An updated evolutionary classification of CRISPR–Cas systems


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 bacteria1–3 and provide sequence-specific protection against foreign DNA or, in some cases, RNA4. 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 interference5–11. 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 plasmids12,13,15. Recent studies suggest that CRISPR–Cas systems can also be used for non-defence roles, such as the regulation of collective behaviour and pathogenicity14–16.

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 arrays12,14. Protospacer acquisition in many CRISPR–Cas systems requires recognition of a short protospacer adjacent motif (PAM) in the target DNA14–21. 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 complex22 or via an alternative mechanism that involves bacterial RNase III and an additional RNA species, the tracrRNA (transactivating CRISPR RNA)23. 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 DNA10,13,24 or RNA16–21.

The rapid evolution of most cas genes23–24 and the remarkable variability in the genomic architecture of CRISPR–cas loci poses a major challenge for the
consistent annotation of Cas proteins and for the classification of CRISPR–Cas systems\textsuperscript{33,35}. 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\textsuperscript{31}. 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\textsuperscript{32}. Within each type of CRISPR–Cas system, several subtypes have been delineated based on additional signature genes and characteristic gene arrangements\textsuperscript{33,35}. 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\textsuperscript{16,33,34,36,37,38}. 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\textsuperscript{39,40}.

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.

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)\textsuperscript{41} 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\textsuperscript{31}. 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 (irrespective 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

Author addresses

1National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland 20894, USA.
2Bioinformatics group, Department of Computer Science, University of Freiberg, Georges-Kohler-Allee 106, 79110 Freiberg, Germany.
3Archea Centre, Department of Biology, Copenhagen University, Ole Maaløes Vej 5, DK2200 Copenhagen N, Denmark.
4Department of Food, Bioprocessing and Nutrition Sciences, North Carolina State University, Raleigh, North Carolina 27606, USA.
5Laboratory of Microbiology, Wageningen University, Dreijenplein 10, 6703HB Wageningen, Netherlands.
6Department of Regulation in Infection Biology, Helmholz Centre for Infection Research, D-38124 Braunschweig, Germany.
7DuPont Nutrition and Health, BP10, Dangé-Saint-Romain 68220, France.
8Département de Biochimie, de Microbiologie et de Bio-informatique, Faculté des Sciences et de Génie, Groupe de Recherche en Ecologie Buccale, Félix d’Hérelle Reference Center for Bacterial Viruses, Faculté de médecine dentaire, Université Laval, Québec City, Québec, Canada.
9Departamento de Fisiología, Genética y Microbiología, Universidad de Alicante, 03080-Alicante, Spain.
10Biochemistry and Molecular Biology, Genetics and Microbiology, University of Georgia, Davison Life Sciences Complex, Green Street, Athens, Georgia 30602, USA.
11Biomedical Sciences Research Complex, University of St Andrews, North Haugh, St Andrews, KY16 9TZ, UK.
12Department of Chemical Engineering and Applied Chemistry, University of Toronto, Toronto, M5S 3E5, Canada.
13BIOSS Centre for Biological Signalling Studies, Cluster of Excellence, University of Freiburg, Germany.
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 repeats17,29,31. The role of Cas2, which is a homologue of the mRNA interference toxins of numerous toxin–antitoxin systems, is less well understood17,29,34. 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 and DNase activities, its catalytic residues are dispensable for adaptation17, 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 complexes16,18,30,32,41–47,57,58 (BOX 2) or, in type II systems, by a single large protein, Cas9 (REFS 24,25,59). 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 a RNA endonuclease16 (typically Cas6; in type I and type III systems) or an alternative mechanism that involves RNase III and a translocating CRISPR RNA (tracrRNA)42 (in type II systems). However, in at least one type II CRISPR–Cas system, that of Neisseria meningitidis, crRNAs with mature 3’ ends are directly transcribed from internal promoters, and crRNA processing does not occur42.

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 systems17,33 by the combined action of the Cas7 and Cas10 proteins in type III systems18,30,32,103–105 by or by Cas9 in type II systems42. 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 interference101. 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 DNA113,106. 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 death107. Other notable components of the ancillary module include: a diverse set of proteins containing the CRISPR-associated Rossmann fold (CARF) domain107,108, which have been hypothesized to regulate CRISPR–Cas activity109 (in many type I and type II systems); and the inactivated P-loop ATPase Cas2, which forms a homotetrameric ring that accommodates linear double-stranded DNA in the central hole (in type II systems)106–111. Cas2 is not required for interference but apparently has a role in spacer integration, possibly preventing damage from the double-strand break in the chromosomal DNA110. Ancillary module genes are often found outside of CRISPR–cas loci, but the functions of these stand-alone genes have not been characterized in depth116,119.

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.
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. Often, the helicase domain is fused to a HD family endonuclease domain that is involved in the cleavage of the target DNA. 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. 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. 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. 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)14,15. 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 II-A and type II-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. The RAMP’s 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.

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. In type III systems, the large subunit is the putative helicase-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 helicase-like Pfam 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, but its activity remains to be characterized in detail. Although it has been speculated that Cas3 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 required for interference. However, whether such activity is a universal feature of the large subunit remains to be determined.

The small subunit proteins are encoded by Cmr2 (subtypes II-A and II-D), Cmr5 (subtypes II-B and II-C), Cas6e (subtype I-C) or Cas6f (subtype I-A). They are α-helical proteins that have detectable homologues, although a structural comparison suggests that the small subunit proteins of type I and III systems are homologous to one another.114 Despite differences in structural details, the overall shapes and architectures of the Cascade102,103, Cmr and Csm complexes101,102,103 are remarkably similar, as can be seen from electron microscopy images of Escherichia coli Cascade complexes115 (comprising Cas5, Cas6e and six Cas7 proteins, together with Cas8e as the large subunit and two Cas2 proteins as the small subunits; see the figure, part e) and Thermus thermophilus Cmr complexes116 (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. 56, 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, Tnep and Myxan (see also TIGRFAM directory), and now could be defined through specific Cas8b families, Cas8b1 (for Hmari), Cas8b2 (for Tnep) 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. 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...
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, cmr1 or csa2; 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 Cas37. 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.
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 permuted of the conserved motifs of the domain.

Putative type IV CRISPR-Cas systems. Several bacterial genomes contain putative, functionally uncharacterized type IV systems, often on plasmids, as can be typed 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
Class 2 CRISPR–Cas systems

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, and also contributes to adaptation. In addition to cas9, all identified type II CRISPR–Cas loci contain cas1 and cas2. 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. 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) or subsequently proposed on the basis of a distinct locus organization (II-C). Subtype II-A systems include an additional gene, for which a signature gene is proposed. The long and short variants of Csn2 form compact clusters when superimposed on the Cas9 phylogeny and seem to correspond to two distinct variants of subtype II-A. 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.

Subtype II-B lacks csn2, which is otherwise typical of type I systems. Moreover, subtype II-B lacks 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. Subtype II-C loci only have three protein-coding genes (cas1, cas2, and cas3) and are the most common type II CRISPR–Cas system in bacteria. A notable example of a subtype II-C system is the crRNA-processing-independent system found in Neisseria meningitidis.

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

Putative type V CRISPR–Cas systems. A gene denoted cpf1 in TIGR FAM reference (TIGRFAM) 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 novicida Fx177). 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 (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 IINH nuclease domain that is present in all Cas9 proteins. 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.

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 archaean and bacterial genomes.
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 onnurreneus*\(^3\). Based on some marginal similarities to protein components of crRNA–effector complexes, this locus was previously described as a Csf module\(^9\), 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 (Cas2), 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. onnurreneus* as a derived type I system (notwithstanding the absence of the signature gene cas3 or its variant *cas3*\(^3\)), 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\(^9\). 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\(^9\). 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\(^9\). These loci lack detectable small subunits encoded by csn2 or cmr5 but contain a unique cas gene provisionally denoted csx26. Another variant is typified by the CRISPR–cas locus from *Thermotoga lettingae*\(^6\), 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 59 (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\(^4,5\), CRISPR–Cas systems are much more prevalent in archaea (87% of genomes) than 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)\(^6\). 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).
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 \(cas\) 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\(^{35}\) and CRISPRstrand\(^{74}\). 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 (\(\chi^2\) test \(P\) value of \(7 \times 10^{-5}\)), 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\(^{29-37}\). 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\(^{45}\), the partially selfish mode of reproduction in which toxin–antitoxin systems are maintained in the genome through their addictive properties\(^{46}\), and pervasive horizontal gene transfer\(^{74,80}\). In line with these trends, evidence of extensive horizontal transfer of CRISPR–\(cas\) loci has been reported\(^{43,53,54,45}\).

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 alignment\(^{44}\). 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\(^{44}\), and the Cas1 phylogeny generally correlates with the organization of CRISPR–\(cas\) loci\(^{43}\). Thus, until recently, Cas1 has been considered to be the signature of
the presence of CRISPR–Cas systems in a genome. 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, often within predicted self-synthesizing transposable elements dubbed casposons. 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.

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. By contrast, cas1 genes from the other type I subtypes and type III loci are scattered across the tree, suggestive of primarily horizontal evolution. 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. 6; 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 and thus could be expected to co-evolve with CRISPR arrays. 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.

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 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. 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. 4c; 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. 4d; 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). 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 caspase-like mobile element next to an operon encoding a stand-alone effector complex.

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.

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, 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 CpfI of the putative type V system and TnphB, 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.
52
343
30
9
43
289
58
23
39
5
286
7
41
80

586
21
32
35
10
58

Acknowledgements
K.S.M., Y.I.W., D.H. E 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 R01 GM54682 and R01 GM58976. J.v.d.O. was partly supported by SIAM Gravitation Grant, 026.002.002 from the Netherlands Organization for Scientific Research (N.W.O.). S.J.J.B. was financially supported by an NWO Vidi grant (864.11.006) and European Research Council (ERC) StG (E37970). 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 Economia y Competitividad (BFU2014-55029-R). R.B. is supported by the Deutsche Forschungsgemeinschaft (DFG) grant (BA 2161/5-2). 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 FOR2180 for helpful discussions.

Competing interests statement
The authors declare no competing interests.

FURTHER INFORMATION
TIGRFAM file directory: http://tigr.umich.edu/krfram/

SUPPLEMENTARY INFORMATION
See online article: 31 (table), 32 (table), 33 (box), 34 (table), 35 (box), 36 (table), 37 (box), 38 (box), 39 (table), 40 (box and figure), 41 (table), 42 (figure).

ALL LINKS ARE ACTIVE IN THE ONLINE PDF.