher's note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Supplementary information

Supplementary information is available for this paper at https://doi.org/10.1038/s415XX-XXX-XXXX-X

Figures

Figure 1. Updated classification of class 1 CRISPR–Cas systems.
The figure schematically shows representative (typical) CRISPR–*cas* loci of each class 1 subtype and selected distinct variants, with the dendrogram on the left showing the likely evolutionary relationships between the types and subtypes. The column on the right indicates the organism and the corresponding gene range. Homologous genes are colour-coded and identified by a family name. The gene names follow the previous classification[18]. Where both a systematic

name and a legacy name are commonly used, the legacy name is given under the systematic name. The small subunit is encoded by *csm2*, *cmr5*, *cse2*, *csa5* and several additional families of homologous genes that are collectively denoted *cas11*. The adaptation module genes *cas1* and *cas2* are dispensable in subtype III-A and subtype III-B (dashed lines). Gene regions colored cream represent the HD nuclease domain; the HD domain in Cas10 is distinct from that of Cas3 and Cas3′. Functionally uncharacterized genes are shown in grey. The pink shading shows the effector module. The grey shading of different hues shows the two levels of classification, subtype and variants. Most of the subtype III-B, III-C, III-E, III-F loci as well as IV-B and IV-C loci lack CRISPR arrays and are shown accordingly, although for each of the type III subtypes exceptions have been detected. CHAT, protease domain of the caspase family; RT, reverse transcriptase; TPR, Tetratricopeptide repeats.

Figure 2. Updated classification of class 2 CRISPR–Cas systems.
The figure schematically shows representative (typical) CRISPR–*cas* loci of each class 2 subtype and selected distinct variants, with the dendrogram on the left showing the likely evolutionary relationships between the types and subtypes. The column on the right indicates the organism and the corresponding gene range. Homologous genes are colour-coded and identified by a family name following the previous classification[18]. Where both a systematic name and a legacy name are commonly used, the legacy name is given under the systematic name. The grey shading of different hues shows the two levels of classification, subtypes and variants. The adaptation module genes *cas1* and *cas2* are present in only a subset of the subtype V-D, VI-A and VI-D loci and are accordingly shown by dashed lines. The WYL-domain-encoding genes and *csx27* genes are also dispensable and shown by dashed lines. Additional genes encoding components of the interference module, such as tracrRNA, are shown. The domains of the effector proteins are color-coded: RuvC-like nuclease, yellow; HNH nuclease, green; HEPN RNase, purple; transmembrane domains, blue.

Figure 3. Distribution of the six types of CRISPR–Cas system in the major archaeal and bacterial phyla.
The heat map shows the weighted fraction (between 0 and 1.0) of the genomes in each of the major archaeal and bacterial phyla in which CRISPR–Cas systems of the respective type were

detected. Each CRISPR–*cas* locus of a given type within a taxon was assigned a weight, equal to the weight of the respective genome (see Supplementary Materials and Methods for details); additionally, the weights of the genomes that lack CRISPR–Cas loci were collected. The sum of the weights of the CRISPR–*cas* loci of each type was normalized by the sum total of the weights across the taxon. Partial or unknown indicates CRISPR–*cas* loci that could not be assigned to any of the known types.

Figure 4. Ancillary genes in CRISPR–Cas systems.
The basic molecular machinery of the CRISPR–Cas systems consists of the *cas* core genes. The core genes are often accompanied by diverse ancillary genes that perform additional or regulatory functions. The ancillary genes are typically present only in subsets of the CRISPR–*cas* loci of the respective types and subtypes, and often also occur in other, non-*cas* genomic contexts. The prediction of the ancillary genes was performed using the 'CRISPRicity' protocol as previously described [32,98]. Operationally, the list of ancillary genes includes families, labeled as 'associated' in the profFam.tab column in Supplementary Dataset 2. The number of occurrences (count) of ancillary genes in each unambiguously classified CRISPR-*cas* locus was averaged across the system subtypes using genome weights, calculated as described in the Supplementary Materials and Methods. Occurrence of ancillary genes across the types and subtypes of CRISPR–Cas systems is shown (**part a**). The vertical axis shows the weighted mean number of ancillary genes per locus in different subtypes. The common ancillary genes and their distribution among CRISPR–Cas types and subtypes is shown (**part b**). Gene families are denoted with the corresponding profile names (Supplementary Dataset 2). The weighted mean number of ancillary genes per locus in different subtypes is colour-coded as per the scale shown in the bottom.

Figure 5. Outline of a complete scenario for the origins and evolution of CRISPR–Cas systems.
The figure depicts a hypothetical scenario of the origin of CRISPR–Cas systems from an ancestral signaling system (possibly, an abortive infection defense system (Abi)). This putative ancestral Abi module shares a cyclic oligoA polymerase Palm domain (RNA recognition motif (RRM) fold) with Cas10 and is proposed to function analogously to type III CRISPR–Cas systems. Specifically, cyclic oligoA molecules that are synthesized in response to virus infection

bind to the CARF domain of the second protein in this system, resulting in activation of the RNase activity of the HEPN domain which induces dormancy through indiscriminate RNA cleavage. This putative ancestral Abi module would give rise to the type III-like CRISPR–Cas effector module via the duplication of the RRM domain, with subsequent inactivation of one the copies (the two RRM domain are denoted RRM1 and RRM2). The ancestral class 1 CRISPR–Cas system is inferred to have evolved through the merger of two modules, the adaptation module, including the CRISPR repeats, derived from a casposon, and the type III-like effector module likely derived from the ancestral Abi system. The subsequent acquisition of the HD nuclease domain by the effector module provided for RNA-guided DNA cleavage. Inactivation of the oligoA polymerase domain in the effector complex or, possibly, replacement of Cas10 by an unrelated protein and acquisition of the Cas3 helicase led to the emergence of type I systems which lack the cyclic oligoA-dependent signaling pathway and exclusively cleave dsDNA. Class 2 systems of type II and different subtypes of type V appear to have evolved independently by recruitment of distinct TnpB nucleases that are encoded by IS605-like transposable elements. Type VI likely originated from an RNA-cleaving, HEPN-domain-containing abortive infection or toxin-antitoxin system. Some CRISPR-Cas systems, such as type IV and Tn7-linked systems I-F3 and V-K, were subsequently recruited by mobile genetic elements and lost their interference capacity along with the original defense function. The key evolutionary events are described to the right of the images. The typical CRISPR–*cas* operon organization is shown for each CRISPR–Cas subtype and selected, distinct variants. Homologous genes are colour-coded and identified by a family name following the previous classification[18]. The multi-forking arrows denote events that have been inferred to have occurred on multiple, independent occasions during the evolution of CRISPR–Cas systems. Additional abbreviations: "GGDD", key catalytic motif of the cyclase or polymerase domain of Cas10 that is involved in the synthesis of cyclic oligoA signaling molecules; TR, terminal repeats; TSD, target site duplication, the likely source of the ancestral repeats[88].

Box 1. The two classes of CRISPR–Cas systems and their modular organization.
Class 1 CRISPR–Cas systems have effector modules composed of multiple Cas proteins that form a crRNA-binding complex and function together in the target binding and processing .
Class 2 systems have a single, multidomain crRNA-binding protein that is functionally

analogous to the entire effector complex of class 1. The top panel of the figure illustrates the generic organizations of class 1 and class 2 CRISPR–Cas loci. The bottom panel of the figure shows the functional modules of CRISPR–Cas systems. The scheme shows the typical relationships between genetic, structural and functional organization for the six types of CRISPR–Cas systems. Protein names follow the current nomenclature. An asterisk indicates the putative small subunit that might be fused to the large subunit in several type I subtypes. The pound symbol (#) indicates that other unknown sensor, effector, and Ring nuclease protein families could be involved in the same signaling pathway. Dispensable (and/or missing in some subtypes and variants) components are indicated by dashed outlines. Cas6 is shown with a thin solid outline for type I because it is dispensable in some but not most systems and with a dashed line for type III because most of these apparently use the Cas6 protein provided *in trans* by other CRISPR–*cas* loci. The three colors for Cas9, Cas10, Cas12 and Cas13 reflect the fact that these proteins contribute to different stages of the CRISPR–Cas response. The CARF and HEPN domain proteins are the most common sensors and effectors, respectively, in the type III ancillary modules, but several alternative sensors and effectors have been identified as well[43]. Ring nucleases are a distinct variety of CARF domain proteins that cleave cyclic oligoA produced by Cas10 and thus control the indiscriminate RNase activity of the HEPN domain of Csx6 (ref. [99]). Figure modified from Ref. 18.

Box 2. Strategies for classification and principles of nomenclature of CRISPR–Cas systems. The top panel of the figure shows the hierarchy of the main sources of information that are used for classification of CRISPR–Cas systems. Computational strategies exploit a combination of comparative genomic and experimental evidence, aiming to analyze the components of the *cas* loci, establish their organization and place them within the classification scheme. Given the fast evolution resulting in extensive sequence divergence of most Cas proteins, sensitive sequence similarity search and phylogenetic analysis methods are crucial for the correct assignment of the individual components; neighborhood analysis is necessary for understanding the architecture of the specific variants of the system. Experimental data are often essential to determine distinct features of CRISPR–Cas systems and molecular details of their mechanisms. Experimental results guide additional computational analyses by providing information on functional similarity between components of different CRISPR–Cas systems, and on the contributions of different

components to the system function. The bottom panel illustrates the 3-level gene nomenclature scheme and the evidence used for the classification of a variant of subtype VI-B are shown. Gene neighborhood analysis allows unambiguous classification of this system as Class 2. Motif search and profile-profile comparison of HEPN domains result in classification of as Type VI. However, PSI-BLAST searches do not detect sequence similarity to any of the previously identified type VI effector proteins. Moreover, these loci encompass distinct ancillary genes, supporting their classification as a separate subtype (VI-B). The phylogenetic tree of Cas13b contains two strongly supported branches that are associated with distinct ancillary genes. Accordingly, subtype VI-B is subdivided into two variants [25].

Glossary terms

CRISPR, clustered regularly interspaced short palindromic repeats present in most archaeal and many bacterial genomes

CRISPR array, genomic locus containing multiple, tandem CRISPR

Cas, CRISPR-associated (proteins)

CRISPR–Cas, archaeal and bacterial system of adaptive immunity that consists of a CRISPR array and *cas* genes

Spacers, unique segments of DNA inserted between CRISPR units

Protospacers, segments of DNA (typically, from a virus or plasmid) that are acquired by the CRISPR-Cas systems via the activity of the adaptation complex

PAM, protospacer adjacent motif, a short nucleotide sequence next to the protospacer that is required for target recognition by the crRNA-effector

crRNA, short RNA molecule containing the spacer sequence and parts of CRISPR and used as the guide to target and cleave the cognate foreign DNA or RNA

pre-crRNA, long transcript of a CRISPR locus that is processed to yield the crRNA

CRISPR-Cas system to become spacers

Adaptation, first stage of the CRISPR-Cas response that involves spacer acquisition

Interference, final stage of the CRISPR-Cas response that involves recognition and cleavage of the target DNA or RNA

Transposon, a mobile genetic element, typically flanked by inverted terminal repeats, that changes its location in the host genome by inserting into new sites with the help of a transposon-encoded enzyme known as transposase, integrase or recombinase

Casposon, a member of a distinct class of transposons that employ a Cas1 homolog as the transposases and are thought to be the ancestors of CRISPR-Cas adaptation modules

Table of contents blurb

The number and diversity of CRISPR–Cas systems has substantially increased in recent years. In this Review, Koonin and colleagues provide an updated evolutionary classification of CRISPR–Cas systems and *cas* genes, with an emphasis on major developments, and outline a complete scenario for the origins and evolution of CRISPR–Cas systems.