CRISPR-Cas: an adaptive immunity system in prokaryotes

Eugene V Koonin* and Kira S Makarova

Address: National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, 8600 Rockville Pike, Bethesda, MD 20894, USA

* Corresponding author: Eugene V Koonin (koonin@ncbi.nlm.nih.gov)

Abstract

Most of the archaea and numerous bacteria possess an elaborate system of adaptive immunity to mobile genetic elements known as the CRISPR (clustered regularly interspaced short palindromic repeats)-associated system (CRISPR-Cas), which consists of arrays of short repeats interspersed with unique DNA spacers and adjacent operons encompassing CRISPR-associated (*cas*) genes with predicted and, in some cases, experimentally validated nuclease, helicase, and polymerase activities. The system functions by integrating fragments of alien DNA between the repeats and employing their transcripts to degrade the DNA of the respective invading elements via an RNA interference-like mechanism. The CRISPR-Cas system is a case of apparent Lamarckian inheritance.

Introduction and context

A brief history of the serendipitous discovery of a prokaryotic immune system

Bacteria and archaea exist in an incessant arms race with various selfish genetic elements (phages, transposons, and plasmids) and have evolved a variety of defense systems. The best known ones probably are the numerous restriction-modification enzyme systems that exploit different methylation patterns of host and infecting agent DNA to eliminate the invader [1]. Recently, a novel widespread defense system that functions on a completely different principle was discovered; it became known as the CRISPR (clustered regularly interspaced short palindromic repeats)-associated system, usually referred to as CRISPR-Cas (where Cas stands for CRISPR-associated proteins) or, alternatively, as CASS [2-4]. The discovery of this system involved considerable intrigue and serendipity. Distinct arrays of short repeats interspersed with unique spacers (CRISPR) have been seen in bacterial and archaeal genomes for years, with no clues as to their possible functions [5,6]. Independently, Cas protein sequences encoded by putative operons adjacent to CRISPR [7] were analyzed in detail and found to contain domains characteristic of several nucleases, a helicase, a

polymerase, and RNA-binding proteins; it has been suggested that these proteins might belong to a novel repair system [8]. A new light was shed on the probable function of the CRISPR when it was observed that some of the unique inserts were (nearly) identical to fragments of phage and plasmids genes, a pivotal observation that immediately led to the idea that CRISPR might be involved in defense against selfish elements [9-11]. These findings were combined with the results of comprehensive computational re-analysis of the Cas proteins to develop a detailed hypothesis on the mechanism of CRISPR-Cas [12]. This hypothesis drew a close analogy between the putative novel prokaryotic defense system and the eukaryotic RNA interference (RNAi) mechanisms [13], with the important difference that CASS mediates integration of a piece of alien DNA into the host genome as the first step in the sequence of events which leads to immunity to the given agent [12]. Specific roles for individual Cas proteins were proposed as well on the basis of their domain composition and by analogy with RNAi components although the proteins involved are not homologous [12]. This hypothesis prompted direct experiments that demonstrated that engineering a specific bacteriophage sequence into the CRISPR locus of the lactic bacterium *Streptococcuss thermophilus* indeed

conferred resistance to the cognate phage, an effect that was abrogated by even a single mismatch between the insert and the target gene [14]. These key experiments clinched the case for the defense role of CRISPR-Cas and triggered an avalanche of further genetic and biochemical experiments.

Major recent advances
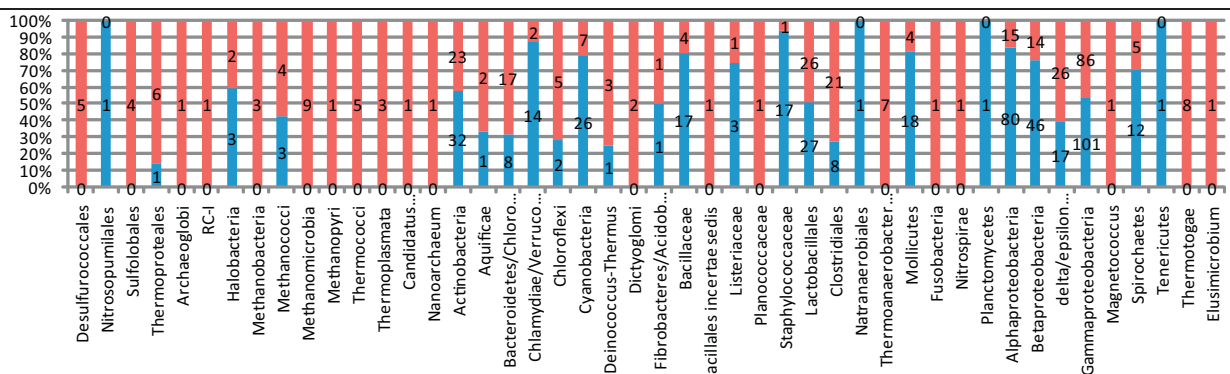### Comparative genomics, diversity, and evolution of CRISPR-Cas systems
CRISPR-Cas is a highly diverse constellation of genes, with the number of CRISPR-*cas* loci, *cas* gene repertoire, and (predicted) operon organization often changing even between closely related strains [3,12,15-18]. Comparative analysis of operon architectures revealed seven distinct types of CRISPR-Cas, each of which is characterized by a distinct signature of genomic architecture [12]. Only two genes, *cas1* and *cas2*, are invariably present in each CRISPR-Cas system so far detected and accordingly can be used as genomic markers of CRISPR-Cas. In addition to the two universal genes, three genes (*cas3*, *cas4*, and *cas5*) are present in the majority of CRISPR-Cas, and approximately 20 other genes are found in various subsets of these systems [12,15]. Cas5 and several other less common CRISPR-Cas components belong to the large and extremely diverged superfamily of repeat-associated mysterious proteins (RAMPs) [12,15].

Using the highly conserved Cas1 protein sequence as a marker, we detected CRISPR-Cas in 297 of the 774 analyzed prokaryotic genomes (37%); among archaea, CRISPR-Cas is much more common than among bacteria: up to 90% of the available archaeal genomes carry CRISPR-Cas (Figure 1). The representation of

CRISPR-Cas in the genomes of diverse groups of archaea and bacteria differs within a broad range, from ubiquity to complete absence (although it should be noted that all groups completely lacking CRISPR-Cas are currently represented by a small number of genomes) (Figure 1).

Phylogenetic analysis of core *cas* genes, such as *cas1* and *cas3*, fails to recover major bacterial and archaeal lineages, an observation that appears to be indicative of extensive mobility of CRISPR-*cas* modules via horizontal gene transfer (HGT). The *cas* genes are not only horizontally mobile but also typically show high rates of sequence evolution, with the partial exception of core genes, in particular *cas1* [12,15]. In many cases, this fast evolution renders sequence conservation between homologous Cas proteins barely detectable, most conspicuously among the RAMPs, which are propagated by both gene duplication and HGT and constitute a large fraction of Cas protein sets in most CRISPR-Cas-carrying prokaryotes [12]. The RAMPs are extremely diverged in sequence, so that the demonstration that different RAMP families were related and possessed the same fold required the careful use of the most sensitive sequence analysis methods (and even so, it is likely that additional RAMPs have been missed). Conceivably, RAMPs and perhaps some other Cas proteins evolve under positive selection dictated by the arms race with selfish elements [19,20]. This possibility is congruent with the observations that, although virus or plasmid origin is apparent for a considerable number of CRISPR spacers, the majority of the spacers are not significantly similar to any sequences in current databases [9,19,20]. In addition, deletion of CRISPR units (a repeat with a spacer) has been demonstrated [21]. Thus, it appears, first, that

Figure 1. Representation of the CRISPR (clustered regularly interspaced short palindromic repeats)-Cas system in archaeal and bacterial lineages

The bars show the fractions of sequenced genomes in each group that possess CRISPR-Cas (red) and lack CRISPR-Cas (blue). The actual numbers of available genomes are also indicated for both parts of each bar. The figure is based on the results of screening of the proteins encoded in 774 archaeal and bacterial genomes for statistically significant sequence similarity to Cas1 sequences using BLASTP [32].

the repertoire of selfish genetic elements encountered by archaea and bacteria is vast, and second, that the CRISPR-Cas-mediated immunity is short-lived; that is, spacers rapidly deteriorate by mutation once the cognate element is no longer a threat [19].

Taken together, these observations identify the CRISPR-Cas as a bona fide component of the prokaryotic mobilome [that is, the totality of genetic elements that are characterized by extensive horizontal mobility and include selfish elements (viruses, plasmids, transposons, and so on) as well as defense and stress response systems] [22,23]. Notably, CRISPR-*cas* loci are often located within 'defense islands' (i.e., regions of bacterial and archaeal genomes that consist primarily of genes encoding defense and stress response systems, such as restriction-modification and toxin-antitoxin modules) [24]. This genomic association permits the prediction of novel prokaryotic defense systems. Comparative-genomic analysis of the CRISPR-*cas* loci is facilitated by the use of specialized databases and accompanying custom software tools for CRISPR detection [25].

Molecular mechanisms of CRISPR-Cas and functions of Cas proteins

CRISPR-Cas systems mediate immunity to invading genetic elements via three distinct stages: (a) adaptation, (b) expression and processing of CRISPR, and (c) interference [3]. The full molecular picture is far from being clear for each of these stages, but recently several fundamental results, particularly on the processing of CRISPR transcripts, have been reported. With regard to the adaptation stage, following the original work that demonstrated the insertion of a phage-specific spacer into the CRISPR locus of *Streptococcus thermophilus*, this process was explored systematically, leading to the conclusion that a phage challenge typically triggers insertion of a single phage-specific resistance-conferring spacer with a characteristic length of 30 base pairs; successive infection of a bacterial culture with multiple phages led to the accumulation of the cognate spacers in the CRISPR loci [21]. Furthermore, it has been shown that insertion of new spacers depends on short PAMs (proto-spacer adjacent motifs), which differ between variants of the CRISPR-Cas system and might determine the identity of the inserted spacer [26].

The original 'prokaryotic RNAi' hypothesis maintained that CRISPR-Cas systems would target mRNAs of invading agents [12]. However, the first experiments that, in general, validated the hypothesis have also shown that both strands of the CRISPR spacer DNA were effective in conferring immunity to the cognate phage, an observation best compatible with a DNA target [14]. A more direct experiment showed that the insertion of a self-splicing intron into the target gene made the respective plasmid resistant to the CRISPR-mediated immunity, a clear indication that the invading DNA itself is targeted [27]. Whether this conclusion is general and applies to all CRISPR-Cas remains to be determined, especially given the extreme diversity of the architectures of these systems.

As of September 2009, biochemical activities and/or crystal structures of several widespread Cas proteins have been determined (Table 1) [3]. In agreement with the computational predictions and nuclease activities, either RNAse or DNAse or both were demonstrated for several Cas proteins. Notably, these novel nucleases include both universal Cas proteins. Specifically, Cas1 has been shown to be a metal-dependent DNAse with no sequence specificity and has been implicated in the integration of the alien DNA into the CRISPR cassettes [28]. Cas2 has been characterized as a metal-dependent endoribonuclease whose role in the CRISPR-Cas mechanism remains unclear [29]. A striking finding is that some of the RAMP proteins that contain a double ferredoxin-fold domain and were originally proposed to be non-enzymatic RNA-binding proteins (considering their extreme sequence divergence [12]), actually possess RNAse activity that is apparently involved in the processing of CRISPR transcripts [30,31]. In particular, a RAMP protein seems to be the active moiety of the CASCADE (CRISPR-associated complex for antiviral defense) complex that consists of five Cas proteins (Table 1) and is the CRISPR-processing machine of *Escherichia coli* [30]. In concert with the Cas3 protein that consists of (predicted) helicase and nuclease domains, CASCADE seems to be involved in the interference stage.

Future directions

Comparative-genomic predictions validated by a rapidly growing body of experimental results indicate that the CRISPR-Cas is an adaptive immunity system that is widely employed by archaea and bacteria for defense against diverse invading elements, in particular, viruses. The system functions by integrating fragments of alien element genes into CRISPR loci and employing the resulting spacers, after transcription and processing, as guide RNAs to abrogate the replication of the cognate elements by cleaving nucleic acid molecules complementary to the guide. In some cases, at least, the target of CRISPR-Cas is the genomic DNA of an invading genetic element. Experiments aimed at molecular dissection of CASS proved the predicted principle of its action and are starting to reveal multiple activities of the protein components of CASS and the molecular architecture of complexes formed by these proteins. However, an

Table 1. Cas proteins and complexes: predicted and experimentally determined biochemical activities and possible functions in CRISPR-mediated immunity

Protein[a]	Representation in CRISPR-Cas[b]	Domain organization[b]	Predicted activity and function[b]	Three-dimensional structure[c]	Experimentally demonstrated activity and/or function
Cas1 (COG1518)	Universal	Highly conserved domain without detectable similarity to other proteins	Nuclease, possibly integrase; role in adaptation (integration of invader DNA)	3GOD: unique mostly α-helical fold [28]	Metal-dependent nuclease, cleaves both DNA and RNA [28]
Cas2 (COG1343, COG3512, and additional small families)	Universal (except for some probably non-functional CASS variants)	Small domain distantly related to VapD, an uncharacterized bacterial protein linked to toxin-antitoxin system; some fusions with Cas3	Nuclease, possible role in adaptation	2IVY, 2I8E: ferredoxin-like fold [29]	Sequence-specific endoribonuclease [29]
Cas3 (COG1203)	Present in a substantial majority of CASS, with the exception of several reduced though possibly functional systems	Superfamily 2 helicase, typically fused to HD nuclease; in some CASS variants, helicase and nuclease are encoded by adjacent genes	Helicase-nuclease, possible roles at all stages of CASS-mediated immunity	None	Interacts with CASCADE, contributes to CASS-mediated interference [30]; endonuclease activity demonstrated from a stand-alone HD-protein from *Sulfolobus* [33]
Cas4 (COG1468, COG4343)	Present in a substantial majority of CASS, with the exception of several reduced though possibly functional systems	RecB-like nuclease domain and an additional metal-binding module	Nuclease, implicated in adaptation	None	None
RAMPs (Cas5 [COG1688], Cas6 [COG1583], COGs 1769, 1567, 1336, 1367, 1604, 1337, 1332, 5551, and additional small families)	Diverse subsets of RAMPs present in all CASS	RAMP	RNA-binding proteins, probably sequence-structure-specific	1WJ9, 3I4H Duplicated ferredoxin-fold domain	Cas5 in *Escherichia coli* is a CASCADE subunit and directly cleaves CRISPR transcripts generating guide RNAs [30]; Cas6 performs the same function in *Pyrococcus* [31]
CASCADE [CasABCDE(Cse1234-Cas5e) complex]	*E. coli*; different combinations of sub-units in other prokaryotes	Cas5e (COG1688) is a RAMP; other domains uncharacterized	Cse4 (COG1857): predicted nuclease; Cas5e: predicted RNA-binding protein	See above for Cas5.	CasC is the principal structural subunit; Cas5e is the nuclease subunit [30].

[a]These are only the most widespread and experimentally characterized Cas proteins; there is no unified nomenclature of Cas proteins [3]; the Cas protein names are accompanied by the numbers of clusters of orthologous genes (COGs) [34] where available. [b]Data are from references [3] and [12]. [c]Structures are identified by Protein Data Bank accession numbers. Cas, CRISPR (clustered regularly interspaced short palindromic repeats)-associated protein; CASCADE, CRISPR (clustered regularly interspaced short palindromic repeats)-associated complex for antiviral defense; CASS, CRISPR (clustered regularly interspaced short palindromic repeats)-associated system; CRISPR, clustered regularly interspaced short palindromic repeats; HD, HD (histidine-aspartate)-family nuclease; RAMP, repeat-associated mysterious protein.

enormous amount of experimental work remains to be done to elucidate the mechanisms of CASS, in particular, the molecular details of spacer incorporation into the CRISPR loci and the specific pathways of RNA-guided destruction of alien genomes. These experiments can be expected to reveal the considerable mechanistic diversity that reflects the extreme diversity of *cas* gene repertoires and operonic organization. Another important direction of future work is the characterization of the arms race between CRISPR-Cas and viruses of prokaryotes and elucidation of putative mechanisms of counterdefense employed by the viruses. Finally, it is worth noting that, by integrating fragments of invaders' genomes into the genomes of the archaeal and bacterial hosts, the CASS effectively operates via a Lamarckian-type inheritance of acquired characters.

Abbreviations

Cas, CRISPR-associated protein; CASCADE, CRISPR-associated complex for antiviral defense; CASS, CRISPR-associated system; CRISPR, clustered regularly interspaced short palindromic repeats; HGT, horizontal gene transfer; PAM, proto-spacer adjacent motif; RAMP, repeat-associated mysterious protein; RNAi, RNA interference.

Competing interests

The authors declare that they have no competing interests.

Acknowledgments

References

1. Kusano K, Naito T, Handa N, Kobayashi I: **Restriction-modification systems as genomic parasites in competition for specific sequences.** *Proc Natl Acad Sci U S A* 1995, **92**:11095-9.

2. Sorek R, Kunin V, Hugenholtz P: **CRISPR–a widespread system that provides acquired resistance against phages in bacteria and archaea.** *Nat Rev Microbiol* 2008, **6**:181-6.

3. van der Oost J, Jore MM, Westra ER, Lundgren M, Brouns SJ: **CRISPR-based adaptive and heritable immunity in prokaryotes.** *Trends Biochem Sci* 2009, **34**:401-7.

4. Waters LS, Storz G: **Regulatory RNAs in bacteria.** *Cell* 2009, **136**:615-28.

5. Groenen PM, Bunschoten AE, van Soolingen D, van Embden JD: **Nature of DNA polymorphism in the direct repeat cluster of Mycobacterium tuberculosis; application for strain differentiation by a novel typing method.** *Mol Microbiol* 1993, **10**:1057-65.

6. Jansen R, van Embden JD, Gaastra W, Schouls LM: **Identification of a novel family of sequence repeats among prokaryotes.** *OMICS* 2002, **6**:23-33.

7. Jansen R, Embden JD, Gaastra W, Schouls LM: **Identification of genes that are associated with DNA repeats in prokaryotes.** *Mol Microbiol* 2002, **43**:1565-75.

8. Makarova KS, Aravind L, Grishin NV, Rogozin IB, Koonin EV: **A DNA repair system specific for thermophilic Archaea and bacteria predicted by genomic context analysis.** *Nucleic Acids Res* 2002, **30**:482-96.

9. Mojica FJ, Diez-Villasenor C, Garcia-Martinez J, Soria E: **Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements.** *J Mol Evol* 2005, **60**:174-82.

10. Pourcel C, Salvignol G, Vergnaud G: **CRISPR elements in Yersinia pestis acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies.** *Microbiology* 2005, **151**:653-63.

11. Bolotin A, Quinquis B, Sorokin A, Ehrlich SD: **Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin.** *Microbiology* 2005, **151**:2551-61.

    F1000 Factor 3.2 *Recommended*
    Evaluated by Peter Williams 22 Sep 2005, Camilla Nesbo 27 Oct 2005

12. Makarova KS, Grishin NV, Shabalina SA, Wolf YI, Koonin EV: **A putative RNA-interference-based immune system in prokaryotes: computational analysis of the predicted enzymatic machinery, functional analogies with eukaryotic RNAi, and hypothetical mechanisms of action.** *Biol Direct* 2006, **1**:7.

    F1000 Factor 3.2 *Recommended*
    Evaluated by Richard Willson 04 Jul 2006, Julian Parkhill 01 Aug 2006

13. Carthew RW, Sontheimer EJ: **Origins and Mechanisms of miRNAs and siRNAs.** *Cell* 2009, **136**:642-55.

14. Barrangou R, Fremaux C, Deveau H, Richards M, Boyaval P, Moineau S, Romero DA, Horvath P: **CRISPR provides acquired resistance against viruses in prokaryotes.** *Science* 2007, **315**:1709-12.

    F1000 Factor 6.6 *Must Read*
    Evaluated by Joachim Kurtz 27 Mar 2007, Julian Parkhill 04 Apr 2007, Judy Wall 16 Aug 2007

15. Haft DH, Selengut J, Mongodin EF, Nelson KE: **A guild of 45 CRISPR-associated (Cas) protein families and multiple CRISPR/Cas subtypes exist in prokaryotic genomes.** *PLoS Comput Biol* 2005, **1**:e60.

16. Horvath P, Coute-Monvoisin AC, Romero DA, Boyaval P, Fremaux C, Barrangou R: **Comparative analysis of CRISPR loci in lactic acid bacteria genomes.** *Int J Food Microbiol* 2009, **131**:62-70.

17. Horvath P, Romero DA, Coute-Monvoisin AC, Richards M, Deveau H, Moineau S, Boyaval P, Fremaux C, Barrangou R: **Diversity, activity, and evolution of CRISPR loci in Streptococcus thermophilus.** *J Bacteriol* 2008, **190**:1401-12.

18. DeBoy RT, Mongodin EF, Emerson JB, Nelson KE: **Chromosome evolution in the Thermotogales: large-scale inversions and strain diversification of CRISPR sequences.** *J Bacteriol* 2006, **188**:2364-74.

    F1000 Factor 3.0 *Recommended*
    Evaluated by Kevin Sowers 03 Apr 2006

19. Andersson AF, Banfield JF: **Virus population dynamics and acquired virus resistance in natural microbial communities.** *Science* 2008, **320**:1047-50.

20. Tyson GW, Banfield JF: **Rapidly evolving CRISPRs implicated in acquired resistance of microorganisms to viruses.** *Environ Microbiol* 2008, **10**:200-7.

    F1000 Factor 3.0 *Recommended*
    Evaluated by Yuri Wolf 07 Aug 2008

21. Deveau H, Barrangou R, Garneau JE, Labonte J, Fremaux C, Boyaval P, Romero DA, Horvath P, Moineau S: **Phage response to CRISPR-encoded resistance in Streptococcus thermophilus.** *J Bacteriol* 2008, **190**:1390-400.

    F1000 Factor 3.0 *Recommended*
    Evaluated by Aidan Coffey 20 Dec 2007

22. Frost LS, Leplae R, Summers AO, Toussaint A: **Mobile genetic elements: the agents of open source evolution.** *Nat Rev Microbiol* 2005, **3**:722-32.

23. Koonin EV, Wolf YI: **Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world.** *Nucleic Acids Res* 2008, **36**:6688-719.

24. Makarova KS, Wolf YI, Van der Oost J, Koonin EV: **Prokaryotic homologs of Argonaute proteins are predicted to function as key components of a novel system of defense against mobile genetic elements.** *Biol Direct* 2009, **4**:29.

25. Grissa I, Vergnaud G, Pourcel C: **CRISPRcompar: a website to compare clustered regularly interspaced short palindromic repeats.** *Nucleic Acids Res* 2008, **36**:W145-8.

26. Mojica FJ, Diez-Villasenor C, Garcia-Martinez J, Almendros C: **Short motif sequences determine the targets of the prokaryotic CRISPR defence system.** *Microbiology* 2009, **155**:733-40.

27. Marraffini LA, Sontheimer EJ: **CRISPR interference limits horizontal gene transfer in staphylococci by targeting DNA.** *Science* 2008, **322**:1843-5.

    F1000 Factor 10.0 *Exceptional*
    Evaluated by Elsebet Lund 23 Jan 2009, Judy Wall 23 Jan 2009, Rino Rappuoli 27 Jan 2009, Vincenzo Scarlato 18 Feb 2009

28. Wiedenheft B, Zhou K, Jinek M, Coyle SM, Ma W, Doudna JA: **Structural basis for DNase activity of a conserved protein implicated in CRISPR-mediated genome defense.** *Structure* 2009, **17**:904-12.

29. Beloglazova N, Brown G, Zimmerman MD, Proudfoot M, Makarova KS, Kudritska M, Kochinyan S, Wang S, Chruszcz M, Minor W, Koonin EV, Edwards AM, Savchenko A, Yakunin AF: **A novel family of sequence-specific endoribonucleases associated with the clustered regularly interspaced short palindromic repeats.** *J Biol Chem* 2008, **283**:20361-71.

30. Brouns SJ, Jore MM, Lundgren M, Westra ER, Slijkhuis RJ, Snijders AP, Dickman MJ, Makarova KS, Koonin EV, van der Oost J: **Small CRISPR RNAs guide antiviral defense in prokaryotes.** *Science* 2008, **321:**960-4.

> F1000 Factor 6.0 *Must Read*
> Evaluated by Malcolm White 15 Sep 2008

31. Carte J, Wang R, Li H, Terns RM, Terns MP: **Cas6 is an endoribonuclease that generates guide RNAs for invader defense in prokaryotes.** *Genes Dev* 2008, **22:**3489-96.

32. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25:**3389-402.

33. Han D, Krauss G: **Characterization of the endonuclease SSO2001 from Sulfolobus solfataricus P2.** *FEBS Lett* 2009, **583:**771-6.

34. Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, Rao BS, Smirnov S, Sverdlov AV, Vasudevan S, Wolf YI, Yin JJ, Natale DA: **The COG database: an updated version includes eukaryotes.** *BMC Bioinformatics* 2003, **4:**41.

> F1000 Factor 3.0 *Recommended*
> Evaluated by Shiladitya DasSarma 15 Oct 2003