An updated evolutionary classification of CRISPR–Cas systems

Kira S. Makarova[1], Yuri I. Wolf[1], Omer S. Alkhnbashi[2], Fabrizio Costa[2], Shiraz A. Shah[3], Sita J. Saunders[2], Rodolphe Barrangou[4], Stan J. J. Brouns[5], Emmanuelle Charpentier[6], Daniel H. Haft[1], Philippe Horvath[7], Sylvain Moineau[8], Francisco J. M. Mojica[9], Rebecca M. Terns[10], Michael P. Terns[10], Malcolm F. White[11], Alexander F. Yakunin[12], Roger A. Garrett[3], John van der Oost[5], Rolf Backofen[2,13] and Eugene V. Koonin[1]

Abstract | The evolution of CRISPR–*cas* loci, which encode adaptive immune systems in archaea and bacteria, involves rapid changes, in particular numerous rearrangements of the locus architecture and horizontal transfer of complete loci or individual modules. These dynamics complicate straightforward phylogenetic classification, but here we present an approach combining the analysis of signature protein families and features of the architecture of *cas* loci that unambiguously partitions most CRISPR–*cas* loci into distinct classes, types and subtypes. The new classification retains the overall structure of the previous version but is expanded to now encompass two classes, five types and 16 subtypes. The relative stability of the classification suggests that the most prevalent variants of CRISPR–Cas systems are already known. However, the existence of rare, currently unclassifiable variants implies that additional types and subtypes remain to be characterized.

The CRISPR–Cas modules are adaptive immune systems that are present in most archaea and many bacteria[1–5] and provide sequence-specific protection against foreign DNA or, in some cases, RNA[6]. A CRISPR locus consists of a CRISPR array, comprising short direct repeats separated by short variable DNA sequences (called 'spacers'), which is flanked by diverse *cas* genes. CRISPR–Cas immunity involves three distinct mechanistic stages: adaptation, expression and interference[7–11]. The adaptation stage involves the incorporation of fragments of foreign DNA (known as 'protospacers') from invading viruses and plasmids into the CRISPR array as new spacers. These spacers provide the sequence memory for a targeted defence against subsequent invasions by the corresponding virus or plasmid. During the expression stage, the CRISPR array is transcribed as a precursor transcript (pre-crRNA), which is processed and matured to produce CRISPR RNAs (crRNAs). During the interference stage, crRNAs, aided by Cas proteins, function as guides to specifically target and cleave the nucleic acids of cognate viruses or plasmids[7,9,12,13]. Recent studies suggest that CRISPR–Cas systems can also be used for non-defence roles, such as the regulation of collective behaviour and pathogenicity[14–16].

Numerous, highly diverse Cas proteins are involved in the different stages of CRISPR activity (BOX 1; see Supplementary information S1 (table)). Briefly, Cas1 and Cas2, which are present in most known CRISPR–Cas systems, form a complex that represents the adaptation module and is required for the insertion of spacers into CRISPR arrays[17,18]. Protospacer acquisition in many CRISPR–Cas systems requires recognition of a short protospacer adjacent motif (PAM) in the target DNA[19–22]. During the expression stage, the pre-crRNA molecule is bound to either Cas9 (which is a single, multidomain protein) or to a multisubunit complex, forming the crRNA–effector complex. The pre-crRNA is processed into crRNAs by an endonuclease subunit of the multisubunit effector complex[23] or via an alternative mechanism that involves bacterial RNase III and an additional RNA species, the tracrRNA (transactivating CRISPR RNA)[24]. Finally, at the interference stage, the mature crRNA remains bound to Cas9 or to the multisubunit crRNA–effector complex, which recognizes and cleaves the cognate DNA[10,11,25,26] or RNA[26–31].

The rapid evolution of most *cas* genes[32–34] and the remarkable variability in the genomic architecture of CRISPR–*cas* loci poses a major challenge for the

[1]*National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland 20894, USA.*
Correspondence to E.V.K.
e-mail:
koonin@ncbi.nlm.nih.gov

doi:10.1038/nrmicro3569
Published online
28 September 2015

consistent annotation of Cas proteins and for the classification of CRISPR–Cas systems[13,35]. Nevertheless, a consistent classification scheme is essential for expedient and robust characterization of CRISPR–*cas* loci in new genomes, and thus important for further progress in CRISPR research. Owing to the complexity of the gene composition and genomic architecture of the CRISPR–Cas systems, any single, all-encompassing classification criterion is rendered impractical, and thus a 'polythetic' approach based on combined evidence from phylogenetic, comparative genomic and structural analysis was developed[13]. At the top of the classification hierarchy are the three main types of CRISPR–Cas systems (type I–type III). These three types are readily distinguishable by virtue of the presence of unique signature proteins: Cas3 for type I, Cas9 for type II and Cas10 for type III[13]. Within each type of CRISPR–Cas system, several subtypes have been delineated based on additional signature genes and characteristic gene arrangements[13,35]. Recently, in-depth sequence and structural analysis of the effector complexes from different variants of CRISPR–Cas systems has uncovered common principles of their organization and function[4,30,31,36–46]. In parallel, the biotechnological development of molecular components of type II CRISPR–Cas systems into a powerful new generation of genome editing and engineering tools has triggered intensive research into the functions and mechanisms of these systems, thereby advancing our understanding of the Cas proteins and associated RNAs[47,48].

In this Analysis article, we refine and extend the classification of CRISPR–*cas* loci based on a comprehensive analysis of the available genomic data. As a result of this analysis, we introduce two classes of CRISPR–Cas systems as a new, top level of classification and define two putative new types and five new subtypes within these classes, resulting in a total of five types and 16 subtypes. We employ this classification to analyse the evolutionary relationships between CRISPR–*cas* loci using several measures. The results of this analysis highlight pronounced modularity as an emerging trend in the evolution of CRISPR–Cas systems. Finally, we demonstrate the potential for automated annotation of CRISPR–*cas* loci by developing a computational approach that uses the new classification to assign CRISPR–Cas system subtype with high precision.

Author addresses

[1]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland 20894, USA.
[2]Bioinformatics group, Department of Computer Science, University of Freiberg, Georges-Kohler-Allee 106, 79110 Freiberg, Germany.
[3]Archaea Centre, Department of Biology, Copenhagen University, Ole Maaløes Vej 5, DK2200 Copenhagen N, Denmark
[4]Department of Food, Bioprocessing and Nutrition Sciences, North Carolina State University, Raleigh, North Carolina 27606, USA.
[5]Laboratory of Microbiology, Wageningen University, Dreijenplein 10, 6703HB Wageningen, Netherlands.
[6]Department of Regulation in Infection Biology, Helmholtz Centre for Infection Research, D-38124 Braunschweig, Germany.
[7]DuPont Nutrition and Health, BP10, Dangé-Saint-Romain 86220, France.
[8]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.
[9]Departamento de Fisiología, Genética y Microbiología. Universidad de Alicante. 03080-Alicante, Spain.
[10]Biochemistry and Molecular Biology, Genetics and Microbiology, University of Georgia, Davison Life Sciences Complex, Green Street, Athens, Georgia 30602, USA.
[11]Biomedical Sciences Research Complex, University of St Andrews, North Haugh, St Andrews, KY16 9TZ, UK.
[12]Department of Chemical Engineering and Applied Chemistry, University of Toronto, Toronto, M5S 3E5, Canada.
[13]BIOSS Centre for Biological Signaling Studies, Cluster of Excellence, University of Freiburg, Germany.

Classification of CRISPR–*cas* loci

The classification of CRISPR–Cas systems should ideally represent the evolutionary relationships between CRISPR–*cas* loci. However, the pervasive exchange and divergence of *cas* genes and gene modules has resulted in a complex network of evolutionary relationships that cannot be readily (and cleanly) partitioned into a small number of distinct groupings (although such partitioning might be achievable for individual modules, see below). Therefore, we adopted a two-step classification approach that first identified all *cas* genes in each CRISPR–*cas* locus and then determined the signature genes and distinctive gene architectures that would allow the assignment of these loci to types and subtypes.

To robustly identify *cas* genes, which is a non-trivial task owing to high sequence variability, we developed a library of 394 position-specific scoring matrices (PSSM)[49] for all 93 known protein families associated with CRISPR–Cas systems (see Supplementary information S2 (table)). Importantly, this set included 229 PSSMs for recently characterized families that were not part of the previous CRISPR–Cas classification[13]. The PSSMs were used to search the protein sequences annotated in 2,751 complete archaeal and bacterial genomes that were available at the National Center for Biotechnology Information (NCBI) as of 1 February 2014 (see Supplementary information S3 (box) for a detailed description of the methods). A highly significant similarity threshold was used to identify bona fide *cas* genes. Genes that were located in the same genomic neighbourhood as bona fide *cas* genes (irrespectively of their proximity to a CRISPR array) and that encoded proteins with moderate similarity to Cas PSSMs were then identified as putative *cas* genes. This two-step procedure was devised to minimize the false-positive rate, while allowing the detection of diverged variants of Cas proteins.

Gene neighbourhoods around the identified *cas* genes were merged into 1,949 distinct *cas* loci from 1,302 of the 2,751 analysed genomes, including 1,694 complete loci. A *cas* locus was annotated as 'complete' if it encompassed at least the full complement of genes for the main components of the interference module (the multisubunit crRNA–effector complex or Cas9). This criterion was adopted because, although the adaptation module genes *cas1* and *cas2* are the most common *cas* genes, many otherwise complete (and hence thought to be functionally active) CRISPR–Cas systems lack *cas1* and *cas2* and seem to instead depend on adaptation modules from other loci in the same genome. Within the set of complete loci, 111 composite loci that contained two or more adjacent CRISPR–Cas units (each consisting of at least a full complement of essential effector complex components) were identified and split into distinct units. Each locus or unit was classified by scoring type-specific and subtype-specific PSSMs that were constructed from multiple sequence alignments of the respective signature Cas proteins (see Supplementary information S2,S4 (tables)). For some of the more diverged signature proteins, multiple PSSMs were required for a single protein to capture the entire diversity of the cognate CRISPR–Cas subtype.

Box 1 | **Cas protein families and functional modules**

The Cas proteins can be divided into four distinct functional modules: adaptation (spacer acquisition); expression (crRNA processing and target binding); interference (target cleavage); and ancillary (regulatory and other CRISPR-associated functions) (FIG. 1). In recent years, a wealth of structural and functional information has accumulated for the core Cas proteins (Cas1–Cas10) (see Supplementary information S1 (table)), which allows them to be classified into these modules.

The adaptation module is largely uniform across CRISPR–Cas systems and consists of the Cas1 and Cas2 proteins, with possible additional involvement of the restriction endonuclease superfamily enzyme Cas4 (REF. 91) and, in type II systems, Cas9 (REFS 63,64). Cas1, which adopts a unique α-helical fold, is an integrase that mediates the insertion of new spacers into CRISPR arrays by cleaving specific sites within the repeats[17,89,92]. The role of Cas2, which is a homologue of the mRNA interferase toxins of numerous toxin–antitoxin systems, is less well understood[3,72,93,94]. Cas2 has been shown to form a complex with Cas1 in the *Escherichia coli* type I CRISPR–Cas system and is required for adaptation. However, although Cas2 has RNase[95] and DNase activities[96], its catalytic residues are dispensable for adaptation[17], indicating that these activities are not directly involved in this process, at least in this species.

The expression and interference modules are represented by multisubunit CRISPR RNA (crRNA)–effector complexes[36,38,39,43–46,97,98] (BOX 2) or, in type II systems, by a single large protein, Cas9 (REFS 24,25,99). In the expression stage, pre-crRNA is bound to the multisubunit crRNA–effector complex, or to Cas9, and processed into a mature crRNA in a step catalysed by an RNA endonuclease[23] (typically Cas6; in type I and type III systems) or an alternative mechanism that involves RNase III and a transactivating CRISPR RNA (tracrRNA)[24] (in type II systems). However, in at least one type II CRISPR–Cas system, that of *Neisseria meningitidis*, crRNAs with mature 5′ ends are directly transcribed from internal promoters, and crRNA processing does not occur[69].

In the interference module, the crRNA–effector complex (in type I and type III systems) or Cas9 (in type II systems) combines nuclease activity with dedicated RNA-binding domains. Target binding relies on base pair formation with the spacer region of the crRNA. Cleavage of the target is catalysed by the HD family nuclease (Cas3″ or a domain in Cas3) in type I systems[52,100], by the combined action of the Cas7 and Cas10 proteins in type III systems[26,39,46,101–104] or by Cas9 in type II systems[25]. In type I systems, the HD nuclease domain is either fused to the superfamily 2 helicase Cas3′ (REFS 50–52) or is encoded by a separate gene, *cas3″*, whereas in type III systems a distinct HD nuclease domain is fused to Cas10 and is thought to cleave single-stranded DNA during interference[105]. In type II systems, the RuvC-like nuclease (RNase H fold) domain and the HNH (McrA-like) nuclease domain of Cas9 each cleave one of the strands of the target DNA[25,106]. Remarkably, the large (~950–1,400 amino acids) multidomain Cas9 protein is required for all three of the functional steps of CRISPR-based immunity (adaptation, expression and interference) in type II systems and thus concentrates much of the CRISPR–Cas system's function in a single protein.

The ancillary module is a combination of various proteins and domains that, with the exception of Cas4, are much less common than the core Cas proteins in CRISPR–Cas systems. Aside from its putative role in adaptation, Cas4 is thought to contribute to CRISPR–Cas-coupled programmed cell death[3,94]. Other notable components of the ancillary module include: a diverse set of proteins containing the CRISPR-associated Rossmann fold (CARF) domain[35,107], which have been hypothesized to regulate CRISPR–Cas activity[107] (in many type I and type III systems); and the inactivated P-loop ATPase Csn2, which forms a homotetrameric ring that accommodates linear double-stranded DNA in the central hole (in type II systems)[108–111]. Csn2 is not required for interference but apparently has a role in spacer integration, possibly preventing damage from the double-strand break in the chromosomal DNA[6,110]. Ancillary module genes are often found outside of CRISPR–*cas* loci, but the functions of these stand-alone genes have not been characterized in depth[72,94].

Of the single-unit complete loci, 1,574 (93%) were assigned to a specific subtype or the newly defined putative types IV and V, which are not split into subtypes, eight were identified up to the type only and one remained unclassified by our procedure (a subtype I-D system operon that is adjacent to the remnants of a subtype III-B system operon disrupted by recombination).

Our analysis suggests that the CRISPR–Cas systems can be divided on the basis of the genes encoding the effector modules; that is, whether the systems have several variants of a multisubunit complex (the CRISPR-associated complex for antiviral defence (Cascade) complex, the Csm complex or the Cmr complex) or Cas9. Thus, we introduce a new, broadest level of classification of CRISPR–Cas systems, which divides them into 'class 1' and 'class 2'. Class 1 systems possess multisubunit crRNA–effector complexes, whereas in class 2 systems all functions of the effector complex are carried out by a single protein, such as Cas9. We also find evidence for two putative new types, type IV and type V, which belong to class 1 and class 2, respectively. These observations result in a new classification system in which CRISPR–Cas systems are clustered into five types, each with a distinctive composition of expression, interference and adaptation modules (FIG. 1). These five types are divided into 16 subtypes, including five new subtypes (II-C, III-C and III-D, together with the single subtypes of type IV and type V systems), as detailed below.

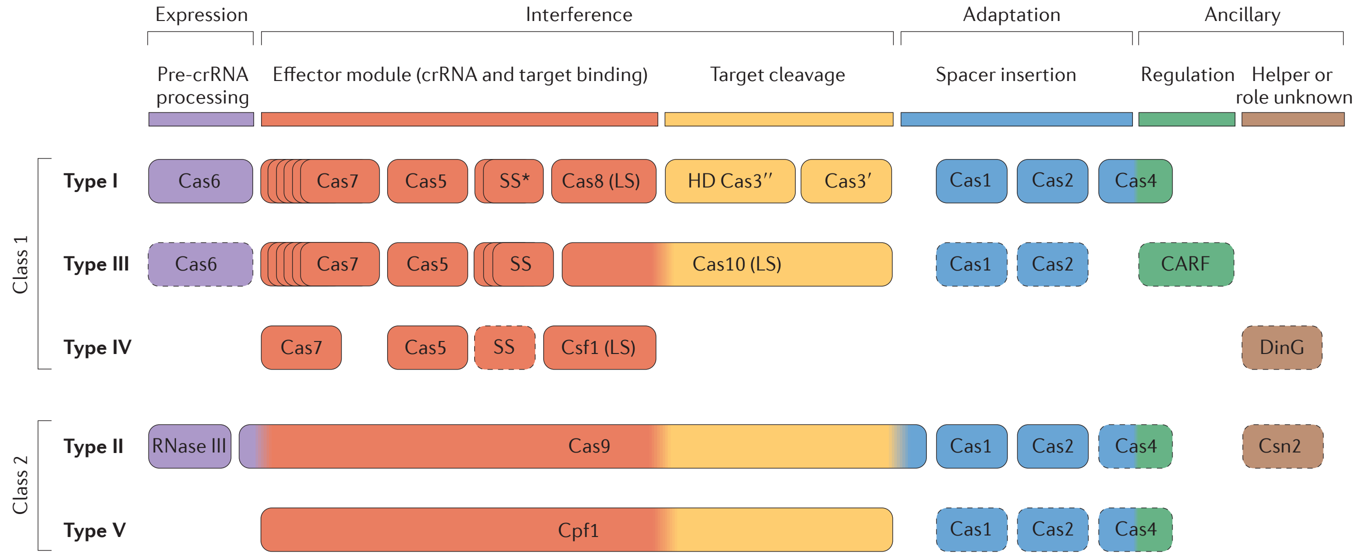


Figure 1 | **Functional classification of Cas proteins.** Protein names follow the current nomenclature and classification[13]. An asterisk indicates that the putative small subunit (SS) protein is instead fused to Cas8 (the type I system large subunit (LS)) in several type I subtypes[33]. The type III system LS and type IV system LS are Cas10 and Csf1 (a Cas8 family protein), respectively. Dispensable components are indicated by dashed outlines. Cas6 is shown with a solid outline for type I because it is dispensable in some but not most systems and by a dashed line for type III because most systems lack this gene and use the Cas6 provided *in trans* by other CRISPR–*cas* loci. The two colours for Cas4 and three colours for Cas9 reflect that these proteins contribute to different stages of the CRISPR–Cas response. The functions shown for type IV and type V system components are proposed based on homology to the cognate components of other systems, and have not yet been experimentally verified. The functional assignments for Cpf1 are tentatively inferred by analogy with Cas9 (only the RuvC (and TnpB)-like domains of the two proteins are homologous). CARF, CRISPR-associated Rossmann fold; pre-crRNA, pre-CRISPR RNA. This research was originally published in *Biochem. Soc. Trans.* Makarova K. S., Wolf Y. I., & Koonin E. V. The basic building blocks and evolution of CRISPR–Cas systems. *Biochem. Soc. Trans.* 2013; 41: 1392–1400 © The Biochemical Society.

Class 1 CRISPR–Cas systems

Class 1 CRISPR–Cas systems are defined by the presence of a multisubunit crRNA–effector complex. The class includes type I and type III CRISPR–Cas systems, as well as the putative new type IV.

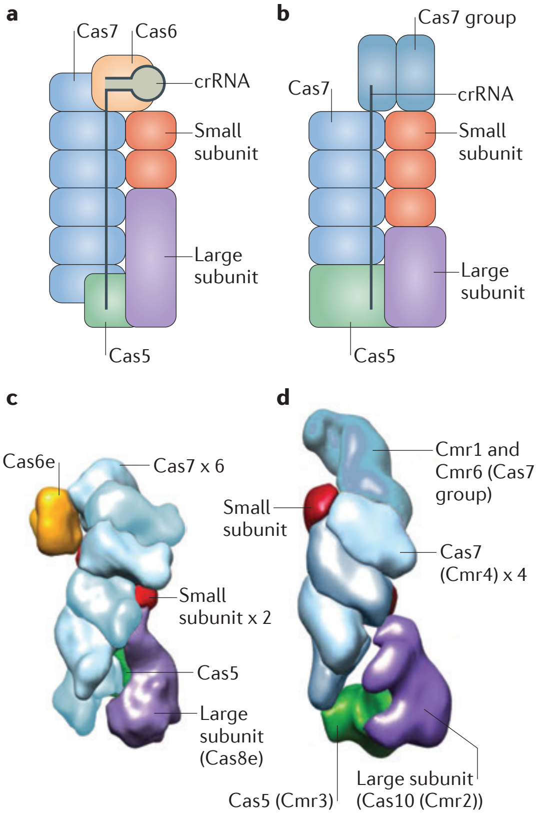
Type I CRISPR–Cas systems. All type I loci contain the signature gene *cas3* (or its variant *cas3′*), which encodes a single-stranded DNA (ssDNA)-stimulated superfamily 2 helicase with a demonstrated capacity to unwind double-stranded DNA (dsDNA) and RNA–DNA duplexes[50–52]. Often, the helicase domain is fused to a HD family endonuclease domain that is involved in the cleavage of the target DNA[50,53]. The HD domain is typically located at the amino terminus of Cas3 proteins (with the exception of subtype I-U and several subtype I-A systems, in which the HD domain is at the carboxyl terminus of Cas3) or is encoded by a separate gene (*cas3″*) that is usually adjacent to *cas3′* (FIG. 1).

Type I systems are currently divided into seven subtypes, I-A to I-F and I-U, all of which have been defined previously[13]. In the case of subtype I-U, U stands for uncharacterized because the mechanism of pre-crRNA cleavage and the architecture of the effector complex for this system remain unknown[33]. The type I-C, I-D, I-E and I-F CRISPR–Cas systems are typically encoded by a single (predicted) operon that encompasses the *cas1*, *cas2* and *cas3* genes together with the genes for the subunits of the Cascade complex (BOX 2). By contrast, many type I-A and I-B loci seem to have a different organization in which the *cas* genes are clustered in two or more (predicted) operons[35]. In most type I loci, each of the *cas* gene families is represented by a single gene.

Each type I subtype has a defined combination of signature genes and distinct features of operon organization (FIG. 2; see Supplementary information S4 (table)). Notably, *cas4* is absent in I-E and I-F systems, and *cas3* is fused to *cas2* in I-F systems. Subtypes I-E and I-F are monophyletic (that is, all systems of the respective subtype are descended from a single ancestor) in phylogenetic trees of Cas1 and Cas3, and each has one or more distinct signature genes (see Supplementary information S4,S5,S6 (table, box, box)).

Subtypes I-A, I-B and I-C seem to be descendants of the ancestral type I gene arrangement (*cas1–cas2–cas3–cas4–cas5–cas6–cas7–cas8*)[4,54]. This arrangement is preserved in subtype I-B, whereas subtypes I-A and I-C are diverged derivatives of I-B with differential gene loss and rearranged gene orders. A single signature gene for each of these subtypes could not be defined. The only protein that shows no significant sequence similarity between the subtypes is Cas8. However, the Cas8 sequence is highly diverged even within subtypes, so that consistent application of the signature gene approach would result in numerous new subtypes. For example, there are at least 10 distinct Cas8b families within subtype I-B

Box 2 | Structural composition of multiprotein crRNA–effector complexes



In type I and type III CRISPR–Cas systems, multiprotein CRISPR RNA (crRNA)–effector complexes mediate the processing and interference stages of the CRISPR defence system. In type I systems, this complex is known as the CRISPR-associated complex for antiviral defence (Cascade; see the figure, part **a**) complex, whereas in type III-A and type III-B systems the complexes are respectively known as Csm and Cmr (see the figure, part **b**) complexes. A common structural feature among the Cas proteins found in crRNA–effector complexes is the RNA recognition motif (RRM), a nucleic acid-binding domain that is the core fold of the extremely diverse RAMP protein superfamily[4,32,34]. The RAMPs Cas5 and Cas7 comprise the skeleton of the crRNA–effector complexes. In type I systems, Cas6 is typically the active endonuclease that is responsible for crRNA processing, and Cas5 and Cas7 are non-catalytic RNA-binding proteins; however, in type I-C systems, crRNA processing is catalysed by Cas5 (REF. 55). In type III systems, the enzyme that is responsible for processing has not been directly identified but is generally assumed to be Cas6 (REFS 38–40; however, Cas6 is not a subunit of the effector complex in these systems, and in some cases is provided *in trans* by other CRISPR–Cas loci), whereas Cas7 is involved in co-transcriptional RNA degradation during the interference stage[26].

In addition to Cas5, Cas6 and Cas7, crRNA–effector complexes typically contain two proteins that are designated, according to their size, the large subunit and the small subunit. The large subunit is present in all known type I and type III crRNA–effector complexes, whereas the small subunit is missing in some type I loci; a carboxy-terminal domain of the large subunit is predicted to functionally replace the small subunit in complexes where the small subunit is absent[33]. In type III systems, the large subunit is the putative cyclase-related enzyme encoded by *cas10*, whereas in type I systems the large subunit is encoded by diverse *cas8* genes that adopt a complex structure and show no readily detectable similarity to other proteins. Cas10 contains two cyclase-like Palm domains (a form of the RRM domain)[112,113], and the conservation of catalytic amino acid residues implies that one of these domains is active whereas the other is inactivated; the catalytic site of the active domain is required for cleavage of double-stranded DNA during interference[26], but its activity remains to be characterized in detail. Although it has been speculated that Cas8 is a highly derived homologue of Cas10 (REFS 4,33), and the similarity between the organizations of the types I and III crRNA–effector complexes is consistent with this possibility, sequence and structural comparisons fail to provide clear evidence. Some Cas8 proteins of subtype I-B have been shown to possess the single-stranded DNA-specific nuclease activity[114] required for interference[115]. However, whether such activity is a universal feature of the large subunit remains to be determined.

The small subunit proteins are encoded by *csm2* (subtypes III-A and III-D), *cmr5* (subtypes III-B and III-C), *cse2* (subtype I-E) or *csa5* (subtype I-A). They are α-helical proteins that have no detectable homologues, although a structural comparison suggests that the small subunit proteins of type I and III systems are homologous to one another[116].
Despite differences in structural details, the overall shapes and architectures of the Cascade[43,45,97], Cmr and Csm complexes[36,38,41,98,117] are remarkably similar, as can be seen from electron microscopy images of *Escherichia coli* Cascade complexes[31] (comprising Cas5, Cas6e and six Cas7 proteins, together with Cas8e as the large subunit and two Cse2 proteins as the small subunits; see the figure, part **c**) and *Thermus thermophilus* Cmr complexes[36] (comprising a Cas5 group protein known as Cmr3 and six Cas7 group proteins, namely Cmr1, Cmr6 and 4 copies of Cmr4, together with a Cas10 group protein known as Cmr2 as the large subunit and Cmr5 as the small subunit; see the figure, part **d**). This suggests that the ancestral multisubunit effector complex evolved before the divergence of type I and type III CRISPR–Cas systems.
Figure part **c** from REF. 31, Nature Publishing Group. Figure part **d** adapted with permission from REF. 36, Cell Press.

and at least 8 Cas8a families within subtype I-A (see Supplementary information S2 (table)). Thus, notwithstanding its complex evolution, we retain subtype I-B, which is best defined by the ancestral type I gene composition. The three main subdivisions within subtype I-B roughly correspond to the previously described subtypes Hmari, Tneap and Myxan[32] (see also TIGRFAM directory), and now could be defined through specific Cas8b families, Cas8b1 (for Hmari), Cas8b2 (for Tneap) and Cas8b3 (for Myxan), with a few exceptions.

A subset of subtype I-B systems defined by the presence of the *cas8b1* gene has been described as subtype I-G in the recent classification of archaeal CRISPR–Cas systems[35]. However, inclusion of bacterial CRISPR–Cas leads to increased diversity within subtype I-B so that if subtype I-G is recognized, consistency would require

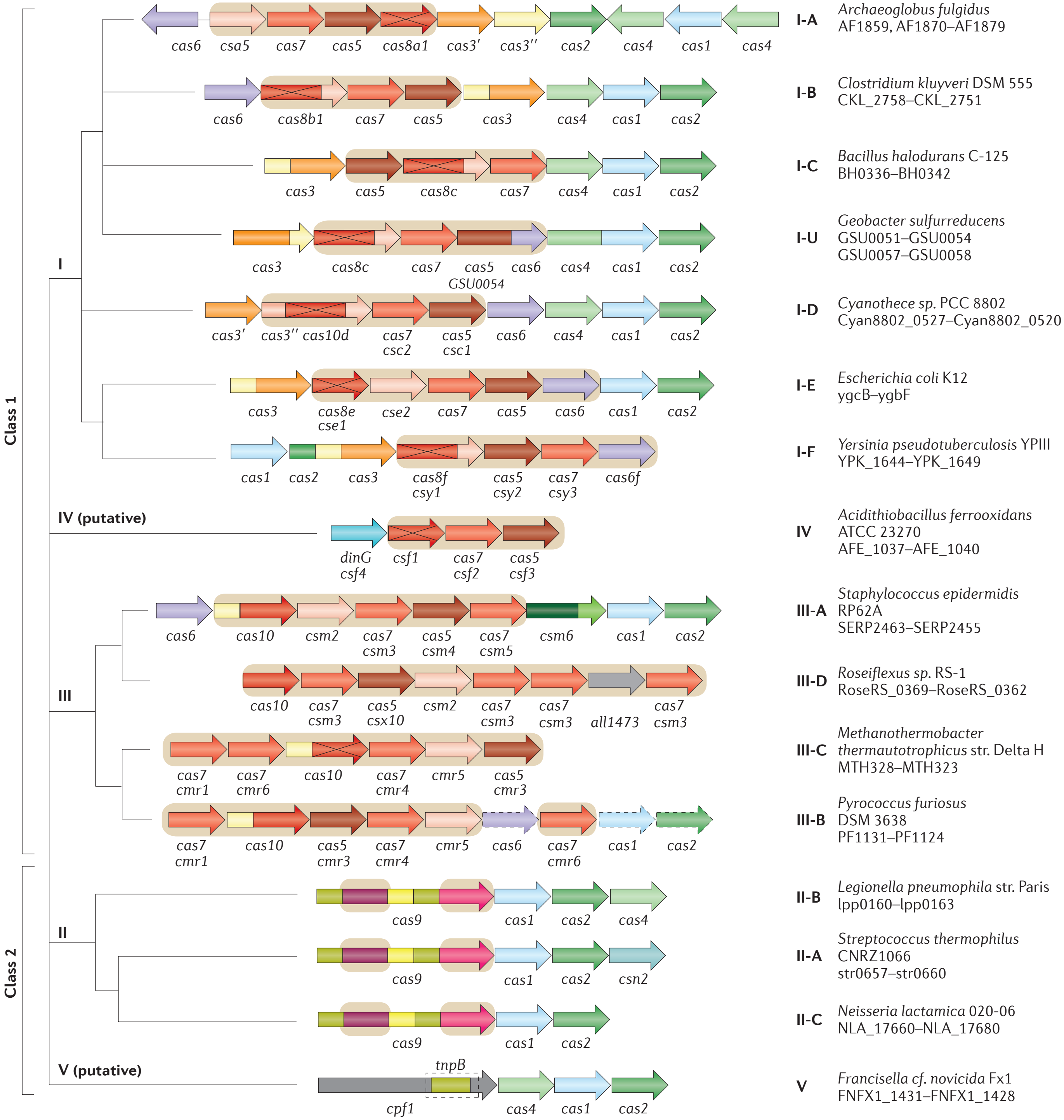

Class 1
I
cas6 csa5 cas7 cas5 cas8a1 cas3′ cas3″ cas2 cas4 cas1 cas4
I-A Archaeoglobus fulgidus AF1859, AF1870–AF1879
cas6 cas8b1 cas7 cas5 cas3 cas4 cas1 cas2
I-B Clostridium kluyveri DSM 555 CKL_2758–CKL_2751
cas3 cas5 cas8c cas7 cas4 cas1 cas2
I-C Bacillus halodurans C-125 BH0336–BH0342
cas3 cas8c cas7 cas5 cas6 cas4 cas1 cas2
GSU0054
I-U Geobacter sulfurreducens GSU0051–GSU0054 GSU0057–GSU0058
cas3′ cas3″ cas10d cas7 csc2 cas5 csc1 cas6 cas4 cas1 cas2
I-D Cyanothece sp. PCC 8802 Cyan8802_0527–Cyan8802_0520
cas3 cas8e cse1 cse2 cas7 cas5 cas6 cas1 cas2
I-E Escherichia coli K12 ygcB–ygbF
cas1 cas2 cas3 cas8f csy1 cas5 csy2 cas7 csy3 cas6f
I-F Yersinia pseudotuberculosis YPIII YPK_1644–YPK_1649
IV (putative)
dinG csf4 csf1 cas7 csf2 cas5 csf3
IV Acidithiobacillus ferrooxidans ATCC 23270 AFE_1037–AFE_1040
III
cas6 cas10 csm2 cas7 csm3 cas5 csm4 cas7 csm5 csm6 cas1 cas2
III-A Staphylococcus epidermidis RP62A SERP2463–SERP2455
cas10 cas7 csm3 cas5 csx10 csm2 cas7 csm3 cas7 csm3 all1473 cas7 csm3
III-D Roseiflexus sp. RS-1 RoseRS_0369–RoseRS_0362
cas7 cmr1 cas7 cmr6 cas10 cas7 cmr4 cmr5 cas5 cmr3
III-C Methanothermobacter thermautotrophicus str. Delta H MTH328–MTH323
cas7 cmr1 cas10 cas5 cmr3 cas7 cmr4 cmr5 cas6 cas7 cmr6 cas1 cas2
III-B Pyrococcus furiosus DSM 3638 PF1131–PF1124
Class 2
II
cas9 cas1 cas2 cas4
II-B Legionella pneumophila str. Paris lpp0160–lpp0163
cas9 cas1 cas2 csn2
II-A Streptococcus thermophilus CNRZ1066 str0657–str0660
cas9 cas1 cas2
II-C Neisseria lactamica 020-06 NLA_17660–NLA_17680
V (putative)
tnpB
cpf1 cas4 cas1 cas2
V Francisella cf. novicida Fx1 FNFX1_1431–FNFX1_1428

Figure 2 | **Architectures of the genomic loci for the subtypes of CRISPR–Cas systems.** Typical operon organization is shown for each CRISPR–Cas system subtype. For each representative genome, the respective gene locus tag names are indicated for each subunit. Homologous genes are colour-coded and identified by a family name. The gene names follow the classification from REF. 13. 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 either *csm2*, *cmr5*, *cse2* or *csa5*; no all-encompassing name has been proposed to collectively describe this gene family to date. Crosses through genes encoding the large subunit (Cas8 or Cas10 family members) indicate inactivation of the respective catalytic sites. Genes and gene regions encoding components of the interference module (CRISPR RNA (crRNA)–effector complexes or Cas9 proteins) are highlighted with a beige background. The adaptation module (*cas1* and *cas2*) and *cas6* are dispensable in subtypes III-A and III-B; in particular, they are rarely present in subtype III-B (dashed lines). Dark green denotes the CARF domain. Gene regions coloured cream represent the HD nuclease domain; the HD domain in Cas10 is distinct from that of Cas3 and Cas3″. Also coloured are the regions of *cas9* that roughly correspond to the RuvC-like nuclease (lime green), HNH nuclease (yellow), recognition lobe (purple) and protospacer adjacent motif (PAM)-interacting domains (pink). The regions of *cpf1* aside from the RuvC-like domain are functionally uncharacterized and are shown in grey, as is the functionally uncharacterized *all1473* gene in subtype III-D.

splitting I-B into several subtypes. Therefore, at present, we classify these variants within subtype I-B.

Subtype I-C seems to be a derivative of subtype I-B that lacks Cas6, which seems to be functionally replaced by Cas5 (REF. 55). Subtype I-A is another derivative of subtype I-B and is typically characterized by the fission of *cas8* into two genes that encode degraded large and small subunits, respectively, as well as fission of *cas3* into *cas3′* and *cas3′′*.

Subtype I-D also has several unique features, including Cas10d (instead of a Cas8 family protein) and a distinct variant of Cas3 (REF. 13) (FIG. 2; see Supplementary information S2,S4 (tables)). Subtype I-U is typified by the presence of an uncharacterized signature gene (*GSU0054*; TIGRFAM reference TIGR02165) and several other distinctive features that have been analysed in detail previously[33] (see Supplementary information S4 (table)). This group is monophyletic in the Cas3 tree and mostly monophyletic in the Cas1 tree (see Supplementary information S5,S6 (boxes)).

The phylogenetic tree of the type I signature protein Cas3′ (and the homologous region of Cas3) has been reported to accurately reflect the subtype classification[43], which is suggestive of a degree of evolutionary coherence between the phylogenies of the different genes in the operons of each subtype. However, re-analysis of the Cas3 phylogeny using a larger, more diverse sequence set (see Supplementary information S6 (box)) reveals a complex picture in which subtypes I-A, I-B and I-C are polyphyletic (that is, not descended from a common ancestor). Conceivably, this discrepancy results from a combination of accelerated evolution of many Cas3 variants and horizontal gene transfer.

In addition to the complete type I CRISPR–*cas* loci, analysis of sequenced genomes has revealed a variety of putative type I-related operons that encode effector complexes but are not associated with *cas1, cas2* or *cas3* genes and are only in some cases adjacent to CRISPR arrays (see Supplementary information S4 (table)). These solo effector complexes are often encoded on plasmids and/or associated with transposon-related genes. Many of these operons are derivatives of subtype I-F, whereas others are derivatives of subtype I-B (see Supplementary information S4,S7 (tables)). Some of the genomes that have these incomplete type I systems encode Cas1–Cas2 as parts of other CRISPR–*cas* loci but others lack these genes altogether (see Supplementary information S7 (table)). The functionality of solo effector complexes has not been investigated.

Type III CRISPR–Cas systems. All type III systems possess the signature gene *cas10*, which encodes a multidomain protein containing a Palm domain (a variant of the RNA recognition motif (RRM)) that is homologous to the core domain of numerous nucleic acid polymerases and cyclases and that is the largest subunit of type III crRNA–effector complexes (BOX 2). Cas10 proteins show extensive sequence variation among the diverse type III CRISPR–Cas systems, which means that several PSSMs are required to identify these loci. All type III loci also encode the small subunit protein (see below), one Cas5 protein and typically several paralogous Cas7 proteins (FIG. 1). Often, Cas10 is fused to an HD family nuclease domain that is distinct from the HD domains of type I CRISPR–Cas systems and, unlike the latter, contains a circular permutation of the conserved motifs of the domain[34,56].

Type III systems have been previously classified into two subtypes, III-A (previously known as Mtube subtype or Csm module) and III-B (previously known as Cmr module or RAMP module), that can be distinguished by the presence of distinct genes encoding small subunits, *csm2* (in the case of subtype III-A) and *cmr5* (in the case of subtype III-B) (FIG. 2; see Supplementary information S4 (table)). Subtype III-A loci usually contain *cas1*, *cas2* and *cas6* genes, whereas most of the III-B loci lack these genes and therefore depend on other CRISPR–Cas systems present in the same genome[4], providing strong evidence for the modularity of CRISPR–Cas systems[35] (FIG. 2). Both subtype III-A and subtype III-B CRISPR–Cas systems have been shown to co-transcriptionally target RNA[26,27,37–39,57] and DNA[26,58–61].

The composition and organization of type III CRISPR–*cas* loci are more diverse than those of type I systems — although there are fewer type III subtypes, each of these is more polymorphic than type I subtypes. This diversity is due to gene duplications and deletions, domain insertions and fusions, and the presence of additional, poorly characterized domains that could be involved either in crRNA–effector complex functions or in associated immunity. At least two type III variants (one from subtype III-A and one from subtype III-B) are common and are here upgraded to subtypes III-D and III-C, respectively, as proposed earlier for archaea[35] (FIG. 3; see Supplementary information S8 (table)). The distinctive feature of subtype III-C (previously known as MTH326-like[33]) is the apparent inactivation of the cyclase-like domain of Cas10 accompanied by extreme divergence of the sequence of this protein. Subtype III-D loci typically encode a Cas10 protein that lacks the HD domain. They also contain a distinct *cas5*-like gene known as *csx10* and often an uncharacterized gene that is homologous to *all1473* from *Nostoc sp.* PCC 7120 (REF. 33). Both of these new subtypes lack *cas1* and *cas2* genes (FIG. 2) and accordingly are predicted to recruit adaptation modules *in trans*. The phylogeny of Cas10, the signature gene of type III CRISPR–Cas, is consistent with the subtype classification, with each subtype representing a distinct clade (see Supplementary information S9 (box)).

Putative type IV CRISPR–Cas systems. Several bacterial genomes contain putative, functionally uncharacterized type IV systems, often on plasmids, as can be typified by the AFE_1037-AFE_1040 operon in *Acidithiobacillus ferrooxidans* ATCC 23270. Similar to most subtype III-B loci, this system lacks *cas1* and *cas2* genes and is often not in proximity to a CRISPR array or, in many cases, is encoded in a genome that has no detectable CRISPR arrays (it might be more appropriate to denote the respective loci Cas systems rather than CRISPR–Cas). Type IV systems encode a predicted minimal multisubunit crRNA–effector complex that consists of a partially degraded large subunit, Csf1, Cas5 and — as a single copy — Cas7, and in some cases, a putative small subunit[33] (FIG. 1); *csf1* can serve as a signature gene for this system. The minimalist architecture of type IV loci is distinct from those of all type I and type III subtypes (FIG. 2; see Supplementary information S4 (table)), which together with the unique large subunit (Csf1) justifies their status as a new type.

There are two distinct variants of type IV CRISPR–Cas systems, one of which contains a DinG family helicase (REF. 62), and a second one that lacks DinG but typically contains a gene encoding a small α-helical protein, which is a putative small subunit [33]. Type IV systems could be mobile modules that, similar to subtype III-B systems, use crRNAs from different CRISPR arrays once these become available. This possibility is consistent with the occasional localization of type IV loci adjacent to CRISPR arrays, *cas6* genes and (less often) adaptation genes[35].

Class 2 CRISPR–Cas systems

Class 2 CRISPR–Cas systems are defined by the presence of a single subunit crRNA–effector module. This class includes type II CRISPR–Cas systems, as well as a putative new classification, type V.

Type II CRISPR–Cas systems. Type II CRISPR–Cas systems dramatically differ from types I and III, and are by far the simplest in terms of the number of genes. The signature gene for type II is *cas9*, which encodes a multidomain protein that combines the functions of the crRNA–effector complex with target DNA cleavage[25], and also contributes to adaptation[63,64]. In addition to *cas9*, all identified type II CRISPR–*cas* loci contain *cas1* and *cas2* (see REF. 65 for a detailed comparative analysis of type II systems) (FIG. 1) and most type II loci also encode a tracrRNA, which is partially complementary to the repeats within the respective CRISPR array[65–67].

The core of Cas9, which includes both nuclease domains and a characteristic Arg-rich cluster, most likely evolved from genes of transposable elements that are not associated with CRISPR[65]. Thus, owing to the significant sequence similarity between Cas9 and its homologues that are unrelated to CRISPR–Cas, Cas9 cannot be used as the only signature for identification of type II systems. Nevertheless, the presence of *cas9* in the vicinity of *cas1* and *cas2* genes is a hallmark of type II loci.

Type II CRISPR–Cas systems are currently classified into three subtypes, which were introduced in the previous classification (II-A and II-B)[13] or subsequently proposed on the basis of a distinct locus organization (II-C)[65,66,68] (FIG. 2; see Supplementary information S4 (table)). Subtype II-A systems include an additional gene, *csn2* (FIG. 2), which is considered a signature gene for this subtype. The long and short variants of Csn2 form compact clusters when superimposed over the Cas9 phylogeny and seem to correspond to two distinct variants of subtype II-A[65]. However, as with subtype I-B, we chose to keep these two variants within subtype II-A. It was recently shown that all four subtype II-A Cas proteins are involved in spacer acquisition[63].

Subtype II-B lacks *csn2* but includes *cas4*, which is otherwise typical of type I systems (FIG. 2). Moreover, subtype II-B *cas1* and *cas2* are more closely related to type I homologues than to subtype II-A, which is suggestive of a recombinant origin of subtype II-B[65]. Subtype II-C loci only have three protein-coding genes (*cas1*, *cas2* and *cas9*) and are the most common type II CRISPR–Cas system in bacteria[3,65,66]. A notable example of a subtype II-C system is the crRNA-processing-independent system found in *Neisseria meningitidis*[69] (BOX 1).

In the Cas9 phylogeny, subtypes II-A and II-B are monophyletic whereas subtype II-C is paraphyletic with respect to II-A (that is, subtype II-A originates from within II-C)[65]. Nevertheless, II-C was retained as a single subtype given the minimalist architecture of the effector modules shared by all II-C loci.

Putative type V CRISPR–Cas systems. A gene denoted *cpf1* (TIGRFAM reference TIGR04330) is present in several bacterial genomes and one archaeal genome, adjacent to *cas1*, *cas2* and a CRISPR array (for example, in the FNFX1_1431–FNFX1_1428 locus of *Francisella cf. novicida* Fx1)[70] (FIG. 2). These observations led us to putatively define a fifth type of CRISPR–Cas system, type V, which combines Cpf1 (the interference module) with an adaptor module (FIG. 1; see Supplementary information S4 (table)). Cpf1 is a large protein (about 1,300 amino acids) that contains a RuvC-like nuclease domain homologous to the respective domain of Cas9 and the TnpB protein of IS605 family transposons, along with putative counterparts to the characteristic Arg-rich region of Cas9 and the Zn finger of TnpB. However, Cpf1 lacks the HNH nuclease domain that is present in all Cas9 proteins[54,65]. Given the presence of a predicted single-subunit crRNA–effector complex, the putative type V systems are assigned to class 2 CRISPR–Cas. Some of the putative type V loci also encode Cas4 and accordingly resemble subtype II-B loci, whereas others lack Cas4 and are more similar in architecture to subtype II-C. Unlike Cas9, Cpf1 is encoded outside the CRISPR–Cas context in several genomes, and its high similarity with TnpB suggests that *cpf1* is a recent recruitment from transposable elements.

If future experiments were to show that these loci encode bona fide CRISPR–Cas systems and that Cpf1 is a functional analogue of Cas9, then these systems would arguably qualify as a novel type of CRISPR–Cas. Despite the overall similarity to type II CRISPR–Cas systems, the putative type V loci clearly differ from the established type II subtypes more than type II subtypes differ from each other, most notably in the distinct domain architectures of Cpf1 and Cas9. Furthermore, whereas type II systems are specific to bacteria, a putative type V system is present in at least one archaeon, *Candidatus* Methanomethylophilus alvus[35].

Rare, unclassifiable CRISPR–Cas systems

The classification of CRISPR–Cas systems outlined above covers nearly all of the CRISPR–*cas* loci identified in the currently sequenced archaeal and bacterial genomes

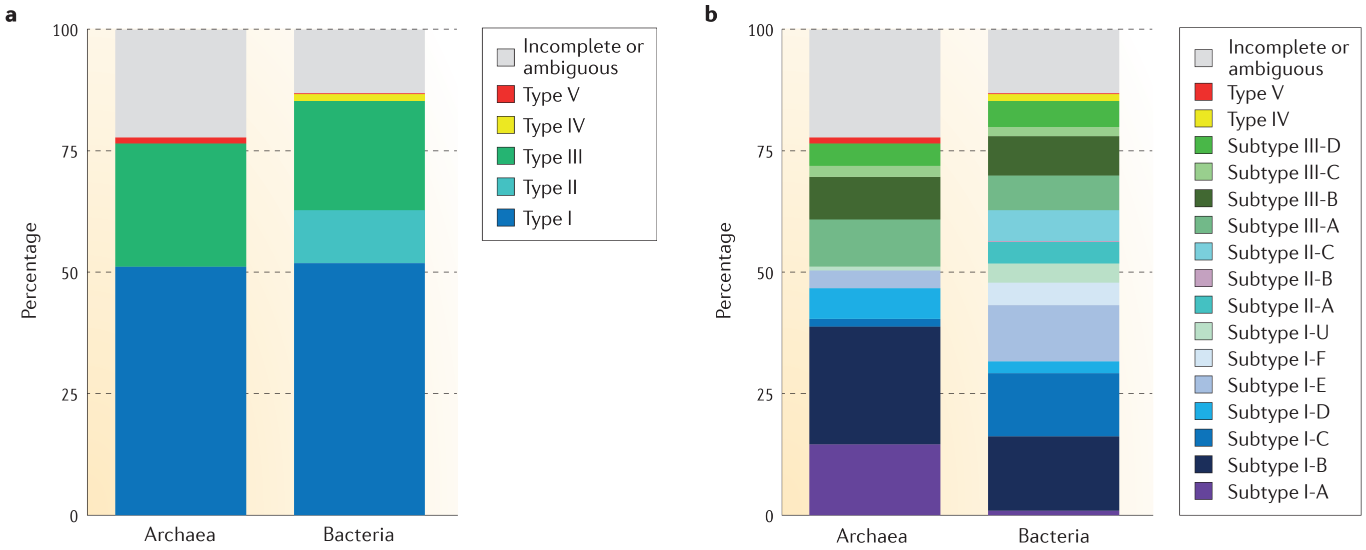


Figure 3 | **Distribution of CRISPR–Cas systems in sequenced archaeal and bacterial genomes. a** | Distribution by types. Chart showing the proportions of identified CRISPR–*cas* loci in bacterial or archaeal genomes that encode type I, type II, type III, type IV or type V CRISPR–Cas systems. The proportion of loci that encode incomplete systems or that we could not classify unambiguously is also shown. **b** | Distribution by subtypes. Chart showing the proportions of identified CRISPR–*cas* loci in bacterial or archaeal genomes that encode each of the subtypes of CRISPR–Cas systems included in the new classification described in this article. Note that type IV and V loci each encompass a single subtype. The proportion of loci that encode incomplete systems or that we could not classify unambiguously is also shown.

(FIG. 3). Nonetheless, owing to the rapid evolution of CRISPR–*cas* loci, which involves extensive recombination, it was not possible to account for all variants.

As a case in point, a putative CRISPR–Cas system was recently identified in *Thermococcus onnurineus*[71]. Based on some marginal similarities to protein components of crRNA–effector complexes, this locus was previously described as a Csf module[35], which here is classified as type IV. However, only the putative Cas7 protein from this locus (TON_0323) is most similar to the variant characteristic of type IV systems (Csf2), whereas Cas2 and Cas4 are uncharacteristic of type IV loci, and an uncharacterized large protein containing an HD domain is present instead of Csf1. These features suggest classification of the *T. onnurineus* as a derived type I system (notwithstanding the absence of the signature gene *cas3* or its variant *cas3′*), although it could not be assigned to any known subtype.

Several unusual variants of type III systems also posed a challenge for our classification. For example, the 15-gene locus in *Ignisphaera aggregans* has previously been classified as subtype III-D[35]. However, the III-D signature gene *csx10*, which encodes Cas5, is missing, and the other Cas proteins encoded by this locus show limited similarities to different type III subtypes[71]. Therefore, the *I. aggregans* locus seems to encode a type III system but cannot be unequivocally assigned to any subtype. Another distinct type III variant has been identified in several Crenarchaeota, primarily from the order Sulfolobales[35]. These loci lack detectable small subunits encoded by *csm2* or *cmr5* but contain a unique *cas* gene provisionally denoted *csx26*. Another variant is typified by the CRISPR–*cas* locus from *Thermotoga lettingae*[35], which is the only known type III system to encode a single Cas7 protein, a feature of type IV systems. These two type III variants share more similarity with subtype III-A than with other subtypes and are currently assigned to this subtype (see Supplementary information S9 (box)); however, subsequent analysis of new genomes along with experimental study might prompt their reclassification into separate subtypes.

This accumulation of unclassifiable variants suggests that the current approaches to CRISPR–Cas system classification will need to be further refined to cope with the challenge of ever increasing diversity.

Distribution in archaea and bacteria

Approximately 47% of analysed bacterial and archaeal genomes encode CRISPR–*cas* loci. As reported previously[13,72], CRISPR–Cas systems are much more prevalent in archaea (87% of genomes) than they are in bacteria (50% of genomes). For those genomes encoding CRISPR–*cas* loci, the rate of incomplete loci is similar for archaeal and bacterial genomes (17% and 12%, respectively). Complete single-unit loci are most commonly type I systems in both archaeal and bacterial genomes (64% and 60% of the loci, respectively), whereas putative type IV and type V systems are rare (<2% overall). Archaea possess significantly more type III systems than bacteria (34% versus 25% of the complete single-unit CRISPR–*cas* loci) but lack type II systems (13% in bacteria) (FIG. 3a). Thus, class 2 CRISPR–Cas systems are represented in archaea only by a single instance of the putative type V.

Overall, the most abundant CRISPR–Cas system is subtype I-B (20% of complete single-unit loci), followed by subtypes I-C and I-E (13% and 12%, respectively). In archaea, subtype I-A is the second most abundant after subtype I-B (18% and 30%, respectively), followed by subtypes III-A and III-B; subtype I-F is missing[35] (FIG. 3b). Among the three type II subtypes, subtypes II-C and II-A are the most abundant, comprising 7% and 5% of bacterial single-unit *cas* loci, respectively; subtype II-B is a minority, with only six loci that are restricted to Proteobacteria (0.3%). Finally, archaea encompass a significantly greater fraction of multi-unit loci than bacteria (14% versus 6%). Of the 13% of all CRISPR–*cas* loci that are incomplete or unclassified, 48% are partial type I loci and 25% are partial type III loci.

Different archaeal and bacterial phyla show distinct trends in the distribution of CRISPR–Cas systems (see Supplementary information S8 (table)). Notably, the Crenarchaeota lack subtypes I-B and I-C systems, which are abundant in other archaea and bacteria, whereas the Euryarchaeota are enriched in subtype I-B loci[35]. The Actinobacteria show a strong preference for subtype I-E systems, and the Cyanobacteria for subtype III-B systems, whereas the Firmicutes account for most of the subtype II-A systems. Finally, the Proteobacteria lack subtype I-A systems but are strongly enriched in subtype I-F loci. Considering the extraordinary importance of type II CRISPR–Cas systems in biotechnology, it is worth emphasizing that these systems represent a minority of CRISPR–*cas* loci. They also seem to be specific to bacteria and are significantly over-represented in the Proteobacteria and the Firmicutes.

We expect that the bias of available sequence data towards cultivable microorganisms, especially those of medical or biotechnological importance, affects the currently observed distribution of CRISPR–Cas systems. Nevertheless, the remarkable stability of the overall fraction of CRISPR-possessing microorganisms over several years of observation seems to imply that at least the main trends are captured by the present analysis.

Modular organization and evolution

Similarly to other defence systems, CRISPR–*cas* loci evolve under strong selection pressure exerted by changing pathogens, resulting in rapid evolution that is largely uncoupled from the evolution of the rest of the respective genomes. Here we examine the evolutionary relationships between different components of the CRISPR–Cas systems and put forward the concept of modular organization, with semi-independent evolution of each module.

cas loci and CRISPR arrays. For the purpose of comparative analysis of CRISPR–Cas systems, CRISPR arrays were predicted in all genomes using CRISPRfinder[73,74] following the procedure described in CRISPRmap[75] and CRISPRstrand[76]. For each of the 1,949 *cas* loci, the nearest CRISPR array was identified, which showed a natural cut-off of 530 base pairs for the distance between *cas* loci and proximal CRISPR arrays (Supplementary information S8 (table)). Using this cut-off, 1,484 *cas* loci (75%) were classified as adjacent to a CRISPR array, 383 loci (22%) were present in CRISPR-positive genomes but far from any array, and 82 loci (54 complete and 28 incomplete, 3% total) were present in CRISPR-negative genomes. Although, as expected, the fraction of *cas* loci in CRISPR-negative genomes was significantly higher for incomplete (6.5%) than complete (2.3%) *cas* loci (χ^2 test *P* value of 7×10^{-5}), the existence of complete *cas* loci that were not accompanied by a recognizable CRISPR array anywhere in the genome was notable, as it defies the principle that crRNA–effector complexes are universally associated with CRISPR immunity. These CRISPR-less loci could be remnants of recently inactivated CRISPR–Cas systems or might function in a different way to the characterized CRISPR–Cas systems.

Conversely, of the 4,210 detected CRISPR arrays, 1,382 (33%) are adjacent (within 530 base pairs) to a *cas* locus, 2,365 arrays (56%) are located outside of *cas* loci in *cas*-positive genomes, and the remaining 463 arrays (11%) are orphans, present in genomes without detected *cas* loci. The orphan CRISPR arrays are probably remnants of formerly functional CRISPR–Cas systems.

CRISPR arrays are themselves classified into 18 structural families and 24 sequence families (only 23 were used here because one family could not be associated with any *cas* loci in our dataset), including unclassified repeats[75–77]. Both structural and sequence families of CRISPR show significant preferential association with particular types and subtypes of *cas* loci, although in most cases associations with other types or subtypes can also occur (FIG. 4; see Supplementary information S3,S10 (boxes)).

CRISPR–Cas systems and the species tree. Defence systems of bacteria and archaea evolve under extreme selection pressure from pathogens, particularly viruses, often using non-classic evolutionary processes, such as the seemingly Lamarckian adaptations represented by spacer integrations in CRISPR arrays[78], the partially selfish mode of reproduction in which toxin–antitoxin systems are maintained in the genome through their addictive properties[79], and pervasive horizontal gene transfer[72,80]. In line with these trends, evidence of extensive horizontal transfer of CRISPR–*cas* loci has been reported[8,13,34,81–83].

To quantify the propensity of CRISPR–Cas systems to evolve via horizontal — as opposed to vertical — transmission, we compared various system features with a provisional species tree of bacteria and archaea that was reconstructed from concatenated ribosomal protein alignments[84]. As expected, the classification of the *cas* loci showed only weak consistency with the species tree (FIG. 4). The association between the species tree and CRISPR repeat types was also weak for both structure-based and sequence-based repeat classification (FIG. 4; see Supplementary information S11 (table)). These observations quantitatively show that horizontal transfer dominates the evolution of CRISPR–*cas* loci.

Cas1 phylogeny, CRISPR–Cas classification and architecture of* cas *loci. We examined the key evolutionary trends of the CRISPR–Cas systems in connection with the classification outlined above. Cas1 is the most conserved Cas protein, in terms of both representation in CRISPR–*cas* loci and amino acid sequence conservation[85], and the Cas1 phylogeny generally correlates with the organization of CRISPR–*cas* loci[13]. Thus, until recently, Cas1 has been considered to be the signature of the presence of CRISPR–Cas systems in a genome[13,32,34]. However, in this analysis we identified 86 genomes containing complete (and by inference, functional) effector modules but that lacked *cas1*. These include genomes encoding the putative type IV systems, most subtype III-B, III-C and III-D systems and rare variants of subtypes I-C and I-F; 14 of these genomes also lack readily identifiable CRISPR arrays (FIG. 2; see Supplementary information S7 (table)).

Conversely, in some archaea and bacteria *cas1* genes are located outside CRISPR–*cas* loci[4], often within predicted self-synthesizing transposable elements dubbed casposons[86]. Casposon-encoded Cas1 proteins probably function as integrases that mediate the mobility of these transposons. The discovery of casposons suggests that the CRISPR–Cas adaptive immunity system arose from the insertion of a casposon near an innate immunity locus that encoded an effector complex[87].

Of the 1,949 CRISPR–*cas* loci analysed, 1,404 encompass at least one *cas1* gene. We constructed a phylogenetic tree of all 1,418 Cas1 sequences (some composite loci contain at least two *cas1* genes) and rooted the tree using the modified midpoint procedure (FIG. 5; see Supplementary information S5 (box)). Mapping CRISPR–*cas* loci onto the Cas1 tree (FIG. 5) demonstrates a considerable agreement between the phylogeny of Cas1 and locus types and subtypes, consistent with previous observations. Thus, *cas1* genes of subtypes I-E, I-F, II-B and putative type V are strictly monophyletic, and *cas1* genes of subtypes I-C, I-U and II-A are largely monophyletic, with a few exceptions. In addition, *cas1* genes of subtypes II-A and II-C form a mostly homogeneous clade, in agreement with a previous analysis[65]. By contrast, *cas1* genes from the other type I subtypes and type III loci are scattered across the tree, suggestive of primarily horizontal evolution[13,34,35,88]. Thus, although substantial recombination occurs between the adaptation module and the other modules of the *cas* loci, the combination of the adaptation module with other modules is far from random.

As expected, the phylogeny of Cas1 is a poor match to the species tree of archaea and bacteria. The correlation of the distances between species with those between the corresponding *cas1* genes in the tree is much weaker than the correlation between the Cas1 phylogeny and CRISPR–*cas* locus classification (FIG. 4; see Supplementary information S11 (table)). These observations imply an extensive history of horizontal transfers, many of which involved complete CRISPR–*cas* loci, whereas a smaller number included the adaptation module alone.

Cas1 is crucial to the adaptation stage of the CRISPR-mediated immune response[17,89] and thus could be expected to co-evolve with CRISPR arrays[83,88]. We mapped structure-based and sequence-based repeat classification of CRISPR arrays adjacent to *cas* loci to the Cas1 tree. When only fully classified CRISPR repeats are considered, a high degree of consistency is observed between the Cas1 tree topology and repeat classification (FIG. 4; see Supplementary information S11 (table)), which probably reflects the direct recognition of repeats by Cas1 and its mechanistic involvement in the formation of the CRISPR arrays[89].

We also developed a quantitative measure to compare the architectures of the *cas* loci to one another and to generate a similarity dendrogram (see Supplementary information S12 (box)). Overall, the topology of the dendrogram is consistent with the subtype classification of CRISPR–Cas systems (FIG. 4; see Supplementary information S11 (table)). However, the clusters obtained by this method are much narrower than the respective subtypes, which is consistent with a frequent rearrangement of CRISPR–Cas loci. By contrast, clusters obtained from protein similarity searches, using proteins from the interference module, are broader and often directly correspond to individual subtypes (see Supplementary information S12,S13 (boxes)). As expected, the clustering of CRISPR–Cas systems by locus architecture is substantially more compatible with the Cas1 phylogeny than with the species tree (FIG. 4), in agreement with the considerable evolutionary coherence of the CRISPR–Cas systems despite frequent horizontal gene transfer of CRISPR–*cas* loci and of individual modules.

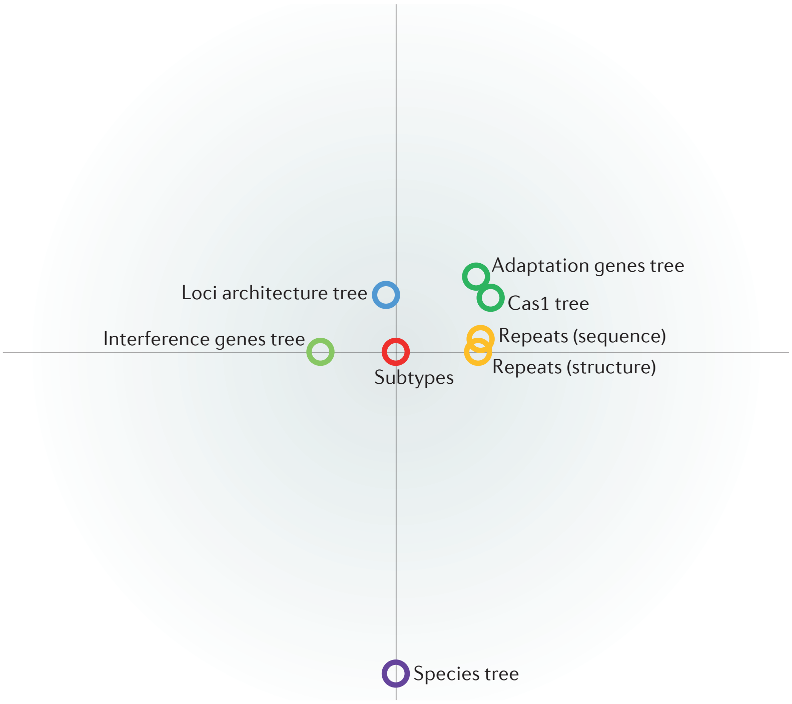


Figure 4 | **Comparison of different classifications of CRISPR–Cas systems.** This graph shows the strength of correlation between the new classification of CRISPR–Cas systems described here ('subtypes'; in the centre of the graph) and other classification measures. 'Interference genes tree' represents a phylogeny of interference module genes, which encode multisubunit CRISPR RNA (crRNA)–effector complexes or Cas9 proteins. This tree was created using a simple clustering approach based on aggregate protein sequence similarity. 'Adaptation genes tree' represents clustering produced by the same method but based on both components of the adaptation module, Cas1 and Cas2. 'Cas1 phylogeny' is the phylogenetic tree of Cas1 proteins shown in FIG. 5. 'Loci architecture tree' represents clustering based on a quantitative measure we developed to compare the architectures of CRISPR–*cas* loci. The measure is based on a weighted similarity index of the order of *cas* genes. 'Repeats (sequence)' denotes the classification of CRISPR sequences into 24 families on the basis of sequence similarity. 'Repeats (structure)' denotes the classification of CRISPR sequences into 18 families on the basis of structural similarity. The species tree represents the phylogeny of bacterial and archaeal translation systems. The distances depicted are inversely proportional to the degree of similarity. The full similarity matrix is shown in Supplementary information S11 (table).

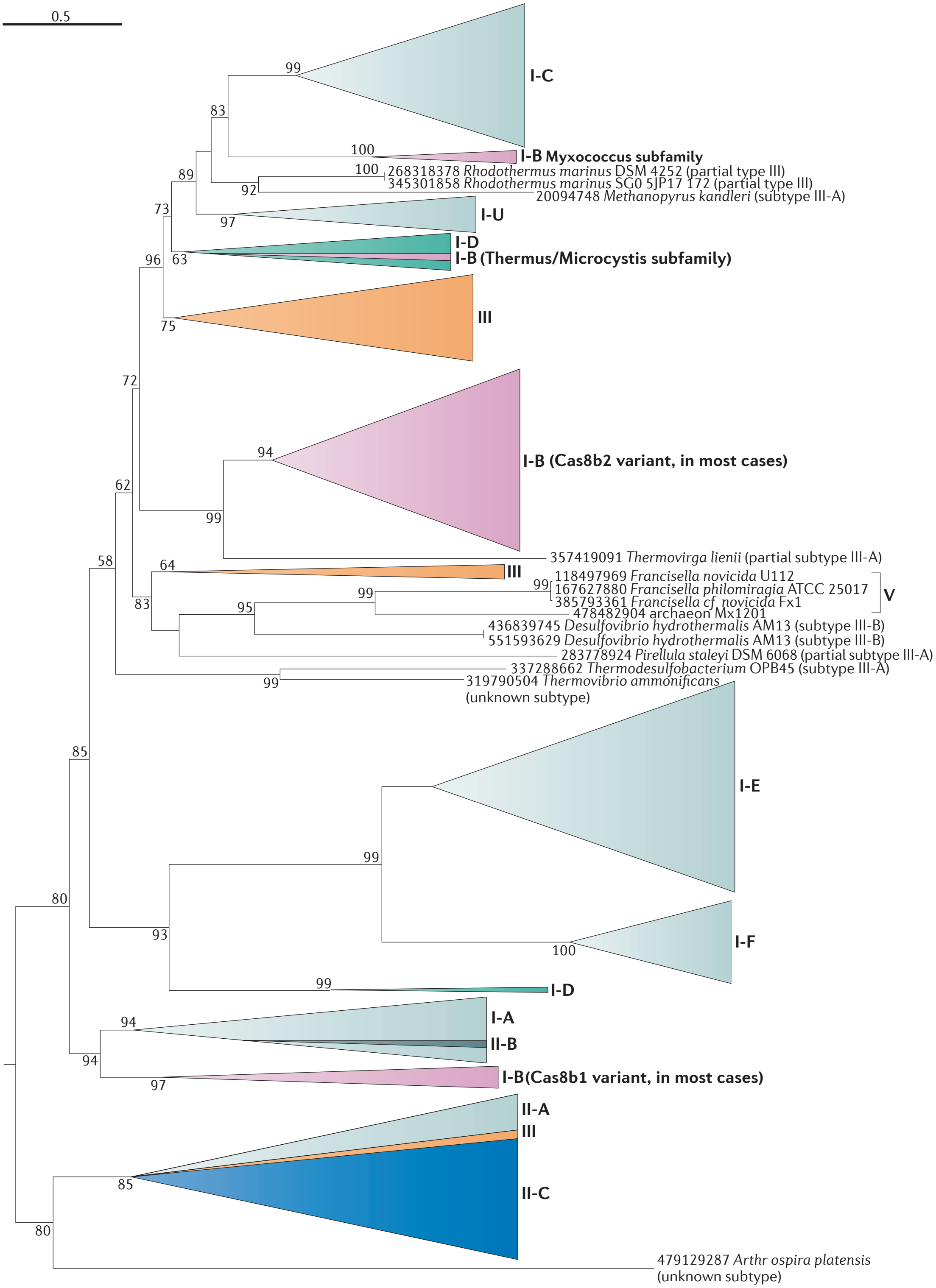

0.5
99
I-C
83
100
I-B Myxococcus subfamily
100
268318378 *Rhodothermus marinus* DSM 4252 (partial type III)
345301858 *Rhodothermus marinus* SG0 5JP17 172 (partial type III)
89
92
20094748 *Methanopyrus kandleri* (subtype III-A)
73
97
I-U
I-D
I-B (Thermus/Microcystis subfamily)
96
63
III
75
72
94
I-B (Cas8b2 variant, in most cases)
62
99
357419091 *Thermovirga lienii* (partial subtype III-A)
58
64
III
99
118497969 *Francisella novicida* U112
167627880 *Francisella philomiragia* ATCC 25017
385793361 *Francisella cf. novicida* Fx1
V
99
478482904 archaeon Mx1201
83
95
436839745 *Desulfovibrio hydrothermalis* AM13 (subtype III-B)
551593629 *Desulfovibrio hydrothermalis* AM13 (subtype III-B)
283778924 *Pirellula staleyi* DSM 6068 (partial subtype III-A)
337288662 *Thermodesulfobacterium* OPB45 (subtype III-A)
99
319790504 *Thermovibrio ammonificans*
(unknown subtype)
85
I-E
99
80
93
100
I-F
99
I-D
94
I-A
II-B
94
97
I-B(Cas8b1 variant, in most cases)
II-A
III
85
II-C
80
479129287 *Arthr ospira platensis*
(unknown subtype)

Figure 5 | **Mapping of the CRISPR–Cas classification onto the phylogenetic tree of Cas1.** Subtypes from the new classification of CRISPR–Cas systems described here were mapped onto a sequence-based phylogenetic reconstruction of 1,418 proteins from the Cas1 family, which is the most conserved Cas protein family. The phylogeny shows a close agreement with the subtype classification, as subtypes I-A, I-C, I-E, I-F, I-U, II-A, II-B, and putative type V are mostly or strictly monophyletic and are shown in gradients of light grey, except for II-B, which is shown in dark grey to indicate its origin from within I-A. The more discordant distribution of Cas1 for other subtypes probably results from horizontal transfer. None of the type III subtypes is monophyletic (in contrast to the Cas10 tree shown in Supplementary information S9 (box)), and so type III subtypes are not indicated. Note that Cas1 is absent in type IV loci and so these putative CRISPR–Cas systems are not shown. Triangles denote multiple collapsed branches. Individual genes are labelled with species names and gene identification numbers. Bootstrap values are indicated as percentage points; values below 50% are not shown.

Automated annotation of CRISPR–*cas* loci

Given the rapid pace of microbial genome sequencing, tools for the automated annotation of CRISPR–*cas* locus subtypes in newly sequenced genomes would be highly valuable. Although a careful inspection of combined features is required for accurate subtype annotation, we investigated whether an automated annotation method based on the similarity of the protein sequences of interference modules can faithfully reproduce the existing locus annotation.

To assess the value of the interference module as a proxy for the distribution of CRISPR–*cas* loci in our classification, we adopted a simple clustering approach based on aggregate sequence protein similarity[35]. This approach was chosen because of the lack of a universal marker suitable for phylogenetic analysis, as there is great variability in gene composition and module architecture between subtypes. The resulting cluster dendrogram (see Supplementary information S13 (box)) showed a high correlation with the subtype classification (FIG. 4; see Supplementary information S10 (box)). A similar cluster dendrogram constructed for Cas1 and Cas2 (see Supplementary information S14 (text)) showed a strong correlation with the Cas1 phylogeny but a considerably weaker correlation with the classification and architecture of CRISPR–*cas* loci than observed for the crRNA–effector complex dendrogram (FIG. 4; see Supplementary information S11 (table)). This difference supports our rationale in classifying CRISPR–*cas* loci on the basis of the interference module rather than Cas1 and demonstrates the ability of interference module protein clustering to closely reflect the new classification.

Having established the strong agreement between the clustering of interference module proteins and our classification, we constructed an automated classifier using prior information on the association between sequence PSSMs and CRISPR–*cas* loci and the corresponding classification of the effector modules. The classifier achieved 0.998 accuracy, which means that only 4 of 1,942 subtypes were incorrectly assigned (see Supplementary information S4,S15 (table, figure)). However, the accuracy of the method depends on the level of sequence similarity of the analysed Cas proteins to those available in the modelling phase, and predictably drops when the variants are only distantly related to the existing subtypes. Thus, the automated classifier described here has only limited applicability when annotating divergent variants of CRISPR–Cas subtypes.

Conclusions

The principal conclusion from the comparative analysis of the CRISPR–*cas* loci described here is the dynamic character and pronounced modularity of the evolution of this adaptive immunity system, which is conceivably driven by a perpetual arms race between the host genome and invading plasmids and viruses (dynamic evolution is a general theme in the evolution of defence systems[72,80]). In particular, the Cas1–Cas2 adaptation module evolved, to a large extent, independently of the operational modules (in particular, crRNA–effector complexes) of CRISPR–Cas systems, in agreement with the probable origin of the system as the result of the integration of a casposon-like mobile element next to an operon encoding a stand-alone effector complex[87]. The dynamic, modular evolution of CRISPR–Cas is also manifested at the level of the architecture of *cas* loci and the combination of different families of CRISPR arrays with different *cas* loci. However, a complementary trend is the frequent horizontal transfer of complete CRISPR–*cas* loci, which confers a degree of coherence to these systems and ensures that there is almost no congruence between the evolution of CRISPR–Cas and the species phylogeny as represented by the translation system[90].

The dynamic and modular character of CRISPR–Cas evolution hampers a straightforward classification based on evolutionary relationships. However, the classification approach we propose here, which combines signature genes with elements of the architecture of *cas* loci, assigned nearly all of the detected CRISPR–*cas* loci to specific subtypes. Furthermore, the resulting classification is largely compatible with the results of sequence-based clustering of crRNA–effector complexes, which can be adopted for automated classification of CRISPR–Cas systems from new genomes. The refinement of automated classification using more sophisticated machine learning and other computational techniques could lead to the development of fully automated classification of CRISPR–Cas systems.

In many respects, the new classification closely resembles the 2011 version[13], suggesting that the most common variants of CRISPR–Cas systems have already been discovered. However, we introduced a new top level, class, to account for the key differences between multisubunit and single-subunit crRNA–effector modules, as well as two new putative types (type IV and type V) and five new subtypes (II-C, III-C and III-D, together with the single subtypes of type IV and type V systems). Furthermore, the existence of currently unclassifiable variants implies that rare types and subtypes remain to be discovered and characterized, and the number of these is expected to substantially increase with the sequencing of new bacterial and archaeal genomes and metagenomes. In particular, the similarity between Cpf1 of the putative type V system and TnpB, which is usually found in transposons, suggests that multiple variants of single-subunit effector modules, and thus class 2 systems, might have evolved on independent occasions.

The classification of CRISPR–Cas systems and the principles of CRISPR–Cas evolution outlined here are expected to help the identification and focused discovery of new variants, some of which could become novel tools for genome engineering.

1. Deveau, H., Garneau, J. E. & Moineau, S. CRISPR/Cas system and its role in phage-bacteria interactions. *Annu. Rev. Microbiol.* **64**, 475–493 (2010).
2. Marraffini, L. A. & Sontheimer, E. J. CRISPR interference: RNA-directed adaptive immunity in bacteria and archaea. *Nat. Rev. Genet.* **11**, 181–190 (2010).
3. Koonin, E. V. & Makarova, K. S. CRISPR–Cas: evolution of an RNA-based adaptive immunity system in prokaryotes. *RNA Biol.* **10**, 679–686 (2013).
4. 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 (2013).
5. Barrangou, R. & Marraffini, L. A. CRISPR-Cas systems: prokaryotes upgrade to adaptive immunity. *Mol. Cell* **54**, 234–244 (2014).
6. Barrangou, R. *et al.* CRISPR provides acquired resistance against viruses in prokaryotes. *Science* **315**, 1709–1712 (2007).

7. Barrangou, R. CRISPR–Cas systems and RNA-guided interference. *Wiley Interdiscip. Rev. RNA* **4**, 267–278 (2013).
8. Westra, E. R. *et al.* The CRISPRs, they are a-changin': how prokaryotes generate adaptive immunity. *Annu. Rev. Genet.* **46**, 311–339 (2012).
9. Wiedenheft, B., Sternberg, S. H. & Doudna, J. A. RNA-guided genetic silencing systems in bacteria and archaea. *Nature* **482**, 331–338 (2012).
10. Garneau, J. E. *et al.* The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. *Nature* **468**, 67–71 (2010).
11. Magadán, A. H., Dupuis, M. E., Villion, M. & Moineau, S. Cleavage of phage DNA by the *Streptococcus thermophilus* CRISPR3–Cas system. *PLoS ONE* **7**, e40913 (2012).
12. van der Oost, J., Jore, M. M., Westra, E. R., Lundgren, M. & Brouns, S. J. CRISPR-based adaptive and heritable immunity in prokaryotes. *Trends Biochem. Sci.* **34**, 401–407 (2009).
13. Makarova, K. S. *et al.* Evolution and classification of the CRISPR–Cas systems. *Nat. Rev. Microbiol.* **9**, 467–477 (2011).
14. Westra, E. R., Buckling, A. & Fineran, P. C. CRISPR–Cas systems: beyond adaptive immunity. *Nat. Rev. Microbiol.* **12**, 317–326 (2014).
15. Sampson, T. R. & Weiss, D. S. CRISPR–Cas systems: new players in gene regulation and bacterial physiology. *Front. Cell. Infect. Microbiol.* **4**, 37 (2014).
16. Louwen, R., Staals, R. H., Endtz, H. P., van Baarlen, P. & van der Oost, J. The role of CRISPR–Cas systems in virulence of pathogenic bacteria. *Microbiol. Mol. Biol. Rev.* **78**, 74–88 (2014).
17. Nunez, J. K. *et al.* Cas1–Cas2 complex formation mediates spacer acquisition during CRISPR–Cas adaptive immunity. *Nat. Struct. Mol. Biol.* **21**, 528–534 (2014).
18. Yosef, I., Goren, M. G. & Qimron, U. Proteins and DNA elements essential for the CRISPR adaptation process in *Escherichia coli*. *Nucleic Acids Res.* **40**, 5569–5576 (2012).
19. Deveau, H. *et al.* Phage response to CRISPR-encoded resistance in *Streptococcus thermophilus*. *J. Bacteriol.* **190**, 1390–1400 (2008).
20. Shah, S. A., Erdmann, S., Mojica, F. J. & Garrett, R. A. Protospacer recognition motifs: mixed identities and functional diversity. *RNA Biol.* **10**, 891–899 (2013).
21. Bolotin, A., Quinquis, B., Sorokin, A. & Ehrlich, S. D. Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. *Microbiology* **151**, 2551–2561 (2005).
22. Mojica, F. J., Díez-Villaseñor, C., García-Martínez, J. & Almendros, C. Short motif sequences determine the targets of the prokaryotic CRISPR defence system. *Microbiology* **155**, 733–740 (2009).
23. Wang, R., Preamplume, G., Terns, M. P., Terns, R. M. & Li, H. Interaction of the Cas6 riboendonuclease with CRISPR RNAs: recognition and cleavage. *Structure* **19**, 257–264 (2011).
24. Deltcheva, E. *et al.* CRISPR RNA maturation by *trans*-encoded small RNA and host factor RNase III. *Nature* **471**, 602–607 (2011).
25. Jinek, M. *et al.* A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **337**, 816–821 (2012).
26. Samai, P. *et al.* Co-transcriptional DNA and RNA cleavage during type III CRISPR–Cas immunity. *Cell* **161**, 1164–1174 (2015).
27. Hale, C. R. *et al.* Essential features and rational design of CRISPR RNAs that function with the Cas RAMP module complex to cleave RNAs. *Mol. Cell* **45**, 292–302 (2012).
28. Sashital, D. G., Wiedenheft, B. & Doudna, J. A. Mechanism of foreign DNA selection in a bacterial adaptive immune system. *Mol. Cell* **46**, 606–615 (2012).
29. van Duijn, E. *et al.* Native tandem and ion mobility mass spectrometry highlight structural and modular similarities in clustered-regularly-interspaced short-palindromic-repeats (CRISPR)-associated protein complexes from *Escherichia coli* and *Pseudomonas aeruginosa*. *Mol. Cell Proteom.* **11**, 1430–1441 (2012).
30. Zhang, J. *et al.* Structure and mechanism of the CMR complex for CRISPR-mediated antiviral immunity. *Mol. Cell* **45**, 303–313 (2012).
31. Wiedenheft, B. *et al.* Structures of the RNA-guided surveillance complex from a bacterial immune system. *Nature* **477**, 486–489 (2011).
32. Haft, D. H., Selengut, J., Mongodin, E. F. & Nelson, K. E. A guild of 45 CRISPR-associated (Cas) protein families and multiple CRISPR/Cas subtypes exist in prokaryotic genomes. *PLoS Comput. Biol.* **1**, e60 (2005).
33. 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 (2011).
34. Makarova, K. S., Grishin, N. V., Shabalina, S. A., Wolf, Y. I. & Koonin, E. V. A putative RNA-interference-based immune system in prokaryotes: computational analysis of the predicted enzymatic machinery, functional analogies with eukaryotic RNAi, and hypothetical mechanisms of action. *Biol. Direct* **1**, 7 (2006).
35. Vestergaard, G., Garrett, R. A. & Shah, S. A. CRISPR adaptive immune systems of Archaea. *RNA Biol.* **11**, 156–167 (2014).
36. Staals, R. H. *et al.* Structure and activity of the RNA-targeting Type III-B CRISPR–Cas complex of *Thermus thermophilus*. *Mol. Cell* **52**, 135–145 (2013).
37. Spilman, M. *et al.* Structure of an RNA silencing complex of the CRISPR–Cas immune system. *Mol. Cell* **52**, 146–152 (2013).
38. Staals, R. H. *et al.* RNA targeting by the type III-A CRISPR–Cas Csm complex of *Thermus thermophilus*. *Mol. Cell* **56**, 518–530 (2014).
39. Tamulaitis, G. *et al.* Programmable RNA shredding by the type III-A CRISPR–Cas system of *Streptococcus thermophilus*. *Mol. Cell* **56**, 506–517 (2014).
40. Benda, C. *et al.* Structural model of a CRISPR RNA-silencing complex reveals the RNA-target cleavage activity in Cmr4. *Mol. Cell* **56**, 43–54 (2014).
41. Hale, C. R., Cocozaki, A., Li, H., Terns, R. M. & Terns, M. P. Target RNA capture and cleavage by the Cmr type III-B CRISPR–Cas effector complex. *Genes Dev.* **28**, 2432–2443 (2014).
42. van der Oost, J., Westra, E. R., Jackson, R. N. & Wiedenheft, B. Unravelling the structural and mechanistic basis of CRISPR–Cas systems. *Nat. Rev. Microbiol.* **12**, 479–492 (2014).
43. Jackson, R. N. Lavin, M., Carter, J., & Wiedenheft, B. Fitting CRISPR-associated Cas3 into the helicase family tree. *Curr Opin Struct Biol.* **24**, 106–114 (2014).
44. Mulepati, S., Heroux, A. & Bailey, S. Crystal structure of a CRISPR RNA-guided surveillance complex bound to a ssDNA target. *Science* **345**, 1479–1484 (2014).
45. Zhao, H. *et al.* Crystal structure of the RNA-guided immune surveillance Cascade complex in *Escherichia coli*. *Nature* **515**, 147–150 (2014).
46. Taylor, D. W. *et al.* Structures of the CRISPR–Cmr complex reveal mode of RNA target positioning. *Science* **348**, 581–585 (2015).
47. Mali, P., Esvelt, K. M. & Church, G. M. Cas9 as a versatile tool for engineering biology. *Nat. Methods* **10**, 957–963 (2013).
48. Sander, J. D. & Joung, J. K. CRISPR–Cas systems for editing, regulating and targeting genomes. *Nat. Biotech.* **32**, 347–355 (2014).
49. Altschul, S. F. & Koonin, E. V. PSI-BLAST — a tool for making discoveries in sequence databases. *Trends Biochem. Sci.* **23**, 444–447 (1998).
50. 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 (2011).
51. Gong, B. *et al.* Molecular insights into DNA interference by CRISPR-associated nuclease-helicase Cas3. *Proc. Natl Acad. Sci. USA* **111**, 16359–16364 (2014).
52. Huo, Y. *et al.* Structures of CRISPR Cas3 offer mechanistic insights into Cascade-activated DNA unwinding and degradation. *Nat. Struct. Mol. Biol.* **21**, 771–777 (2014).
53. Mulepati, S. & Bailey, S. Structural and biochemical analysis of nuclease domain of clustered regularly interspaced short palindromic repeat (CRISPR)-associated protein 3 (Cas3). *J. Biol. Chem.* **286**, 31896–31903 (2011).
54. Makarova, K. S. & Koonin, E. V. Annotation and classification of CRISPR–Cas systems. *Methods Mol. Biol.* **1311**, 47–75 (2015).
55. Nam, K. H. *et al.* Cas5d protein processes pre-crRNA and assembles into a cascade-like interference complex in subtype I-C/Dvulg CRISPR–Cas system. *Structure* **20**, 1574–1584 (2012).
56. Makarova, K. S., Aravind, L., Grishin, N. V., Rogozin, I. B. & Koonin, E. V. A DNA repair system specific for thermophilic Archaea and bacteria predicted by genomic context analysis. *Nucleic Acids Res.* **30**, 482–496 (2002).
57. Hale, C. R. *et al.* RNA-guided RNA cleavage by a CRISPR RNA–Cas protein complex. *Cell* **139**, 945–956 (2009).
58. Marraffini, L. A. & Sontheimer, E. J. CRISPR interference limits horizontal gene transfer in staphylococci by targeting DNA. *Science* **322**, 1843–1845 (2008).
59. Goldberg, G. W., Jiang, W., Bikard, D. & Marraffini, L. A. Conditional tolerance of temperate phages via transcription-dependent CRISPR–Cas targeting. *Nature* **514**, 633–637 (2014).
60. Deng, L., Garrett, R. A., Shah, S. A., Peng, X. & She, Q. A novel interference mechanism by a type IIIB CRISPR–Cmr module in *Sulfolobus*. *Mol. Microbiol.* **87**, 1088–1099 (2013).
61. Peng, W., Feng, M., Feng, X., Liang, Y. X. & She, Q. An archaeal CRISPR type III-B system exhibiting distinctive RNA targeting features and mediating dual RNA and DNA interference. *Nucleic Acids Res.* **43**, 406–417 (2015).
62. White, M. F. Structure, function and evolution of the XPD family of iron-sulfur-containing 5′→3′ DNA helicases. *Biochem. Soc. Trans.* **37**, 547–551 (2009).
63. Heler, R. *et al.* Cas9 specifies functional viral targets during CRISPR–Cas adaptation. **519**, 199–202 *Nature* (2015).
64. Wei, Y., Terns, R. M. & Terns, M. P. Cas9 function and host genome sampling in Type II-A CRISPR–Cas adaptation. *Genes Dev.* **29**, 356–361 (2015).
65. 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 (2014).
66. Chylinski, K., Le Rhun, A. & Charpentier, E. The tracrRNA and Cas9 families of type II CRISPR–Cas immunity systems. *RNA Biol.* **10**, 726–737 (2013).
67. Briner, A. E. *et al.* Guide RNA functional modules direct Cas9 activity and orthogonality. *Mol. Cell* **56**, 333–339 (2014).
68. Fonfara, I. *et al.* Phylogeny of Cas9 determines functional exchangeability of dual-RNA and Cas9 among orthologous type II CRISPR–Cas systems. *Nucleic Acids Res.* **42**, 2577–2590 (2014).
69. Zhang, Y. *et al.* Processing-independent CRISPR RNAs limit natural transformation in *Neisseria meningitidis*. *Mol. Cell* **50**, 488–503 (2013).
70. Schunder, E., Rydzewski, K., Grunow, R. & Heuner, K. First indication for a functional CRISPR/Cas system in *Francisella tularensis*. *Int. J. Med. Microbiol.* **303**, 51–60 (2013).
71. Makarova, K. S. *et al.* Dark matter in archaeal genomes: a rich source of novel mobile elements, defense systems and secretory complexes. *Extremophiles* **18**, 877–893 (2014).
72. Makarova, K. S., Wolf, Y. I. & Koonin, E. V. Comparative genomics of defense systems in archaea and bacteria. *Nucleic Acids Res.* **41**, 4360–4377 (2013).
73. Grissa, I., Vergnaud, G. & Pourcel, C. The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats. *BMC Bioinformatics* **8**, 172 (2007).
74. Bland, C. *et al.* CRISPR recognition tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. *BMC Bioinformatics* **8**, 209 (2007).
75. 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 (2013).
76. Alkhnbashi, O. S. *et al.* CRISPRstrand: predicting repeat orientations to determine the crRNA-encoding strand at CRISPR loci. *Bioinformatics* **30**, i489–i496 (2014).
77. Kunin, V., Sorek, R. & Hugenholtz, P. Evolutionary conservation of sequence and secondary structures in CRISPR repeats. *Genome Biol.* **8**, R61 (2007).
78. Koonin, E. V. & Wolf, Y. I. Is evolution Darwinian or/and Lamarckian? *Biol. Direct* **4**, 42 (2009).
79. Leplae, R. *et al.* Diversity of bacterial type II toxin-antitoxin systems: a comprehensive search and functional analysis of novel families. *Nucleic Acids Res.* **39**, 5513–5525 (2011).
80. Koonin, E. V. & Wolf, Y. I. Evolution of microbes and viruses: a paradigm shift in evolutionary biology? *Front. Cell Infect. Microbiol.* **2**, 119 (2012).
81. Godde, J. S. & Bickerton, A. The repetitive DNA elements called CRISPRs and their associated genes: evidence of horizontal transfer among prokaryotes. *J. Mol. Evol.* **62**, 718–729 (2006).

82. Almendros, C., Mojica, F. J., Díez-Villaseñor, C., Guzmán, N. M. & García-Martínez, J. CRISPR–Cas functional module exchange in *Escherichia coli*. *mBio* **5**, e00767–e00713 (2014).
83. Shah, S. A. & Garrett, R. A. CRISPR/Cas and Cmr modules, mobility and evolution of adaptive immune systems. *Res. Microbiol.* **162**, 27–38 (2011).
84. Yutin, N., Puigbo, P., Koonin, E. V. & Wolf, Y. I. Phylogenomics of prokaryotic ribosomal proteins. *PLoS ONE* **7**, e36972 (2012).
85. Takeuchi, N., Wolf, Y. I., Makarova, K. S. & Koonin, E. V. Nature and intensity of selection pressure on CRISPR-associated genes. *J. Bacteriol.* **194**, 1216–1225 (2012).
86. 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 Biol.* **12**, 36 (2014).
87. Koonin, E. V. & Krupovic, M. Evolution of adaptive immunity from transposable elements combined with innate immune systems. *Nat. Rev. Genet.* **16**, 184–192 (2015).
88. Garrett, R. A., Vestergaard, G. & Shah, S. A. Archaeal CRISPR-based immune systems: exchangeable functional modules. *Trends Microbiol.* **19**, 549–556 (2011).
89. Nunez, J. K., Lee, A. S., Engelman, A. & Doudna, J. A. Integrase-mediated spacer acquisition during CRISPR–Cas adaptive immunity. *Nature* **519**, 193–198 (2015).
90. Puigbo, P., Wolf, Y. I. & Koonin, E. V. Search for a 'Tree of Life' in the thicket of the phylogenetic forest. *J. Biol.* **8**, 59 (2009).
91. Hooton, S. P. & Connerton, I. F. *Campylobacter jejuni* acquire new host-derived CRISPR spacers when in association with bacteriophages harboring a CRISPR-like Cas4 protein. *Front. Microbiol.* **5**, 744 (2014).
92. Wiedenheft, B. *et al.* Structural basis for DNase activity of a conserved protein implicated in CRISPR-mediated genome defense. *Structure* **17**, 904–912 (2009).
93. Kwon, A. R. *et al.* Structural and biochemical characterization of HP0315 from *Helicobacter pylori* as a VapD protein with an endoribonuclease activity. *Nucleic Acids Res.* **40**, 4216–4228 (2012).
94. Makarova, K. S., Anantharaman, V., Aravind, L. & Koonin, E. V. Live virus-free or die: coupling of antivirus immunity and programmed suicide or dormancy in prokaryotes. *Biol. Direct* **7**, 40 (2012).
95. Beloglazova, N. *et al.* A novel family of sequence-specific endoribonucleases associated with the clustered regularly interspaced short palindromic repeats. *J. Biol. Chem.* **283**, 20361–20371 (2008).
96. Nam, K. H. *et al.* Double-stranded endonuclease activity in *Bacillus halodurans* clustered regularly interspaced short palindromic repeats (CRISPR)-associated Cas2 protein. *J. Biol. Chem.* **287**, 35943–35952 (2012).
97. Brouns, S. J. *et al.* Small CRISPR RNAs guide antiviral defense in prokaryotes. *Science* **321**, 960–964 (2008).
98. Rouillon, C. *et al.* Structure of the CRISPR interference complex CSM reveals key similarities with cascade. *Mol. Cell* **52**, 124–134 (2013). 99. Jinek, M. *et al.* Structures of Cas9 endonucleases reveal RNA-mediated conformational activation. *Science* **343**, 1247997 (2014).
100. Beloglazova, N. *et al.* Structure and activity of the Cas3 HD nuclease MJ0384, an effector enzyme of the CRISPR interference. *EMBO J.* **30**, 4616–4627 (2011).
101. Ramia, N. F. *et al.* Essential structural and functional roles of the Cmr4 subunit in RNA cleavage by the Cmr CRISPR–Cas complex. *Cell Rep.* **9**, 1610–1617 (2014).
102. Zhu, X. & Ye, K. Cmr4 is the slicer in the RNA-targeting Cmr CRISPR complex. *Nucleic Acids Res.* **43**, 1257–1267 (2015).
103. Brendel, J. *et al.* A complex of Cas proteins 5, 6, and 7 is required for the biogenesis and stability of clustered regularly interspaced short palindromic repeats (crispr)-derived rnas (crrnas) in *Haloferax volcanii*. *J. Biol. Chem.* **289**, 7164–7177 (2014).
104. Osawa, T., Inanaga, H., Sato, C. & Numata, T. Crystal structure of the CRISPR–Cas RNA silencing Cmr complex bound to a target analog. *Mol. Cell* **58**, 418–430 (2015).
105. Jung, T. Y. *et al.* Crystal structure of the Csm1 subunit of the Csm complex and its single-stranded DNA-specific nuclease activity. *Structure* **23**, 782–790 (2015).
106. Sapranauskas, R. *et al.* The *Streptococcus thermophilus* CRISPR/Cas system provides immunity in *Escherichia coli*. *Nucleic Acids Res.* **39**, 9275–9282 (2011).
107. Makarova, K. S., Anantharaman, V., Grishin, N. V., Koonin, E. V. & Aravind, L. CARF and WYL domains: ligand-binding regulators of prokaryotic defense systems. *Front. Genet.* **5**, 102 (2014).
108. Nam, K. H., Kurinov, I. & Ke, A. Crystal structure of clustered regularly interspaced short palindromic repeats (CRISPR)-associated Csn2 protein revealed Ca^{2+}-dependent double-stranded DNA binding activity. *J. Biol. Chem.* **286**, 30759–30768 (2011).
109. Koo, Y., Jung, D. K. & Bae, E. Crystal structure of *Streptococcus pyogenes* Csn2 reveals calcium-dependent conformational changes in its tertiary and quaternary structure. *PLoS ONE* **7**, e33401 (2012).
110. Arslan, Z. *et al.* Double-strand DNA end-binding and sliding of the toroidal CRISPR-associated protein Csn2. *Nucleic Acids Res.* **41**, 6347–6359 (2013).
111. Lee, K. H. *et al.* Identification, structural, and biochemical characterization of a group of large Csn2 proteins involved in CRISPR-mediated bacterial immunity. *Proteins* **80**, 2573–2582 (2012).
112. Zhu, X. & Ye, K. Crystal structure of Cmr2 suggests a nucleotide cyclase-related enzyme in type III CRISPR–Cas systems. *FEBS Lett.* **586**, 939–945 (2012).
113. Shao, Y. *et al.* Structure of the Cmr2–Cmr3 subcomplex of the Cmr RNA silencing complex. *Structure* **21**, 376–384 (2013).
114. Guy, C. P., Majernik, A. I., Chong, J. P. & Bolt, E. L. A novel nuclease-ATPase (Nar71) from archaea is part of a proposed thermophilic DNA repair system. *Nucleic Acids Res.* **32**, 6176–6186 (2004).
115. Cass, S. D. *et al.* The role of Cas8 in type I CRISPR interference. *Biosci. Rep.* **35**, e00197 (2015).
116. Reeks, J. *et al.* Structure of the archaeal Cascade subunit Csa5: relating the small subunits of CRISPR effector complexes. *RNA Biol.* **10**, 762–769 (2013).
117. Jackson, R. N. & Wiedenheft, B. A conserved structural chassis for mounting versatile CRISPR RNA-guided immune responses. *Mol. Cell* **58**, 722–728 (2015).

Acknowledgements
K.S.M., Y.I.W., D.H. and E.V.K. are supported by the National Institutes of Health (NIH) Intramural Research Program at the National Library of Medicine, US Department of Health and Human Services. R.M.T. and M.P.T. are supported by NIH grants RO1 GM54682 and RO1 GM99876. J.v.d.O. was partly supported by SIAM Gravitation Grant 024.002.002 from the Netherlands Organization for Scientific Research (N.W.O.). S.J.J.B. was financially supported by an NWO Vidi grant (864.11.005) and European Research Council (ERC) Stg (639707). A.F.Y. is supported by the Natural Sciences and Engineering Research Council (NSERC) Strategic Network Grant IBN and NSERC Discovery grant. S.M. acknowledges funding from Natural Sciences and Engineering Research Council of Canada (Discovery program) and holds a Tier 1 Canada Research Chair in Bacteriophages. F.J.M.M. is supported by the Ministerio de Economía y Competitividad (BIO2014-53029). R.B. is supported by the Deutsche Forschungsgemeinschaft (DFG) grant (BA 2168/5-2). S.A.S. and R.A.G. were funded primarily by the Danish Natural Science Research Council. O.S.A., F.C., S.J.S., R.B., S.A.S and R.A.G. are grateful to all members of the FOR1680 for helpful discussions.

Competing interests statement
The authors declare no competing interests.

FURTHER INFORMATION
TIGRFAM file directory: ftp://ftp.jcvi.org/pub/data/TIGRFAMs
TIGR02165 | TIGR04330

SUPPLEMENTARY INFORMATION
See online article: S1 (table) | S2 (table) | S3 (box) | S4 (table) | S5 (box) | S6 (box) | S7 (table) | S8 (table) | S9 (box) | S10 (box) | S11 (table) | S12 (box) | S13 (box) | S14 (box) | S15 (figure)

ALL LINKS ARE ACTIVE IN THE ONLINE PDF