CRISPR adaptive immune systems of Archaea

Vestergaard, Gisle Alberg; Garrett, Roger Antony; Shah, Shiraz Ali

Published in:
R N A Biology

DOI:
10.4161/rna.27990

Publication date:
2014

Document version
Publisher's PDF, also known as Version of record

Document license:
CC BY-NC

Citation for published version (APA):
CRISPR adaptive immune systems of Archaea

Gisle Vestergaard, Roger A Garrett & Shiraz A Shah

To cite this article: Gisle Vestergaard, Roger A Garrett & Shiraz A Shah (2014) CRISPR adaptive immune systems of Archaea, RNA Biology, 11:2, 156-167, DOI: 10.4161/rna.27990

To link to this article: https://doi.org/10.4161/rna.27990
CRISPR adaptive immune systems of Archaea

Gisle Vestergaard¹,², Roger A Garrett¹, and Shiraz A Shah¹,*

¹Archaean Centre; Department of Biology; University of Copenhagen; Copenhagen, Denmark; ²Molecular Microbial Ecology Group; Department of Biology; University of Copenhagen; Copenhagen, Denmark

Keywords: CRISPR, Cas, Type I, Type III, archaea

CRISPR adaptive immune systems were analyzed for all available completed genomes of archaea, which included representatives of each of the main archaeal phyla. Initially, all proteins encoded within, and proximal to, CRISPR-cas loci were clustered and analyzed using a profile–profile approach. Then cas genes were assigned to gene cassettes and to functional modules for adaptation and interference. CRISPR systems were then classified primarily on the basis of their concatenated Cas protein sequences and gene synteny of the interference modules. With few exceptions, they could be assigned to the universal Type I or Type III systems. For Type I, subtypes I-A, I-B, and I-D dominate but the data support the division of subtype I-B into two subtypes, designated I-B and I-G. About 70% of the Type III systems fall into the universal subtypes III-A and III-B but the remainder, some of which are phyla-specific, diverge significantly in Cas protein sequences, and/or gene synteny, and they are classified separately. Furthermore, a few CRISPR systems that could not be assigned to Type I or Type III are categorized as variant systems. Criteria are presented for assigning newly sequenced archaean CRISPR systems to the different subtypes. Several accessory proteins were identified that show a specific gene linkage, especially to Type III interference modules, and these may be cofunctional with the CRISPR systems. Evidence is presented for extensive exchange having occurred between adaptation and interference modules of different archaean CRISPR systems, indicating the wide compatibility of the functionally diverse interference complexes with the relatively conserved adaptation modules.

Introduction

CRISPR adaptive immune systems are present in most archaea and in many bacteria where they primarily target and degrade invading genetic elements. The immune reaction involves three primary stages. First, adaptation involving selection of 30–45 bp DNA fragments (protopacers) from an invading genetic element and their insertion between repeats of genomic CRISPR arrays as de novo spacers. Second, transcripts of CRISPR arrays are processed generally within the repeat sequences to yield small CRISPR RNAs (crRNAs). Third, crRNAs are assembled into protein interference complexes and guide the complex to matching sequences on nucleic acid(s) of the invading genetic element which are then cleaved.¹

Early classifications of CRISPR adaptive immune systems were based primarily on sequence analyses of CRISPR repeats or CRISPR-associated (Cas) proteins and yielded several distinct groupings.²-⁴ There is now a consensus that CRISPR systems can be classified structurally into major classes denoted Types I, II, and III for bacteria and Types I and III for archaea.³ The initial adaptation step is relatively conserved mechanistically among the three main CRISPR types and requires proteins Cas1, Cas2 and, in many systems, Cas4. Primary processing of CRISPR transcripts is accomplished with Cas6 in Type I and Type III systems and by a combination of RNase III and a tracrRNA in the bacterial Type II system.¹ The interference modules are much more varied with respect to both the number and sequences of the protein components involved, and the nucleic acid targets. Moreover, the high-sequence diversity of interference components, especially for the proteins labeled Cas7 and Cas8, provided a major obstacle to reaching a consensus about their classification. Despite this diversity, however, the three dimensional structures of the different interference complexes (denoted Cascade) appear to be partly conserved.⁶-⁸ Nevertheless, the mechanisms of interference are likely to be diverse, with Type I and Type II systems appearing to target primarily DNA while Type III interference systems can target DNA or RNA.⁹-¹³

Type I and Type III systems have been further classified into subtypes based primarily on sequences of signature Cas proteins Cas1, Cas3, and Cas8 for Type I and Cas10, and the small protein component S (protein S) for Type III systems, and on their gene synteny.¹⁴ However, there remain some limitations in the methods employed for identifying CRISPR subtypes. For example, categorizing systems according to the sequence of the larger conserved Cas1 protein is not always unambiguous and, moreover, Cas8, Cas10, and protein S are unsatisfactory signature proteins because of their variable size and high-sequence diversity.

Type III systems are found in most archaean phyla, and especially among extreme thermophiles, and their interference complexes are commonly encoded in gene cassettes that are not

*Correspondence to: Shiraz Shah; Email: sashah@bio.ku.dk
Submitted: 12/31/2013; Revised: 01/23/2014; Accepted: 01/24/2014
http://dx.doi.org/10.4161/rna.27990
linked genomically to either CRISPR loci or adaptation \textit{cas} gene cassettes.\cite{15} Earlier, we analyzed archaeal CRISPR-Cmr/Csm systems of archaea, based on Cas10 (Cmr2/Csm1) sequences, and concluded that there were five main families A, B, C, D, and E, of which the smallest and least well defined was family C.\cite{15,16} Subsequently, these were defined as Type III systems and were separated into the major subtypes III-A and III-B for bacteria and archaea,\cite{5} where subtypes III-A and III-B corresponded to the archaeal families E and B, respectively. Although the potential existence of families A and D was acknowledged by others earlier,\cite{4} they were omitted from the subsequent general classification.\cite{14}

Previously, there have been reports of other proteins being encoded within or adjacent to archaeal CRISPR-Cas cassettes,

![Figure 1. 16S rRNA phylogenetic tree for all archaeal genomes included in the study. The tree shows the primary archaeal phyla with the total number of genomes analyzed in each phylum indicated (single genomes for a given phylum are not numbered). The putative kingdoms are color-coded on the bar on the far left. Crenarchaeota, green; Aigarchaeota, light green; Korarchaeota, yellow; Thaumarchaeota, light blue; Euryarchaeota, dark blue. Environmental habitats are color-coded on vertical lines to the right of the kingdom bar: hydrothermal, orange; marine, blue; wetland sediment, brown, and hypersaline lake, pink. For each phylum, the Type I and Type III CRISPR subtypes are color-coded on the right and the total numbers of each subtype are given. The branch length ruler corresponds to a 5% difference in 16S RNA sequence.](image-url)
including mobile elements and toxin-antitoxin systems, and in a study of *Pyrococcus* species, evidence was presented for *herA* and *nurA* genes being specifically linked to *cas* genes. Here, we systematically analyzed the numbers and degree of specificity of all genes that are associated physically with archaeal *cas* gene cassettes, and identified 12 additional accessory protein families.

### Results

#### Analysis strategy

First, we downloaded the 159 archaeal genomes available in May 2013 (www.ebi.ac.uk/genomes/archaea.html). The phylogenetic diversity of the archaea and the number of sequenced genomes that were analyzed within each order are illustrated in the 16S rRNA-based phylogenetic tree (Fig. 1). All genomic loci-containing *cas* genes were first extracted automatically and the proteins were then clustered using Markov clustering. Subsequently, the loci were annotated manually aided by profile–profile searches against Conserved Domains (CDD) and TIGRFAMs (TF) databases. We then followed the proposal of Makarova et al., pertaining to the unification of Cas protein families involved in interference complexes of Type I and Type III systems throughout the analysis. Gene cassettes encoding adaptation and interference modules were extracted together with the *cas6* gene of the CRISPR RNA processing enzyme (Table S1, http://crispr.archaea.dk/TableS1.html). A significant fraction of the genetic modules were deficient or otherwise defective and they are also identified and tabulated (Table S1). Whereas 93% of gene cassettes encoding Type I adaptation and interference complexes were found to be linked to one another and to CRISPR loci, for Type III systems, only 55% of the *cas* gene cassettes and CRISPR loci were contiguous, and many interference complexes were encoded as separate genetic units (Table 1), consistent with them sharing adaptation modules and CRISPR loci with Type I systems.

Given that many of the CRISPR systems were non-integral, and that there is strong evidence for occurrence of exchange between adaptation and interference modules, we analyzed the two functional modules separately. First, we determined the operon structures and gene syntenies within each genetic unit to define the modules. Next, we prepared separate dendrograms for the adaptation and interference modules by comparison of sequences of all their protein components and adding up scores for each module, which resulted in module-to-module distance matrices, which were then converted to dendrograms using the neighbor joining method (Figs. S1 and S2, http://crispr.archaea.dk/FigureS1.pdf and http://crispr.archaea.dk/FigureS2.pdf, respectively). Subsequently, a detailed analysis of the different Type I and Type III subtypes was performed employing a combination of properties of the gene cassettes, and the protein components, of the interference modules.

Our criteria for defining subtypes were as follows. A vertical line is drawn near the base of the dendrogram in order to separate optimally the already established CRISPR subtypes, and this line is defined as the subtype threshold. Branches occurring before the line represent separate subtypes whereas branches initiating after the line belong to the same subtype. Exceptions were branches starting within the line. They were defined as separate subtypes if they showed consistent differences in gene synteny (as seen for subtypes I-G and V-2, see below), whereas if they showed similar gene syntenies they are inferred to represent divergent variants of the same subtype. These criteria can be readily applied to classifying newly identified CRISPR systems if a dendrogram is first prepared using the approach outlined above. On applying these criteria to the archaeal genomes, we generated a comprehensive, manually curated catalog of all the archaeal CRISPR systems. The catalog is presented in Table S1 and Figure S2 in a readily accessible form. It can easily be searched for individual organisms and genes, and researchers can both utilize and further analyze the data online. The results obtained were generally consistent with the current classification for Type I and Type III subtypes for bacteria and archaea, but a number of archaea-specific properties emerged, which are summarized below.

#### Adaptation modules

Initially, we generated a dendrogram of all the archaeal adaptation modules employing a combination of concatenated protein sequences and gene synteny of the adaptation Cas proteins of Type I and Type III systems (Table 2; Fig. S1). The threshold for defining adaptation subtypes was then determined to maximize the matches to the universal CRISPR subtypes, which were primarily Type I subtypes because most archaeal adaptation modules are linked genomically to Type I interference complexes.
useful for investigating the phenomenon of adaptation—interference exchange (see below).

Type I subtypes

Currently, all Type I systems of bacteria and archaea are classified into six subtypes I-A to I-F based on Cas1, Cas3, and Cas8 phylogeny and on the gene syntenies. Although this classification scheme has been largely successful, identification of newly sequenced CRISPR systems is not always straightforward, and in particular, employing Cas8 as a signature protein can be problematic. Often, newly sequenced CRISPR systems are not covered by existing Cas8 models. Furthermore, sequence matches to published Cas8 models are often ambiguous or misleading. For example, the Cas8a1_I-A model gives positive matches to Type I-B systems.

Nevertheless, we could assign most archaeal Type I systems to current subtypes using our criteria. The results show a strong archaeal bias to subtypes I-A, I-B, and I-D, with very few examples of subtypes I-C and I-E, and no I-F subtypes were found (Table 3). Our sequence analyses also support the division of the archaeal subtype I-B into distinct subtypes, designated here as subtype I-B and I-G, which correspond closely to the earlier proposed groupings Hmar and Tneap, respectively. Four variant Type I subtypes were identified as V_I-1, V_I-2, V_I-3, and V_I-4 that each occurred in low numbers and tended to be phyla-specific (Table 4). For each subtype, variations in gene order were sometimes observed but the same ORFs with similar sequences were still present. Typical subtype gene syntenies are illustrated on a simplified interference module dendrogram in Figure 3 and the complete archaeal interference dendrogram is presented in Figure S2.

The interference module of subtype I-D is quite distantly related to the other Type I subtypes. In the dendrogram, it lies at the junction of the Type I and Type III systems (Fig. 3; Fig. S2) and it carries a Cas10d protein that shows low but significant similarity to both Cas8 and Cas10 of Type I and Type III systems, respectively (data not shown). Therefore, we infer that subtype I-D may represent an intermediate between Type I and Type III interference modules.

The total numbers of each Type I subtype found associated with different archaeal phyla are indicated on the phylogenetic tree (Fig. 1) and are summarized for the main archaeal kingdoms in Table 3. The results show significant differences between the kingdoms. There is a strong bias to subtype I-A systems among euryarchaea, which lack most other subtypes except I-B (one example) and I-D (7 examples). In contrast, euryarchaea exhibit a more varied composition with multiple examples of subtypes I-A, I-B, I-G, and I-D, but with few instances of subtypes I-C and I-E.

Newly classified Type I subtypes

It is clear from the interference module dendrogram (Fig. S2) that the original subtype I-B contains independent groups, which branch off well before the subtype separation threshold, hence, it does not constitute a homogenous cluster. Therefore, we propose dividing subtype I-B into subtypes I-B and I-G (Fig. 3). Subtype I-G modules are also widespread among bacteria where, for example, they constitute the dominant Type I subtype in thermophilic Clostridium species. Cas8 sequences of the new I-B subtype match the CDD Cas8b_I-B model as predicted, while Cas8 sequences from I-G modules match Cas8a1_I-A, further supporting the division into subtypes I-B and I-G. Furthermore, the Cas8a1_I-A CDD model does not match any Type I-A Cas8 sequences, demonstrating that the model is unsuitable for correct identification of Type I subtypes.

The V_I-1 and V_I-2 variants branch off from subtypes I-A and I-G (Fig. 3; Fig. S2). Both variants carry interference modules with five genes, whereas I-A and I-G modules exhibit six and four genes, respectively. When compared with subtype I-G, the
additional genes in variants Vr-1 and Vr-2 appear to have resulted from partitioning of cas8 into cas8' and cas8". Despite the similarity in gene synteny, variants Vr-1 and Vr-2 are distinct in their amino acid sequences, diverging before the subtype definition threshold (Fig. 3; Fig. S2). Moreover, their Cas8 sequences are not covered by any CDD or Pfam models. However, given that there are so few members and that they appear to be archaea-specific, Vr-1 and Vr-2 are classified as variants rather than subtypes (Table 4).

Vr-3 was also reported earlier as a GSU0053-type module, and more recently, as a variant of subtype I-C. We were unable to confirm any link to subtype I-C and believe it represents an independent subtype, which should retain variant status until more examples are found in archaea and/or bacteria (Fig. 3).

*Methanocella conradii* contains a Type I interference module designated Vr-4, which despite yielding close protein sequence similarity to a group of I-D systems from cyanobacteria, carries a Cas8 protein which does not have an HD domain, a key structural signature of subtype I-D systems. However, the functional Cas3 protein does contain an HD domain and, therefore, Vr-4 may represent an intermediate between subtype I-D and other Type I subtypes (Figs. 3 and 4).

**Type III subtypes**

Earlier, five CRISPR families A to E were proposed for archaea based primarily on the divergent sequences of the Cas10 (Cmr2/Csm1) proteins. Later, Makarova et al. defined the universal subtype III-A corresponding to family E and subtype III-B equivalent to family B. Here, all archaeal genomic copies of Type III systems were identified many of which were limited to interference gene cassettes (Tables 1 and 2). Inspection of their positions within the interference dendrogram (Fig. S2) demonstrated that about 70% of the interference modules fell unambiguously into the current universal subtypes III-A or III-B, which are common to both crenarchaea and euryarchaea (Fig. 1).

Here we reclassify the remaining 30% of the subtypes into subtypes III-C (formerly family A) and III-D (formerly family D); the earlier family C was integrated into subtype III-B. Ten examples of an additional subtype were observed that were specific to the order Sulfolobales and the subtype is classified as variant subtype Vr-1 (Table 4). Typical gene cassettes for each subtype are illustrated in the simplified dendrogram (Fig. 4), and the complete Type III interference dendrogram is included in Figure S2. As for Type I subtypes, gene orders occasionally differ for a given subtype but the same ORFs with closely related sequences are retained. The distribution of Type III subtypes within the different archaeal phyla is illustrated in Figure 1 and their total numbers and kingdom distributions are summarized in Table 3 where subtypes III-C and III-D show strong kingdom biases.

**Newly classified Type III subtypes**

Subtype III-C corresponds to the Type III systems defined earlier as MTH-326-like Type III and archaenal family A. Eight examples of this subtype were found among archaea, and a further 29 were identified in bacteria (data not shown). Although the overall gene synteny of subtype III-C is similar to that of subtype III-B (Fig. 4), the encoded Cas10 analog (Csx11, TIGR02682) is divergent showing no sequence similarity detectable by conventional sequence alignment and search methods. However, when employing sensitive profile–profile alignments, using HHsearch, we obtained a match with a 98% probability score, between most of the N-terminal half of the protein and Csm1, the Cas10 analog of subtype III-A. In contrast to the earlier report, we found that the Cas5 analog of subtype III-C gives a full-length sequence match to Csm4, the Cas5 analog of subtype III-A and that it is not fused to Cas7. Finally, we were unable to detect any sequence similarity between protein S of subtype III-C and Csm2 or Cmr5 of subtypes III-A and III-B, respectively.

Earlier, the SSO1438-type Type III module was documented as being distinct from subtypes III-A and III-B. With 16 examples in diverse archaea and at least one in bacteria (*Rhodothermus*), we propose naming this subtype III-D. The Cas10 analog is quite similar to those of subtypes III-A and III-B but subtype III-D is exceptional in carrying a higher number of RAMP genes, where Cas7 has diverged into four or five paralogs, in addition to the presence of the Cas5 analog (Fig. 4). Protein S of subtype III-D

<table>
<thead>
<tr>
<th>Table 3. Number and kingdom distribution of the major subtypes of archaeal Type I and Type III systems</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td>subtype</td>
</tr>
<tr>
<td>A</td>
</tr>
<tr>
<td>B</td>
</tr>
<tr>
<td>C</td>
</tr>
<tr>
<td>D</td>
</tr>
<tr>
<td>E</td>
</tr>
<tr>
<td>G</td>
</tr>
</tbody>
</table>

cren, crenarchaea; eury, euryarchaea.

<table>
<thead>
<tr>
<th>Table 4. Number and kingdom distribution of variant archaeal interference subtypes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td>subtype</td>
</tr>
<tr>
<td>Vr-1</td>
</tr>
<tr>
<td>Vr-2</td>
</tr>
<tr>
<td>Vr-3</td>
</tr>
<tr>
<td>Vr-4</td>
</tr>
<tr>
<td>Type III</td>
</tr>
<tr>
<td>subtype</td>
</tr>
<tr>
<td>VIII-1</td>
</tr>
<tr>
<td>VIII-2</td>
</tr>
<tr>
<td>Unclassified</td>
</tr>
<tr>
<td>V-1</td>
</tr>
<tr>
<td>V-2</td>
</tr>
<tr>
<td>V-3</td>
</tr>
</tbody>
</table>

cren, crenarchaea; eury, euryarchaea.
ORF (~125 aa) is found in all VIII-1 modules. It is predicted to

Figure 3. Dendrogram and gene syntenies of archaeal interference
modules of different Type I subtypes where the total numbers of identi-

VIII-1 modules carry a Cas5 and a Cas7 analog together with a small protein (89 aa) and a large protein (659 aa). A similar

Another small ORF (~300 aa), which may be a Cas8/10 analog.

A further variant subtype V III-2 was detected in

module also occurs in Thermotoga lettingae, and a profile–profile

Non-core proteins specific for CRISPR systems

Unclassified CRISPR systems

Several interference modules could not be classified as either

Three examples of two different proteins were found, one a pre-

Unclassified CRISPR systems

Non-core proteins specific for CRISPR systems

Genome context analyses revealed that non-core cas genes are often linked to cas gene cassettes of CRISPR-Cas systems and all

Table 5

Table 4

Non-core proteins specific for CRISPR systems

Genome context analyses revealed that non-core cas genes are often linked to cas gene cassettes of CRISPR-Cas systems and all identified examples of non-core cas-associated genes are marked in Table S1 together with their cognate core cas gene modules. These accessory genes encode a variety of proteins, which include Csx1, Csx3, HerA, and NurA, as reported earlier.3,5,17 Several additional genes are listed in Table 5, some of which are specific to CRISPR-Cas systems, which include the genes encoding Csx1 and sRRM. Other linked proteins, including Cas_RecF, contain structural domains which are also encoded elsewhere but we were able to demonstrate a functional link to CRISPR-Cas systems because sequence analyses revealed that the proteins had coevolved significantly with their partner CRISPR-Cas system. Thus, Cas_RecF proteins are more closely related to each other, sequence-wise, than they are to other proteins with recF domains.

There was minimal evidence for the presence of non-core cas genes associated specifically with Type I systems (Table S1). Only a few examples of two different proteins were found, one a predicted ABC ATPase and the other containing an RRM domain (Table 5). Exceptionally, several proteins were found associated with Type I systems of the order Thermococcales and they are listed separately in Table S2. The genes encoding these proteins flank CRISPR systems in Thermococcales regardless of their subtype. This may reflect that the CRISPR systems are borne on integrated elements or specialized genomic loci where the flank-
ing genes, although conserved, have no direct functional link with CRISPR activity.

In contrast, analysis of Type III systems yielded several different

A further variant subtype V III-2 was detected in Ferroglobus placidus that carries a Cas5 and a Cas7 analog together with a small protein (89 aa) and a large protein (659 aa). A similar
their being co-functional (Table S1). Examples of archaeal CRISPR-Cas gene cassettes carrying interspaced accessory protein genes, sometimes present in multiple copies, are color-coded violet in Figure 5. In total, 18 different accessory proteins were identified, 11 of which were specific for Type III systems, and seven were encoded together with combined Type I + Type III systems (Table 5). Many of the proteins are designated Csx1 and they comprise a large group of disparate proteins exhibiting a wide range of sizes and very diverse sequences but share similar N-terminal domains. Significantly, a Csx1 protein of S. islandicus was recently implicated in influencing the nucleic acid targeting specificity of a subtype III-B interference complex.13

The present work did not include an analysis of toxin–antitoxin gene pairs and IS elements that are commonly associated with archaeal CRISPR systems.15 Nor does it cover proteins implicated in modulating CRISPR RNA transcription and processing including Cbp1 of the Sulfolobales19 and Cbp2 of the Thermoproteales,19 which, like RNase III that mediates RNA processing in bacterial Type II CRISPR systems,20 are not linked genomically to CRISPR systems, and are therefore likely to perform additional cellular functions.

CRISPR RNA processing

Cas6 is the primary processing endoribonuclease for all archaeal CRISPR transcripts, which are generally transcribed as single transcripts from within the leader.21,22 The cas6 gene is found either separately, or associated with gene cassettes for adaptation, Type I interference, Type III interference, or combinations thereof (Table S1). Experimental data suggest that the crRNA processing capability of a single Cas6 protein can be shared by different CRISPR systems in the same host.15 Consistent with this finding, we observed that cas6 genes associated with Type III systems are not phylogenetically distinct from those associated with Type I systems. This is visualized on the full archaeal Cas6 dendrogram where Cas6 proteins associated specifically with Type III systems are intermixed with those associated with Type I systems (Fig. S3, http://crispr.archaea.dk/FigureS3.pdf).

Modular exchange

The genomic organization of archaeal adaptation and interference genes as separate cas gene cassettes is consistent with their constituting separate functional modules.15 Moreover, exchange of adaptation and interference modules between different CRISPR systems appears to have been widespread,15,22 and this is exemplified for two methanoarchaea in Figure 6, where almost identical adaptation modules are associated with interference modules of three different Type I subtypes. To quantify the extent of modular exchange, we analyzed the adaptation and interference module dendrograms (Figs. S1 and S2). Groups of closely related adaptation modules were examined to establish whether their cognate Type I interference modules were also color-coded red. In summary, we estimate that about 50% of archaeal Type I systems are susceptible to exchange and that the phenomenon is particularly widespread for subtypes I-B, I-G, I-D, VI-1, and VI-2.

Discussion

In this study we have examined all the currently detectable CRISPR adaptive immune systems in sequenced archaeal genomes and have generated a catalog of different Type I and Type III subtypes found within each archaean (Table S1). The analyses have focused primarily on the properties of interference modules because of the widespread occurrence of modular exchange and the finding that adaptation modules are relatively highly conserved in protein composition and sequence. While most archaean Type I CRISPR systems constitute integral genetic units, many interference modules of Type III systems are encoded separately. The results obtained underline the high structural conservation of archaeal adaptation modules and the relatively broad structural diversity of interference modules.

Most archaean Type I systems fall into a few of the universal subtypes defined earlier,5 albeit using different criteria. In addition, we have assigned some Type III systems to other subtypes III-C, III-D, and VIII-1, the former two of which (families A and
D) were described earlier for archaea. In their more general classification, Makarova et al.14 employed the signature protein Cas10/Cmr2/Csm1 for classifying Type III systems, but this protein is extremely divergent in subtypes III-C, V III-1, and VIII-2, and we suspect, therefore, that some Type III systems remain undetected.

We successfully used the Pfam RAMP model (PF03787) as a signature for Type III systems because this model is quite specific for Type III Cas7 proteins. Therefore, we predict that employing PF03787 as a universal signature for Type III systems will also lead to the discovery of new Type III subtypes in bacteria.

### Table 5. Accessory proteins encoded by non-core cas genes associated primarily with Type III systems

<table>
<thead>
<tr>
<th>Name</th>
<th>Number</th>
<th>Example</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Type III-specific</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Csx1 proper</td>
<td>65</td>
<td>TTX_1228</td>
<td>protein (~450 aa) matching TF Cas_DxTHG and CDD Csx1_III_U models</td>
</tr>
<tr>
<td>Csx1 superfamily (Csx1s)</td>
<td>18</td>
<td>TON_0318</td>
<td>diverse group of proteins matching Csx1_III_U, including TF Cas02710. Does not include CasR and Cmr6</td>
</tr>
<tr>
<td>RecF-associated protein (RFas)</td>
<td>16</td>
<td>P186_0873</td>
<td>diverse group of small proteins (~170 aa), always associated with Cas RecF</td>
</tr>
<tr>
<td>Cas_RecF</td>
<td>12</td>
<td>P186_0874</td>
<td>protein with RecF domain found near crenarchaeal Type III systems. Always associated with RFas</td>
</tr>
<tr>
<td>Cas ABC ATPase (ABC_ATPase)</td>
<td>11</td>
<td>YN1551_2137</td>
<td>protein with ABC ATPase domain found near crenarchaeal Type III systems. Always associated with sRRMs</td>
</tr>
<tr>
<td>small RRM protein (sRRM)</td>
<td>12</td>
<td>YN1551_2138</td>
<td>diverse group of proteins (~150 aa) containing RRM domains and always associated with Cas ABC ATPase</td>
</tr>
<tr>
<td>Aspartic protease (Asp_prot)</td>
<td>11</td>
<td>Saci_1895</td>
<td>small protein (~100 aa) with retropepsin domain, associated with Type III-D and V_1 subtypes in Sulfolobales</td>
</tr>
<tr>
<td>Cmr7</td>
<td>7</td>
<td>SSO1986</td>
<td>protein (~200 aa) with metalloprotease domain, associated with certain variants of Type III-B subtypes</td>
</tr>
<tr>
<td>Csx3</td>
<td>5</td>
<td>Mhar_1706</td>
<td>small (~100 aa) protein matching to the TF Cas_Csx3 model. Unrelated to Csx1</td>
</tr>
<tr>
<td>Membrane dipeptidase (Zn²⁺_pep)</td>
<td>4</td>
<td>MJ_1673</td>
<td>small (~130 aa) protein with Zn²⁺ dipeptidase domain, associated with some Type III-A systems in Methanococcales</td>
</tr>
<tr>
<td>Mvol_05XX-fam</td>
<td>5</td>
<td>Mvol_0536</td>
<td>five proteins spanning three families associated with Type III-B systems in Methanococcales. Also found in Clostridium sp.</td>
</tr>
<tr>
<td><strong>Types I + III-specific</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cas HerA helicase (HerA)</td>
<td>14</td>
<td>Tagg_0815</td>
<td>always associated with NurA</td>
</tr>
<tr>
<td>Cas NurA nuclease (NurA)</td>
<td>14</td>
<td>Tagg_0814</td>
<td>always associated with HerA</td>
</tr>
<tr>
<td>Csx1 minimal (Csx1m)</td>
<td>13</td>
<td>Pyrfu_0517</td>
<td>small (~180 aa) protein containing a minimal Csx1 domain matching TF Cas_NEO113 and CDD Csx1_III_U</td>
</tr>
<tr>
<td>Csx1 with PIN toxin (Csx1p)</td>
<td>8</td>
<td>Metig_1253</td>
<td>~400 aa protein consisting of a minimal Csx1 domain fused to a PIN toxin domain</td>
</tr>
<tr>
<td>Tneu_1160-fam</td>
<td>3</td>
<td>Tneu_1160</td>
<td>large protein of unknown function, associated with CRISPR systems of Thermoproteales</td>
</tr>
<tr>
<td>SbcC repair ATPase (SbcC)</td>
<td>3</td>
<td>CSUB_C0986</td>
<td>large protein with SbcC domain found in cas gene cassettes of diverse archaea</td>
</tr>
<tr>
<td>SbcD nuclease (SbcD)</td>
<td>3</td>
<td>CSUB_C0987</td>
<td>protein with SbcD domain, associated with SbcC</td>
</tr>
<tr>
<td><strong>Type I-specific</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cas ABC ATPase (ABC_ATPase)</td>
<td>5</td>
<td>TTX_1256</td>
<td>specific for Type I systems but related to corresponding protein for Type III systems</td>
</tr>
<tr>
<td>small RRM protein (sRRM)</td>
<td>4</td>
<td>TTX_1257</td>
<td>specific for Type I systems but related to corresponding protein for Type III systems</td>
</tr>
<tr>
<td>Thermococcales specific (see Table S2)</td>
<td>53</td>
<td>PF1132–1135</td>
<td>11 protein families found in Thermococcales adjacent to Type I systems regardless of subtype</td>
</tr>
</tbody>
</table>

 Most of the accessory proteins are specific to CRISPR systems and are not normally found encoded outside of cas gene cassettes. In addition to the most widespread Csx1-type accessory proteins, numerous accessory proteins have helicase and nuclease domains, which tend to be found associated with DNA replication, recombination, and repair. Protease domains, toxin domains, and RRM domains are also found in many proteins, while some larger proteins have non-identifiable domains.
In late 2011, Makarova et al. proposed unifying protein families of the different CRISPR types, Type I Cas and Type III Cmr and Csm, basically inferring that all Type I interference systems were made up of Cas3, Cas5, Cas7, and Cas8, and that Type III interference complexes contained Cas5, Cas7, Cas10, and the small protein component S. This proposal has received strong independent support from a number of recent structural studies. For example, the subtype I-E and subtype III-B interference complexes exhibit some common structural features that are illustrated schematically in Figure 7. Type I interference modules generally encode a Cas3 helicase, Cas5 RAMP, Cas7 RAMP, and Cas8, and in the subtype I-E complex, Cas5 is embedded in Cas8 to form the base of the complex from which a backbone of multiple Cas7 subunits extends (Fig. 7A). For some subtypes, Cas8 is split into Cas8' and Cas8", and Cas3 can also be partitioned. The interference complex of subtype III-B does not carry a Cas3 helicase but it exhibits analogous structural features that are illustrated schematically in Figure 7. Type I interference modules generally encode a Cas3 helicase, Cas5 RAMP, Cas7 RAMP, and Cas8, and in the subtype I-E complex, Cas5 is embedded in Cas8 to form the base of the complex from which a backbone of multiple Cas7 subunits extends (Fig. 7A). For some subtypes, Cas8 is split into Cas8' and Cas8", and Cas3 can also be partitioned. The interference complex of subtype III-B does not carry a Cas3 helicase but it exhibits analogous structural features, where Cas8' is replaced by Cas10 and Cas8" is exchanged with protein S (Fig. 7B). Moreover, Cas9 is also embedded in Cas10 and the shared Cas7 backbone interfaces with other parts of the complex through divergent Cas7 paralogs, the numbers of which differ between different Type III subtypes (Fig. 7B). Although protein sequence similarities were very difficult to detect by conventional sequence comparison methods, for many predicted homologs, using the profile–profile approach employed here we were able to detect and confirm the predicted homologies. The resulting reduction in the number of core Cas protein families considerably facilitated Cas protein identification and annotation of cas genes. It is likely that some additional Cas proteins described in a recent metagenomic study of bacterial CRISPR systems could have been assigned within this unified Cas protein classification if more sensitive sequence comparative methods had been employed.

Extensive evidence supporting the periodic exchange of adaptation and interference modules between different archaeal Type I systems is summarized in Figures S1 and S2 and the data are consistent with earlier evidence presented for archaeal CRISPR modular exchange. Presumably this widespread phenomenon reflects that the structure and function of the adaptation modules are highly conserved. This property also undermines using the phylogeny of the conserved adaptation protein Cas1 for determining CRISPR subtypes. However, the sole employment of interference modules for classifying archaeal CRISPR systems also carries the reservation that they may function with different types of adaptation modules.

We still have limited insights into the detailed mechanistic and functional diversity of the interference modules of the different Type I and Type III subtypes. Closely related S. islandicus strains were shown earlier to carry multiple and different Type III subtypes, which increases the likelihood that their interference mechanisms are functionally diverse. Furthermore, the recent broad unification of core Cas protein families has provided a basis for discerning non-core proteins that are co-functional with...
CRISPR systems. Several accessory protein families were identified in this study that are exclusively or primarily encoded within, or adjacent to, a limited number of archaeal Type III interference gene cassettes. Two of these, HerA and NurA, were identified earlier for CRISPR systems of different Pyrobaculum species.\textsuperscript{17} It is likely that some of the proteins, which include putative proteases and ATPases, can modify or extend functions of the associated CRISPR systems. This supposition recently received experimental support from one of two Type III-B systems in \textit{S. islandicus} REY15A that was demonstrated to share a CRISPR spacer, and a Cas6 RNA processing enzyme, with a Type I-A system, but was functionally dependent on an accessory Csx1 protein.\textsuperscript{13}

Proteins annotated as Csx1 do not comprise a single protein family, but rather a number of diverse protein families exhibiting different sizes and variable domain architectures. What they share is an N-terminal domain of approximately 150 amino acids in length containing a DxTHG motif or variants thereof. This domain, which defines what can be called the Csx1 superfamily of proteins, is also found in the \textit{cas} transcriptional regulators CasR and Csm6 where it is fused to a C-terminal HTH domain. While it is not clear why such a conserved domain should be fused to a wide range of different proteins, the conserved domain could provide a site for interfacing with Type III and Type I interference modules to modify their activities.

CRISPR-based immune systems probably originated in early primitive cellular structures where exchange of genetic material was likely to have been common.\textsuperscript{27,28} Moreover, their widespread presence in most archaea and many bacteria suggests that they predated the inferred branching of the bacterial domain, although the bacterial Type II system is likely to have evolved into its present form later. We have discussed earlier the evidence supporting inter-genomic exchange of CRISPR systems between archaea.\textsuperscript{23} This was considered to be facilitated by the often integral nature of the CRISPR-\textit{cas} gene cassettes, which are often located in variable genomic regions and are sometimes bordered by transposable elements.\textsuperscript{16} Nevertheless, the relative similarity of protein components of a few archaeal and bacterial subtypes, combined with their biased distributions, suggests that some inter-domain exchange of CRISPR systems has occurred, despite the presence of formidable genetic barriers.\textsuperscript{23} For example, the relatively common bacterial subtypes I-C and I-E occur rarely among archaea, and no examples of bacterial subtype I-F, were found (Fig. 1), suggesting that rare transfers from bacteria to archaea may have occurred, and primarily among the euryarchaea.

Crenarchaea, which tend to occupy archaea-rich thermophilic or acidothermophilic environments, are relatively homogeneous in their CRISPR systems. They exhibit predominantly subtypes I-A (90\%) and III-B (53\%), which suggests that they have undergone very few, if any, transfers of CRISPR systems from other archaeal kingdoms or from bacteria. In contrast, the euryarchaea carry several examples of subtypes I-A, I-B, I-D, and I-G, and subtype III-A (52\%) which constitutes the most common Type III system. Euryarchaea frequent a wide range of natural environments and are more diverse phylogenetically than crenarchaea. Moreover, many of their natural habitats are relatively rich in bacteria, possibly rendering their CRISPR systems more susceptible to exchange between archaea and bacteria.
Our aim in this work was to complete a comprehensive analysis of the archaeal CRISPR systems. Here we provide an interactive system, in Table S1, and in Figure S2, which can be used for future analysis and characterization of these systems, including further examination of the different variant systems as well as the possible functional roles of the non-core Cas proteins.

Materials and Methods

Bioinformatical analyses

Genomic loci encoding archaeal adaptation, Type I interference, and Type III interference modules were first identified in the set of 159 complete archaeal genomes by searching for cas1 and cas7 genes using custom HMMs29 and the Pfam30 RAMPs model. The identified cas1 and cas7 genes and the 10 genes flanking them on either side were pooled, and the whole pool was subject to an all-against-all pairwise sequence alignment comparison,31 which was used as an input for Markov clustering32 using custom similarity measures. Each protein family resulting from the Markov clustering was searched against protein family databases CDD,3 COG,4 TIGFAMs,3 and Pfam30 using profile–profile alignments with HHsearch.33 Each genomic locus containing the genes was inspected manually, and using the information from the profile–profile comparison, as well as conventional sequence searches against public databases, cas modules, and cas gene families, were defined. cas cassettes were also defined as consisting of collections of modules (Table S1). Adaptation and interference module dendrograms (Figs. S1 and S2) were created by comparing all protein components of each type of module against corresponding modules from other casettes. By keeping track of the modules corresponding to each protein component, module-to-module similarity scores were calculated by adding up scores for all constituent proteins. Scores between all modules were inverted and normalized to create a module-to-module distance matrix. The distance matrix was used as input for constructing a neighbor-joining tree,34 which was subsequently mid-point rooted. Variant systems that could not be classified as either Type I or Type III were not included in the interference dendrogram.

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Acknowledgments

Qunxin She is thanked for stimulating discussions and suggestions, and meetings with the German FOR 1680 CRISPR Consortium were both productive and helpful. Michael Terns and Kira Makarova provided insightful and encouraging advice and Johannes Söding’s help with the HHsearch profile–profile program was much appreciated. The research was supported by the Danish Natural Science Research Council.

Supplemental Materials

Supplemental materials may be found here: http://www.landesbioscience.com/journals/rnabiology/article/27990/

References

9. Marraffini LA, Sontheimer EJ. CRISPR interference limits horizontal gene transfer in staphylococci by targeting DNA. Science 2008; 322:1843-5; PMID:19095942; http://dx.doi.org/10.1126/science.1165771


