nature REVIEWS MICROBIOLOGY

SUPPLEMENTARY INFORMATION

In format as provided by the authors

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

Kira S. Makarova, Yuri I. Wolf, Jaime Iranzo, Sergey A. Shmakov, Omer S. Alkhnbashi, Stan J. J. Brouns, Emmanuelle Charpentier, David Cheng, Daniel H. Haft, Philippe Horvath, Sylvain Moineau, Francisco J. M. Mojica, David Scott, Shiraz A. Shah, Virginijus Siksnys, Michael P. Terns, Česlovas Venclovas, Malcolm F. White, Alexander F. Yakunin, Winston Yan, Feng Zhang, Roger A. Garrett, Rolf Backofen, John van der Oost, Rodolphe Barrangou and Eugene V. Koonin

https://doi.org/10.1038/s41579-019-0299-x

Evolutionary classification of CRISPR-Cas systems 2019: explosion of Class 2 and derived variants

Makarova KS[1], Wolf YI[1], Iranzo J[1], Shmakov SA[1], Alkhnbashi OS[2], Brouns SJJ[3], Charpentier E[4], Cheng D[5], Haft DH[1], Horvath P[6], Moineau S[7], Mojica FJM[8], Scott D[5], Shah SA[9], Siksnys V[10], Terns MP[11], Venclovas C[10], White MF[12], Yakunin AF[13], Yan W[5], Zhang F[114,15,16,17], Garrett RA[18], Backofen R[2,19], Oost JvD[20], Barrangou R[21], Koonin EV[1,*]

1. National Center for Biotechnology Information, National Library of Medicine, Bethesda, MD 20894

2. Bioinformatics group, Department of Computer Science, University of Freiberg, Georges-Kohler-Allee 106, 79110 Freiberg, Germany.

3. Kavli Institute of Nanoscience, Department of Bionanoscience, Delft University of Technology, Van der Maasweg 9, 2629 HZ Delft, the Netherlands

4. Max Planck Unit for the Science of Pathogens, Humboldt University, Berlin, Germany

5. Arbor Biotechnology, Cambridge, MA, USA

6. DuPont Nutrition and Health, BP10, Dangé-Saint-Romain 86220, France.

7. Département de biochimie, de microbiologie et de bio-informatique, Faculté des sciences et de génie, Groupe de recherche en écologie buccale, Félix d'Hérelle Reference Center for Bacterial Viruses, Faculté de médecine dentaire, Université Laval, Québec City, Québec, Canada

8. Departamento de Fisiología, Genética y Microbiología. Universidad de Alicante. 03080-Alicante. Spain

9. COPSAC, Copenhagen Prospective Studies on Asthma in Childhood, Herlev and Gentofte Hospital, University of Copenhagen, Ledreborg Alle 34, 2820 Gentofte, Denmark

10. Institute of Biotechnology, Vilnius University, Vilnius, Lithuania.

11. Biochemistry and Molecular Biology, Genetics and Microbiology, University of Georgia, Davison Life Sciences Complex, Green Street, Athens, GA 30602

12. Biomedical Sciences Research Complex, University of St Andrews, North Haugh, St Andrews, KY16 9TZ, UK

13. Department of Chemical Engineering and Applied Chemistry, University of Toronto, Toronto, M5S 3E5, Canada.

14. Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142, USA

15. McGovern Institute for Brain Research, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

16. Howard Hughes Medical Institute, Cambridge, MA 02139

17. Department of Brain and Cognitive Sciences and Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139

18. Archaea Centre, Department of Biology, Copenhagen University, Ole Maaløes Vej 5, DK2200 Copenhagen N, Denmark

19. BIOSS Centre for Biological Signaling Studies, Cluster of Excellence, University of Freiburg, Germany.

20. Laboratory of Microbiology, Wageningen University, Stippeneng 4, 6708 WE Wageningen, Netherlands

21. Department of Food, Bioprocessing, and Nutrition Sciences, North Carolina State University, Raleigh, NC 27606, USA.

*For correspondence: koonin@ncbi.nlm.nih.gov

Supplementary Information

Supplementary Materials and Methods

The Prokaryotic Genome Database

A database containing 13,116 completely assembled archaeal and prokaryotic genomes in GenBank [1] was downloaded from the NCBI FTP site (ftp://ftp.ncbi.nlm.nih.gov/genomes/all/) in March 2019. The database contains 21,395,802 protein sequences annotated in 25,485 genome partitions. All proteins were clustered using MMseqs2 [2] with the sequence similarity threshold of 0.75; for every pair of genomes *i* and *j* the number $n_{i,j}$ of sequences, belonging to the same cluster, was determined, then the distances between the genomes were calculated as $d_{i,j} = -\ln n_{i,j}/\min(n_{i,i},n_{j,j})$. A UPGMA (Unweighted Pair Group Method with Arithmetic mean) dendrogram was reconstructed from genome distances, then genome weights were calculated from this tree as previously described [3].

CRISPR array detection

14,634 potential CRISPR arrays were identified in the sequence database using minCED tool (https://github.com/ctSkennerton/minced), derived from the CRT CRISPR recognition tool [4], with default parameters.

Identification of Cas proteins

Protein sequences were annotated using PSI-BLAST [5] (e-value cutoff of 10^{-4} and effective database size of 2×10^7 amino acids). Profiles from the NCBI CDD database [6], as well as a collection of previously described CRISPR-Cas protein family profiles [7-10] were used as queries. In addition, several profiles were constructed anew for groups of CRISPR-linked proteins that were poorly recognized with the existing profiles.

Alignments and phylogenetic analysis

Relationships within diverse protein families were established using the following procedure: initial sequence clusters were obtained using MMseqs2 [2] with the sequence similarity threshold of 0.5, and the sequences within each cluster were aligned using MUSCLE [11]. Alternatively, previously constructed cluster alignments were used. Then, cluster-to-cluster similarity scores were obtained using HHSEARCH [12] (including trivial clusters consisting of a single sequence each) and normalized by the minimum of the self-scores. Relative similarity scores were converted to distances using the $d = -\ln s$ formula, and a UPGMA dendrogram was constructed from the distance matrix. Highly similar clusters (pairwise score to self-score ratio >0.05) were aligned to each other using HHALIGN [12], and the procedure was performed iteratively. At the last step, sequence-based trees were constructed from the cluster alignments using the

FastTree program [13] (WAG evolutionary model, gamma-distributed site rates) and rooted by mid-point; these trees were grafted onto the tips of the profile similarity-based UPGMA dendrogram

Assembly of CRISPR-*cas* genomic islands

CRISPR-Cas genomic islands were assembled in three consecutive iterations. In each iteration, 15 ORFs flanking "anchor" *cas* genes were collected. The anchor *cas* genes were defined as genes belonging to the adaptation module (*cas1, cas2, cas4*) or the effector module (*cas3, cas5, cas6, cas7, cas8, cas10, cmr7, csb3, csx19, csx22, csx24, csx25, csx26* for Class 1; *cas9, cas12, cas13* for CRISPR-Cas Class 2). In the generated islands, CRISPR-Cas systems were identified and classified into types and subtypes using previously described procedures [7,10]. The annotation for Type II islands was manually corrected according to the phylogeny of *cas9*. Due to the high similarity between Type V effector proteins and the TnpB proteins encoded by transposable elements, phylogenetic analysis and manual curation were used to determine sequences that belonged to clades consisting of known type V effectors. Type VI systems were identified using both matches to the respective profiles and manual curation due to a large number of false positive hits. For the first and second iterations, all islands that were annotated as partial were manually analyzed to identify potential novel CRISPR-Cas protein families; all islands selected by this procedure were added to the island set. Temporary profiles were created for all representatives of protein families identified during the manual curation step using MAFFT [14]. To identify representatives of the newly detected protein families, the sequences identified during manual curation were used as PSI-BLAST search queries against the NR database [15], the sequences identified with this search were clustered with MMSeqs2 (with a 0.9 cutoff for sequence similarity), and the first member of each cluster was chosen as the representative of the respective alignment. The entire procedure was iterated using the updated CRISPR-Cas profile set.

Bipartite gene-sharing network of CRISPR-Cas systems

A bipartite gene-sharing network[16,17] of CRISPR-Cas systems was built from 2,077 representative CRISPR-Cas loci encompassing 15,198 protein sequences, which were partitioned into 4,687 protein clusters. To build the network, each protein cluster was connected to all those loci that contained at least one representative from that cluster. Conversely, every locus was connected to all those protein clusters that were represented in that locus. To facilitate the computational analysis, only non-redundant loci and protein clusters that appeared in at least 2 loci were included, yielding a network with 1,946 loci, 1,190 protein clusters, and 10,932 edges. To improve the performance of the community detection pipeline (see below), the network was split according to the classification of the loci into Class 1 and Class 2, and community detection was performed separately on each subnetwork.

Community detection in the bipartite gene-sharing network was carried out with the software MODULAR [18] which partitions the network into modules that maximize the bipartite modularity index [19]. Due to the combinatorial nature of the modularity maximization problem, programs such as MODULAR only can find approximate solutions when applied to large networks. To address this problem, a consensus clustering pipeline was implemented, which has been shown to greatly increase the

performance of modularity-based community detection approaches [20]. First, MODULAR was run 100 times; then, a pairwise similarity matrix was built by recording, for each pair of nodes, the fraction of runs in which both nodes were assigned to the same module; finally, hierarchical clustering of the nodes based on the pairwise similarity matrix (UPGMA method) was carried out. In the last step, the number of modules was selected such that the resulting partition maximized the bipartite modularity index. The software OSLOM [21], with options '-singlet -r 0 -hr 0 -t 0.05', was subsequently used to filter significant modules with a p-value threshold equal to 0.05.

Supplementary Figures

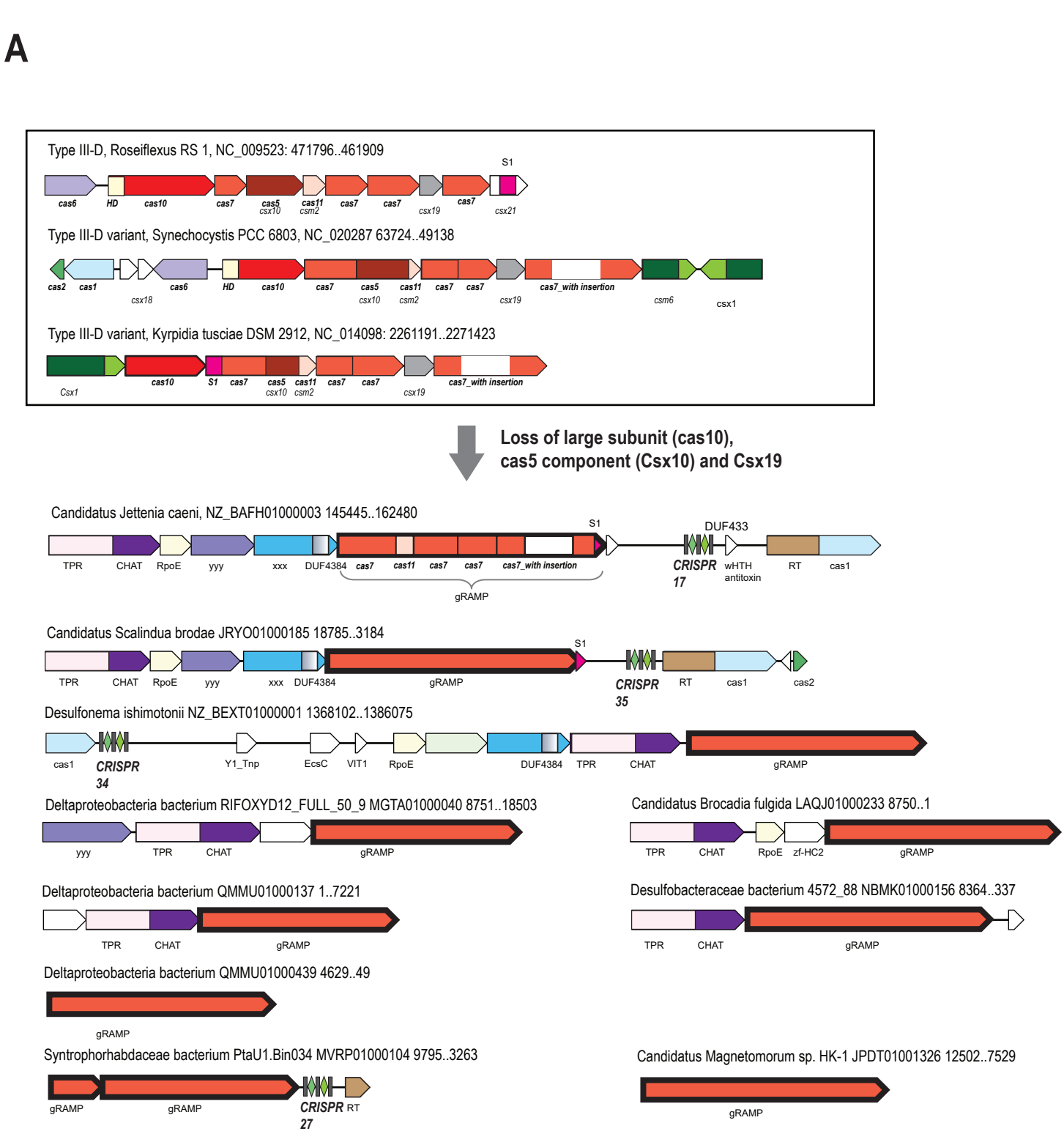
Supplementary Figure 1.

Subtype III-E CRISPR-Cas systems: Origin from subtype III-D and selected gene neighbourhoods

For each gene neighbourhood, the organism name, genome partition and coordinates of the locus are indicated. Genes are shown by block arrows, with the length roughly proportional to the size of the corresponding gene. The *cas* genes are coloured as in Figure 1. Other genes are coloured based on sequence similarity. Gene names or brief annotations are given below the arrows. White arrows correspond to genes that are apparently not related to the CRISPR-Cas function. Abbreviations: gRAMP, predicted multidomain subtype III effector; HD, HD nuclease domain; RT, reverse transcriptase; CHAT, caspase family protease (CHAT domain); TPR, tetratricopeptide repeats; zf, zinc finger; wHTH, winged helix-turn-helix domain; Y1_Tnp, tyrosine transposase; xxx and yyy, uncharacterized proteins associated with the caspase.

A. This panel shows several variants of the subtype III-D system inside the black box, including the one from *Kyrpidia tusciae* that shows the highest similarity with the gRAMP including a unique shared insert in Cas7 (see the domain organization of gRAMP depicted for Candidatus Jettenia caeni gRAMP). The III-D variant from *Kyrpidia tusciae* or a related bacterium likely gave rise to the gRAMP. Next to the arrow showing this origin, the inferred key events are described. Other subtype III-E loci from draft genomes are shown below.

B. Key motifs of the gRAMP protein (see details **in Supplementary Figure 2**).


A
Type III-D, Roseiflexus RS 1, NC_009523: 471796..461909
cas6 HD cas10 cas7 cas5 csx10 cas11 csm2 cas7 cas7 csx19 cas7 S1 csx21
Type III-D variant, Synechocystis PCC 6803, NC_020287 63724..49138
cas2 cas1 csx18 cas6 HD cas10 cas7 cas5 csx10 cas11 csm2 cas7 cas7 csx19 cas7_with insertion csm6 csx1
Type III-D variant, Kyrpidia tusciae DSM 2912, NC_014098: 2261191..2271423
Csx1 cas10 S1 cas7 cas5 csx10 cas11 csm2 cas7 cas7 csx19 cas7_with insertion
Loss of large subunit (cas10), cas5 component (Csx10) and Csx19
Candidatus Jettenia caeni, NZ_BAFH01000003 145445..162480
TPR CHAT RpoE yyy xxx DUF4384 cas7 cas11 cas7 cas7 cas7_with insertion S1 gRAMP DUF433 CRISPR 17 wHTH antitoxin RT cas1
Candidatus Scalindua brodae JRYO01000185 18785..3184
TPR CHAT RpoE yyy xxx DUF4384 gRAMP S1 CRISPR 35 RT cas1 cas2
Desulfonema ishimotonii NZ_BEXT01000001 1368102..1386075
cas1 CRISPR 34 Y1_Tnp EcsC VIT1 RpoE DUF4384 TPR CHAT gRAMP
Deltaproteobacteria bacterium RIFOXYD12_FULL_50_9 MGTA01000040 8751..18503
yyy TPR CHAT gRAMP
Candidatus Brocadia fulgida LAQJ01000233 8750..1
TPR CHAT RpoE zf-HC2 gRAMP
Deltaproteobacteria bacterium QMMU01000137 1..7221
TPR CHAT gRAMP
Desulfobacteraceae bacterium 4572_88 NBMK01000156 8364..337
TPR CHAT gRAMP
Deltaproteobacteria bacterium QMMU01000439 4629..49
gRAMP
Syntrophorhabdaceae bacterium PtaU1.Bin034 MVRP01000104 9795..3263
gRAMP gRAMP CRISPR 27 RT
Candidatus Magnetomorum sp. HK-1 JPDT01001326 12502..7529
gRAMP
B
FeS cluster
abnormal G-loop
"W" motif
FeS cluster
G-loop
"D" catalytic
FeS cluster
"D" catalytic
G-loop
FeS cluster
G-loop
small subunit
insertion
gRAMP (III-E)
WP_007220849.1 Candidatus Jettenia caeni

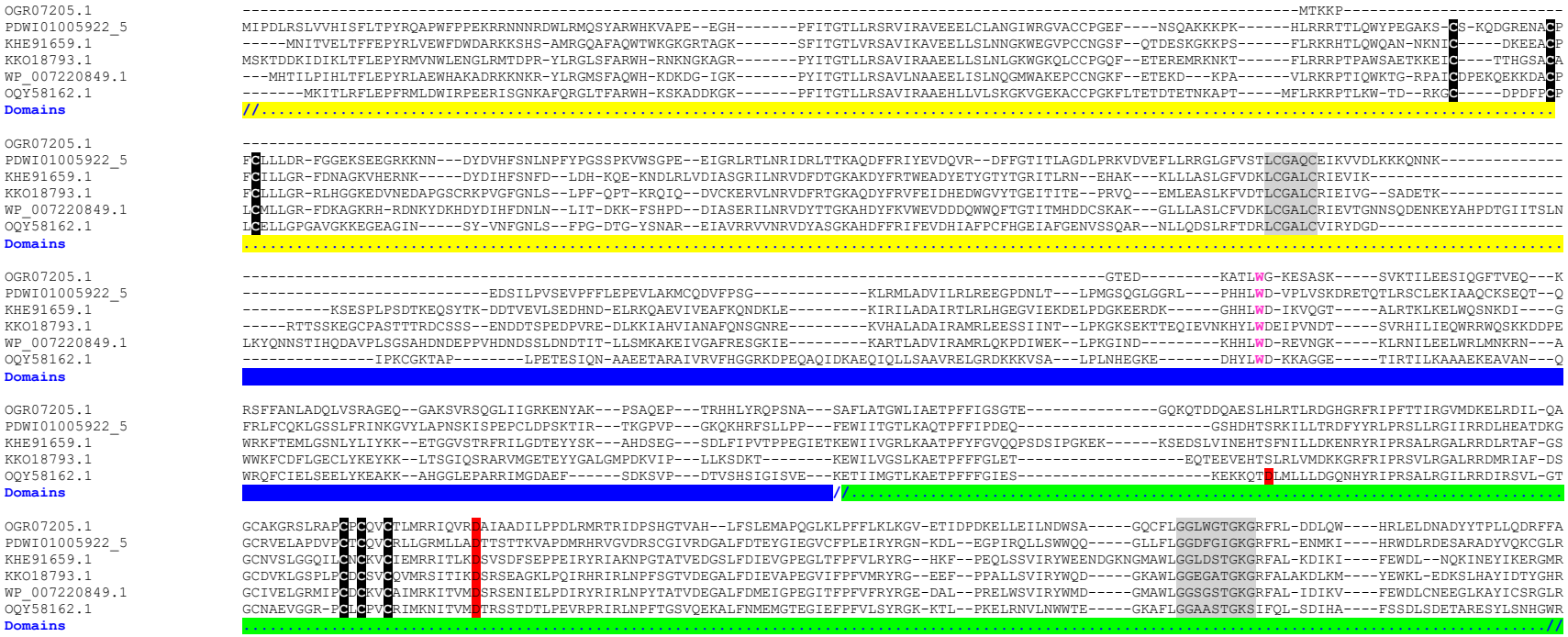
Supplementary Figure 2

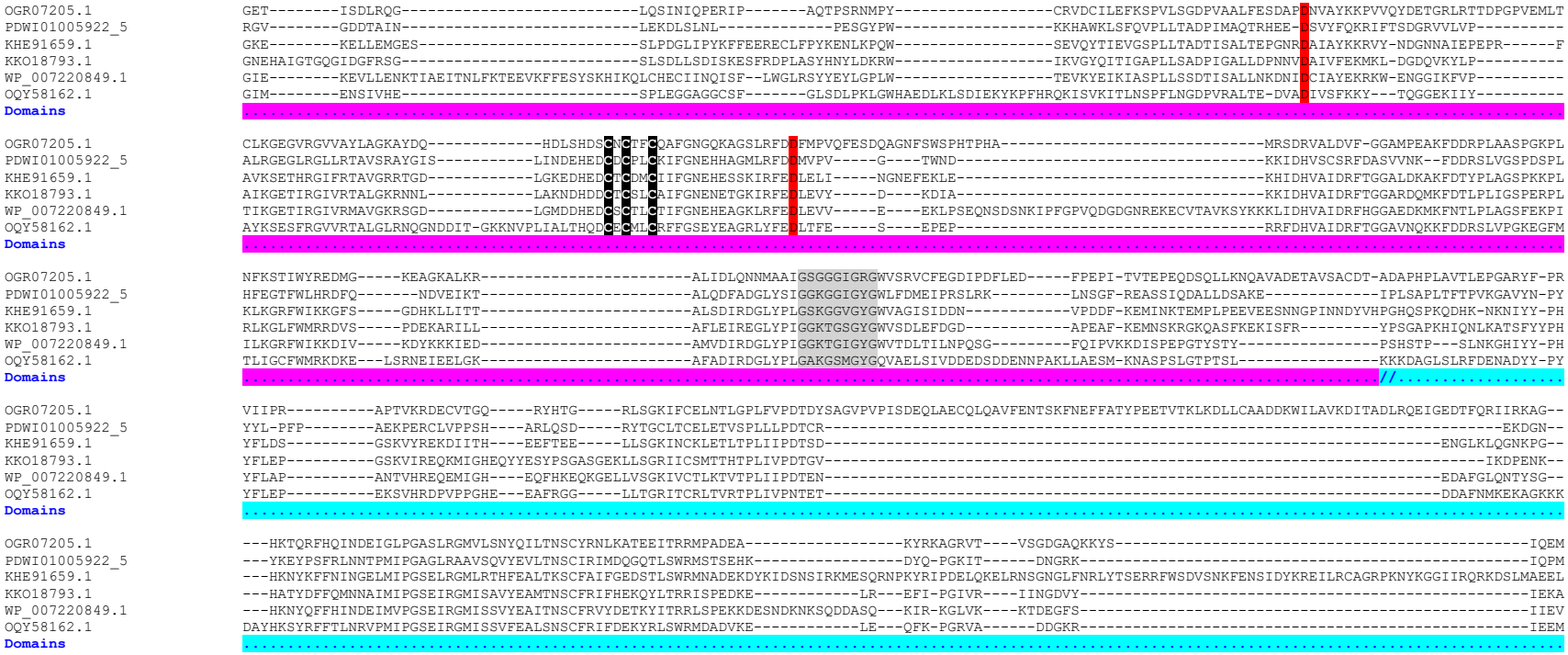
Multiple alignment and HHpred outputs for gRAMP (giant RAMP), the predicted effector module of subtype III-E systems.

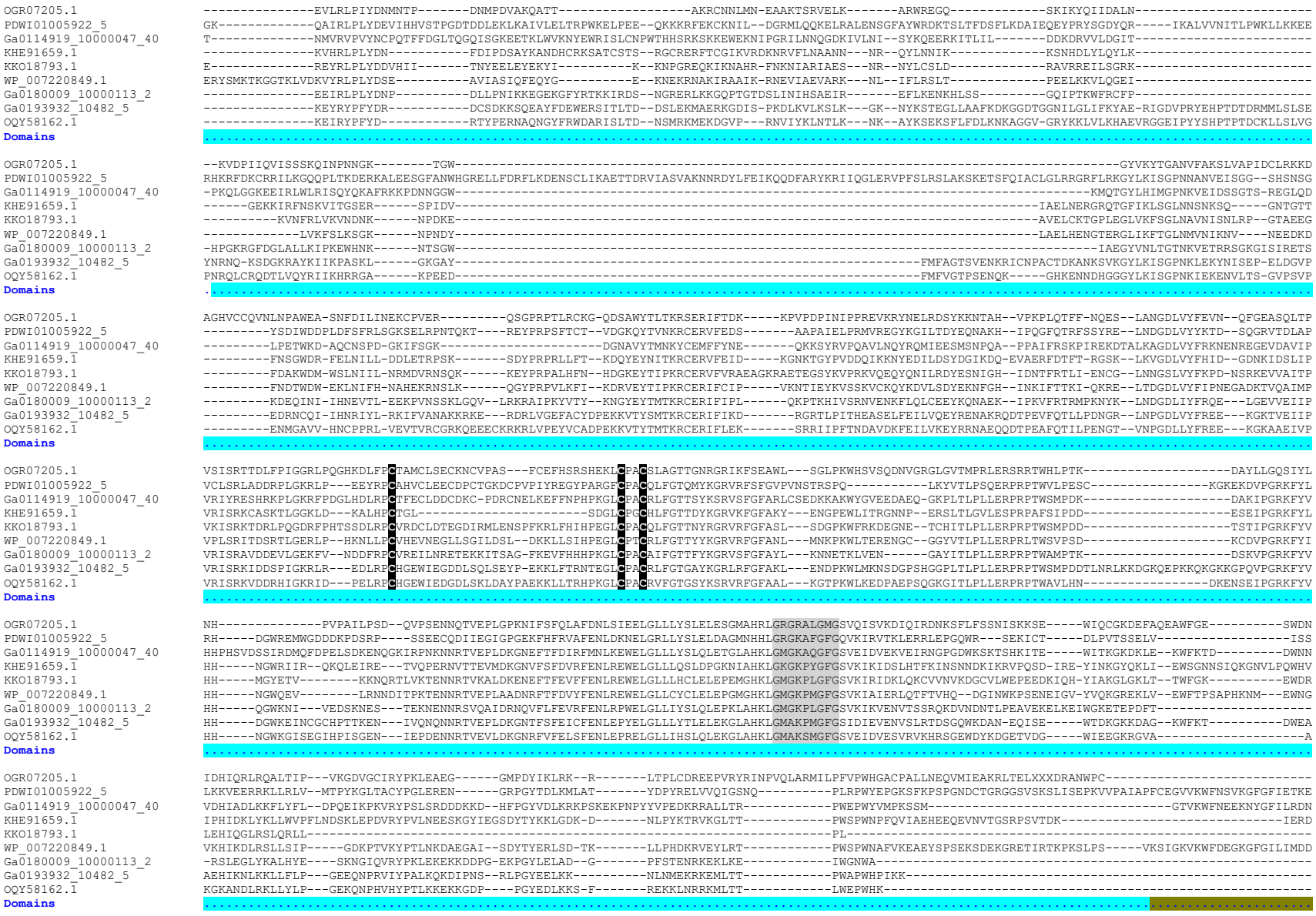
Selected representatives of the gRAMP (predicted III-E effector) family were aligned using MUSCLE and colored using http://www.bioinformatics.org/sms2/color_align_cons.html server with default amino acid groups with 90% consensus. Domains and catalytic residues are colored according to the scheme explained above the alignment.

The standard HHpred server (https://toolkit.tuebingen.mpg.de/#/tools/hhpred) output is shown. The search was done with OQY58162.1 protein as a query. The catalytic residues are highlighted.

Domain 1 - Csm3-like (group 7 RAMP) with abnormal “G-rich” loop
Domain 2 - putative small subunit (no sequence similarity but there is a conserved “W” and four large alpha helices predicted by Jpred in this region)
Domain 3 - Csm3-like (group 7 RAMP), could be catalytic
Domain 4 - Csm4-like (group 7 RAMP), red - catalytic aspartates based on Zhu X, Ye K. Nucleic Acids Res. 2015 Jan;43(2):1257-67; https://doi.org/10.1093/nar/gku1355)
Domain 5 - Csm3-like with large insertion or subdomain (group 7 RAMP)
Domain 6 - Cold shock protein-like domain (RNA-binding OB-fold)
G-rich loop
Fe-S clusters - conserved cysteines




OGR07205.1
PDWI01005922_5
KHE91659.1
KKO18793.1
WP_007220849.1
OQY58162.1
Domains


OGR07205.1
PDWI01005922_5
Ga0114919_10000047_40
KHE91659.1
KKO18793.1
WP_007220849.1
Ga0180009_10000113_2
Ga0193932_10482_5
OQY58162.1
Domains

```
OGR07205.1                ---------------------------------------------
PDWI01005922_5            QRDIFVHFSAIRGEGYKILEPGEKVRFEIGEGRKGPQAINVIRIR
Ga0114919_10000047_40     GEDIFVHRSDINGS-LGTLTEGQKVIFEVKQGPKGLQATNVKVIS
KHE91659.1                GKMV-----------------------------------------
KKO18793.1                ---------------------------------------------
WP_007220849.1            GKEVSISKNSIRGN--ILLKKGQKVTFHIVQGLI-PKAEDIEIAK
Ga0180009_10000113_2      ---------------------------------------------
Ga0193932_10482_5         ---------------------------------------------
OQY58162.1                ---------------------------------------------
Domains                   .............................................
```

//

HHpred search results (OQY58162.1 protein as a query):

```
Csm3 (group 7),
Probability: 95.58    E-value: 0.012    Score: 50.24    Aligned Cols: 144    Identities: 21%    Similarity: 0.244

Q ss_pred                   CCccchHHHHHHHHHHHHHHHHhhhcCccCceecCCceeccccccccCCCCCceeccCCccccCCCCCC
Q Q_4784969         46      GKPFITGTLLRSAVIRAAEHLLVLSKGKVGEKACCPGKFLTETDTETNKAPTMFLRKRPTLKWTDRKGCD  115 (413)
Q Consensus         46      GkPyITGTLLRSaVi~a~E~ll~l~~gk~~~~~CC~g~f~Te~dk~~~~~~p~FLRKRPt~~w~~~~~Cd  115 (413)
                            |.|||-||-++-++..+++.+..........                                 ....|.
T Consensus         41      g~p~IPgsslKG~lR~~~~~~~~~~~~~~~~~~~---------------------------------~~~~~~   77 (204)
T TIGR02582         41      DLPYIPGSSLKGKIRSLLERSLGAVSNKGKY---------------------------------RYHNCD   77 (204)
T ss_pred                   CCeecchhhHHHHHHHHHHHHhhhhcccccc---------------------------------ccCCCC

Q ss_pred                   CCCCCcchhhcCCCCcCccccCCccccceeeecccCCCCCC-----CCCcHHHHHHHHHhHHhhhhccccc
Q Q_4784969         116     PDFPCPLCELLGPGAVGKKEGEAGINSYVNFGNLSFPGDT-----GYSNAREIAVRRVVNRVDYASGKAH  180 (413)
Q Consensus         116     p~~~CplC~lLGr~d~ggK~~~~g~~~~VhF~N~s~~~~~~-----~~~~~dIa~~RILNRVD~~tGkA~  180 (413)
                             ...|+.|.++|........ .......+.|++.......     ...........|.-+++|..||.|.
T Consensus         78      -~~~~~~~~~~lFG~~~~~~~~-~~~~~~g~l~~~D~~~~~~~~~~~~~~~~~~~~~~~~~~~~~idr~t~~~~  145 (204)
T TIGR02582         78      -PGECEVCRLFGSAPENDEE-SIKGPTRLIVRDAFLTEDTKEELARNDTDLPYTEIKFENAIDRITSKAN  145 (204)
T ss_pred                   -CCccHHHHHHCCCCCCCCc-cCCCCceEEEEEeeeCcccHHHHhcCCCCCCcceeeecccccCCCCCcC
Q ss_pred                   cceEEEEEccccCCeeeEEEEcCCchhH---HHHHHHHHHHHHH--HHhhce
Q Q_4784969         181     DFFRIFEVDHIAFPCFHGEIAFGENVSS---QARNLLQDSLRFT--DRLCGA  227 (413)
Q Consensus         181     DyFrv~EvD~~~~~~~ftG~I~i~d~~~S---~~k~LL~~sL~fv--d~LcGa  227 (413)
                            + |..-.|  ..-..|+++|.+......   ....+|..+|..+  ..|.|.
T Consensus         146     ~-~~~e~i--~~g~~f~~~i~~~~~~~~~~~~~~~~~~~~l~~~l~~l~~~~iG~~  194 (204)
T TIGR02582         146     P-RQIERV--PAGAKFKFEIIYSVYEGDEEEEDFKNILEGLKLLEDDYLGGS  194 (204)
T ss_pred                   C-eeEEEe--CCCCEEEEEEEEEEeCCCCCchHHHHHHHHHHHHcccCCCCCC
```

```
4WNZ_B Cmr4; Nuclease, RAMP domain, CRISPR-Cas system; 2.8A {Pyrococcus furiosus DSM 3638}; Related PDB entries: 4W8W_C 4W8W_B 4W8W_A 4W8W_D 4RDP_B 4RDP_A 4WNZ_A
Probab=99.70  E-value=2.1e-19  Score=210.94  Aligned_cols=175  Identities=19%  Similarity=0.156  Sum_probs=0.0  Template_Neff=7.900

Q ss_pred                EEEEEEEcCCeeecCCCCCCCccccchhEeCC-CCCEEeCchHHHHHHHHHHHHHhc--cCcccccCCCCCCCcceeeee
Q Q_4784969       402 IMGTLKAETPFFFGIESKEKKQTDLMLLLDG-QNHYRIPRSALRGILRRDIRSVLG--TGCNAEVGGRPCLCPVCRIMKN  478 (1659)
Q Consensus       402 i~g~Le~lSPLhIGSGe~~~~~~D~~L~RD~-~~G~PyIPGTSLAGaLRa~l~~~~~--~~~~~~LFG~~~~~~~~~~~~~S~  478 (1659)
                        +.+++++.||||||+|.+.. .+|.+|+||. .|.||||||||+|+||+++.....  ......+||.+... .....+.
T Consensus         4 ~~~~i~~~splhiG~G~~~~~-~~D~~v~rd~~~g~P~IPGSslKG~lR~~~~~~~~~~~~~~~~~~~~~~~lFG~~~~~-~~~~~~g~   81 (301)
T 4WNZ_B            4 YLVGLYTLTPTHPGSGTELG-VVDQPIQRERHTGFPVIWGQSLKGVLRSYLKLVEKVDEEKINKIFGPPTEK-AHEQAGL   81 (301)
T ss_dssp              EEEEEEEESSCBCC------------CCCBCTTTCCBCBCHHHHHHHHHHHHHHTSCCCHHHHHHHHTC-----------CC
T ss_pred                EEEEEEEcCCeecCCCCCcC-CccccEeEecCCCCeEeecchHHHHHHHHHHHhccCChHHHHHHHCCCCcc-cccccce


Q ss_pred                EEEEeccCC--C--------------------------------------------------------------------
Q Q_4784969       479 ITVMDTRSS--T--------------------------------------------------------------------  488 (1659)
Q Consensus       479 l~V~Da~l~--~--------------------------------------------------------------------  488 (1659)
                        +.|.|+.+.  .
T Consensus        82 ~~f~da~l~~~p~~~~~~~~~~~~~~~~~vtcp~~l~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~  161 (301)
T 4WNZ_B           82 ISVGDAKILFFPVRSLKGVYAYVTSPLVLNRFKRDLELAGVKNFQTEIPELTDTAIASEEITVDNKVILEEFAILIQKDD  161 (301)
```

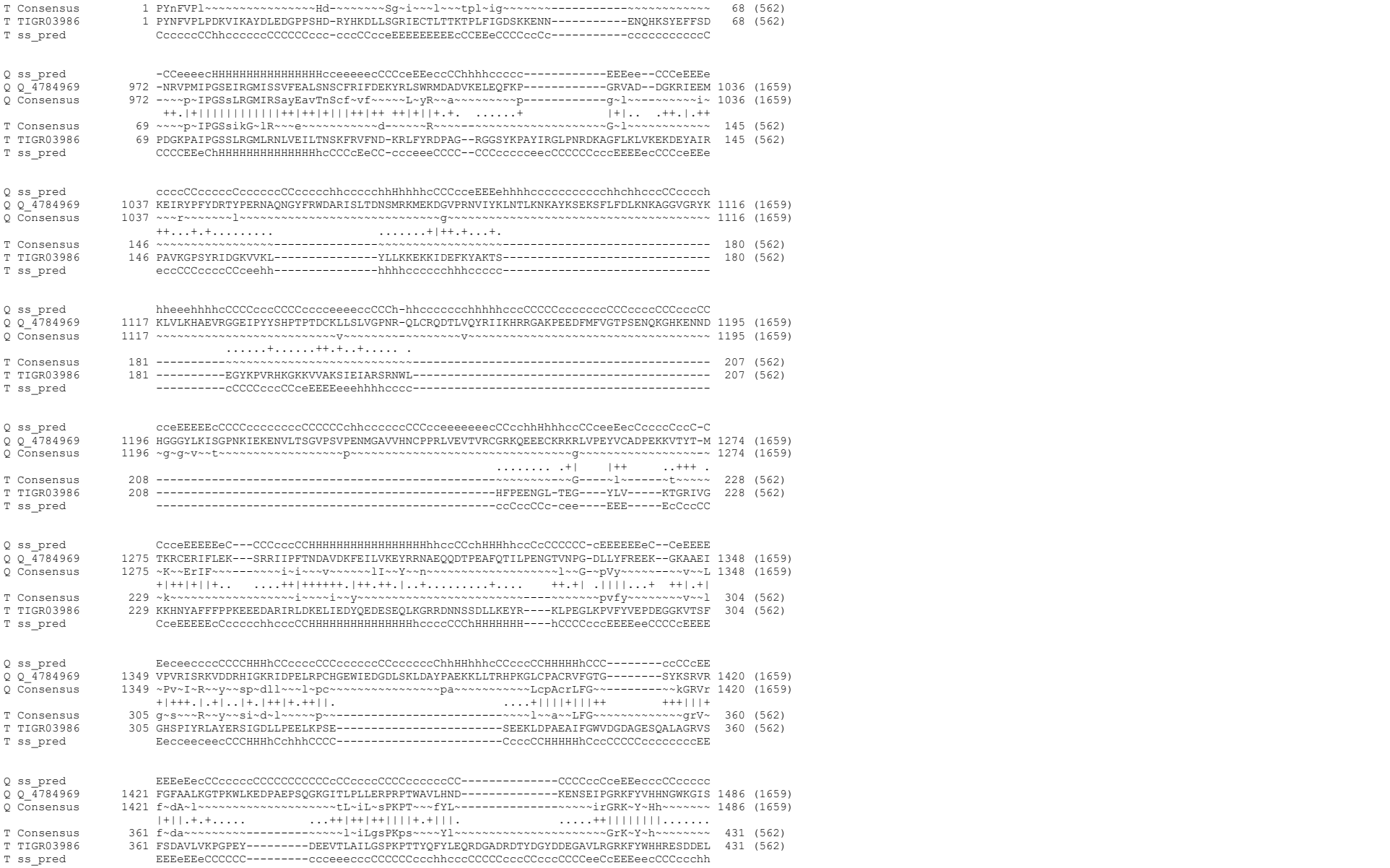
```
T ss_dssp              EEECCBEEEEEEEEEBSSSSEEEEECHHHHHHHHHHHHHHTTC----CCCCCCSSEEEECTTTBSSSEEEETTEEEEECCCT
T ss_pred              EEEeceEEeeeecccCCCCeEEeccHHHHHHHHHHHHHcCCCCCccCCchhcccccccccccCCEEEEEehhheeecCc


Q ss_pred              ---------------------------------------CCCCccccceeecCCCCccccCCeeeeEEEeCCCcEEEEEEE
Q Q_4784969        489 ---------------------------------------DTLPEVRPRIRLNPFTGSVQEKALFNMEMGTEGIEFPFVLS  529 (1659)
Q Consensus        489 ---------------------------------------~~~t~vR~gVrIDr~TGTA~dgaLFd~EvVP~Gt~F~f~Le  529 (1659)
                                                       ...+.+|++|+||+.||+|+++.||+.|+||+|+.|.+.|.
T Consensus        162 ~~~~~~~~~~l~~~f~~~~~~~~~~~~~~~iv~d~~~~~l~~~~~~~~r~~v~Idr~tG~a~~g~Lf~~E~v~~g~~F~~~i~  241 (301)
T 4WNZ_B           162 KGILESVVKAIEQAFGNEMAEKIKGRIAIIPDDVFRDLVELSTEIVARIRINAETGTVETGGLWYEEYIPSDTLFYSLIL  241 (301)
T ss_dssp              TCHHHHHHHHHHHHHCHHHHHHHTTSEEECCHHHHHHHHHSCCC----------------------CBCSSCEEEEEEE
T ss_pred              chhHHHHHHHHHHHhChHHHHHhhCCEEEEehHHHHHHHHhCceEEEecEeeCCCCccccCCCeeEEEEeCCCcEEEEEEE


Q ss_pred              EeCCcc--hhHHHHHHHHHHHcCCCcCCCCcCCCCCEEEEeecEEEeccCC
Q Q_4784969        530 YRGKKT--LPKELRNVLNWWTEGKAFLGGAASTGKSIFQLSDIHAFSSDLS  578 (1659)
Q Consensus        530 l~~~~~~--~~~lL~~LL~~l~~G~irLGg~rsRGfGRvrL~~~~~~~~~~Dl~  578 (1659)
                       +.....  ..+++..+|..+..+..+||+++++|+|+|++.......-+..
T Consensus        242 ~~~~~~~~~~~~~~~l~~~L~~l~~~~~~~lGg~~srG~G~v~i~~~~~~~~~~~~~~~~  292 (301)
T 4WNZ_B           242 VTPRAKDNDMALIKEVLGKINGKYLQIGGNETVGKGFVKVTLKEVTNNGGT  292 (301)
T ss_dssp              ECTTCCHHHHHHHHHHHHTTTTEEEECSSCTTTTCCEEEEEEEEC------
T ss_pred              EccCCCCCchHHHHHHHHHhcCCceEeCCCccccccCEEEEEEeeecCCCcc


>4WNZ_B Cmr4; Nuclease, RAMP domain, CRISPR-Cas system; 2.8A {Pyrococcus furiosus  DSM 3638}; Related PDB entries: 4W8W_C 4W8W_B 4W8W_A 4W8W_D 4RDP_B 4RDP_A 4WNZ_A
Probab=99.23  E-value=5.9e-13  Score=157.06  Aligned_cols=175  Identities=15%  Similarity=0.155  Sum_probs=0.0  Template_Neff=7.900

Q ss_pred              CeEEEEEEEECCCCEEeCCccccCCcccccccccccccCCCC-ceEEEechHHHHHHHHHHHHHHccCCCCccccCccCCchh
Q Q_4784969        643 RQKISVKITLNSPFLNGDPVRALTEDVADIVSFKKYTQGGE-KIIYAYKSESFRGVVRTALGLRNQGNDDITGKKNVPLI  721 (1659)
Q Consensus        643 ~~~I~l~Lt~~sPLLI~~~~~~~~~~~~~~~pD~~~l~~~~~~~g~~~~~~vIPGSSIRGALR~r~~lrtl~~~~~~~~~~~a~~~~  721 (1659)
                       ...+.+.++..+|++|+++.....   .|.+....      .  .+.|+||||||||+||+.+......
T Consensus          1 m~~~~~~i~~~splhiG~G~~~~~~~~~---D~~v~rd--------g~P~IPGSslKG~lR~~~~~~~~~~~~~~-------------   59 (301)
T 4WNZ_B             1 MKAYLVGLYTLTPTHPGSGTELGV---VDQPIQRE-----RHTGFPVIWGQSLKGVLRSYLKLVEKV-------------   59 (301)
T ss_dssp              -CEEEEEEEESSCBCC--------------CCCBC-----TTTCCBCBCHHHHHHHHHHHHHHTSCC-------------
T ss_pred              CceEEEEEEEcCCeecCCCCCcCC---ccccEeEe-----cCCCCeEeecchHHHHHHHHHHhccC-------------


Q ss_pred              hhcccccchhhHHHhCCc-----ceeeeEEEEeeeecCCC----------------------------------------
Q Q_4784969        722 ALTHQDCECMLCRFFGSE-----YEAGRLYFEDLTFESEP----------------------------------------  756 (1659)
Q Consensus        722 ~~~~~~~~~~~~v~~LFGs~~~~~~~~~~~rG~L~vsD~~i~~~~~------------------------------------------  756 (1659)
                             ..+.++.|||..     ..+|.|.|.|+.+...+
T Consensus         60 ------~~~~~~~~lFG~~~~~~~~~~~~~g~~~f~da~l~~~p~~~~~~~~~~~~~vtcp~~l~~~~~~~~~~~~~~~~~~~~~~~~~~  133 (301)
T 4WNZ_B            60 ------DEEKINKIFGPPTEKAHEQAGLISVGDAKILFFPVRSLKGVYAYVTSPLVLNRFKRDLELAGVKNFQTEIPELT  133 (301)
T ss_dssp              ------CHHHHHHTC----------CCEEECCBEEEEEEEEBSSSSEEEEECHHHHHHHHHHHHHHTTC----CCCCCCS
T ss_pred              ------ChHHHHHHCCCCcccccccceEEEeceEEeeeecccCCCCeEEeccHHHHHHHHHHHHHcCCCCCccCCchhc


Q ss_pred              ---------------------------------------------------------------------CcceEecccccc
Q Q_4784969        757 ---------------------------------------------------------------------EPRRFDHVAID  767 (1659)
Q Consensus        757 ---------------------------------------------------------------------~~~~~~~HVAID  767 (1659)
                                                                                     ....++||+||
T Consensus        134 ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~l~~~f~~~~~~~~~~~~~~~iv~d~~~~~l~~~~~~~~r~~v~Id  213 (301)
T 4WNZ_B           134 DTAIASEEITVDNKVILEEFAILIQKDDKGILESVVKAIEQAFGNEMAEKIKGRIAIIPDDVFRDLVELSTEIVARIRIN  213 (301)
T ss_dssp              SEEEECTTTBSSSEEEETTEEEEECCCTTCHHHHHHHHHHHHHCHHHHHHHTTSEEECCHHHHHHHHHSCCC--------
T ss_pred              cccccccccccCCEEEEEehhheeecCcchhHHHHHHHHHHHHhChHHHHHhhCCEEEEehHHHHHHHHhCceEEEecEee


Q ss_pred              CCCCCCCCCCCeeEEEeCCCCccceEEEEEEEecCcccCCHHHHHHHHHHHHHHHHcCCCccCCCCCCcccEEEEEEEeCC
Q Q_4784969        768 RFTGGAVNQKKFDDRSLVPGKEGFMTLIGCFWMRKDKELSRNEIEELGKAFADIRDGLYPLGAKGSMGYGQVAELSIVDD  847 (1659)
Q Consensus        768 RFTGGa~DGALFteeal~~G~~~~~~~~l~l~I~i~~~~~~~~~~~~~~~l~LLllaLrDL~~G~LpLGggtarGyG~v~v~~i~~~~  847 (1659)
                 |.||.+.+|+||+++++.+|.    .|...|.+...  ..++...+|..+|+++..+.+.||++++++|||+|++......
T Consensus        214 r~tG~a~~g~Lf~~E~v~~g~-----F~~~i~~~~~~~~~~~~~~~~~~~l~~~~L~~l~~~~~~~lGg~~srG~G~v~i~~~~~~~  287 (301)
T 4WNZ_B           214 AETGTVETGGLWYEEYIPSDT----LFYSLILVTPR--AKDNDMALIKEVLGKINGKYLQIGGNETVGKGFVKVTLKEVT  287 (301)
T ss_dssp              ---------------CBCSSC----EEEEEEEECTT--CCHHHHHHHHHHHHTTTTEEEECSSCTTTTCCEEEEEEEEC-
T ss_pred              CCCCccccCCCeeEEEEeCCCCc----EEEEEEEEccC--CCCCchHHHHHHHHHhcCCceEeCCCccccccCEEEEEEeeec
```

```
Q ss_pred            CCC
Q Q_4784969       848 EDS  850 (1659)
Q Consensus       848 ~~~  850 (1659)
                      ...
T Consensus       288 ~~~  290 (301)
T 4WNZ_B          288 NNG  290 (301)
T ss_dssp             ---
T ss_pred             CCC
```

```
>cd09683 Csm3_III-A; CRISPR/Cas system-associated RAMP superfamily protein Csm3. CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats) and associated Cas proteins
comprise a system for heritable host defense by prokaryotic cells against phage and other foreign DNA.
Probab=96.63  E-value=0.00011  Score=72.95  Aligned_cols=57  Identities=25%  Similarity=0.390  Sum_probs=0.0  Template_Neff=9.800

Q ss_pred            EEEEEEEEcCCEEccCCCCCchhhhhhhccCccccccccccceee-CCeEeecHHHHHHHHHHHHHHHHcc
Q Q_3743303_3      50 RITCRLTVRTPLIVPNTETDDAFNMKEKAGKKKDAYHKSYRFFTL-NRVPMIPGSEIRGMISSVFEALSNS  119 (782)
Q Consensus        50 ~I~c~Ltt~TPL~I~~~~~~~~~~~~~~~~~~~~hk~y~ff~~-~~~p~IPGSsLRGMIRSayEavTnS  119 (782)
                      .|.++|++.|||.|+.....            .+... .+.|+|||||||+||+++|.+..+
T Consensus         5 ~~~~~l~~~s~l~ig~g~~~~~~~~d~~~~~~~~~~~~~~~~~g~p~IPgSSlKG~lR~~~~~~~~   62 (216)
T cd09683           5 KISGILENQTPLRIGYGKSQEPTGPDD-------------PVIKDANGRPIIPGSSFKGALRSLAERYLKR   62 (216)
T ss_pred            EEEEEEEECCCEEeccCCCCCCCCCCC-------------CceeCCCCCEeeChhHHHHHHHHHHHHHHHH
```

```
>cd09684 Csm3_III-A; CRISPR/Cas system-associated RAMP superfamily protein Csm3. CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats) and associated Cas proteins
comprise a system for heritable host defense by prokaryotic cells against phage and other foreign DNA.
Probab=97.01  E-value=0.00017  Score=71.25  Aligned_cols=126  Identities=24%  Similarity=0.326  Sum_probs=0.0  Template_Neff=10.000

Q ss_pred            ccCCHHHHHhCCccc-------CCCeEEEEEEEcCCcccccCCCCCCCCCCCccCcCcCCCCCcchhhccCCCCCeeCeE
Q Q_3743303_3     524 KGLCPACRVFGTGSY-------KSRVRFGFAALKGTPKWLKEDPAEPSQGKGITLPLLERPRPTWAVLHNDKENSEIPGR  596 (782)
Q Consensus       524 ~~LcpAcrLFG~~a~--------~GRVrFsdA~~~~~~~~~~~~~~~~~~~~~L~iL~sPKPs~~~fYl~~~~~~lrGR  596 (782)
                      ...|+.|+|||....       +|+|+|+||.+..              .....+.      +.         .......
T Consensus        77 ~~~~~~~~lFG~~~~~~~~~~~g~l~~~D~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~  127 (215)
T cd09684          77 QGDCEVCRLFGSAPNAGEGIKGPTRLIVRDAFLTE--------------DTKKEAE------ST---------DLPYTEI  127 (215)
T ss_pred             CCCcchHHHhCCCCCCCCCCCCeeEEEEcceeCh--------------hhhhhhh------cC---------CCCcceE


Q ss_pred            EEEecCCCccccccccccCCCCCCCCcccceeEEEECCCCCEEEEEEEEeCCC----HHHHHHHHHHcCCCCCceeEcCCC
Q Q_3743303_3     597 KFYVHHNGWKGISEGIHPISGENIEPDENNRTVEVLDKGNRFVFELSFENLE----PRELGLLIHSLQLEKGLAHKLGMA  672 (782)
Q Consensus       597 KfY~H~~~~~~~~~~~~~~~~~~~~~~~~~~~~~i~~v~~G~~F~f~I~FeNLs------~ELGaLLwaL~l~~~~~HKLG~G  672 (782)
                      +.+-.....               .......++.+.+|++|+++|.+.+..    +.++.+|+.+|..-...  .||.+
T Consensus       128 ~~~~~id~~~~~~~~~~~~~~~~g~~~~~~~e~i~~g~~f~~~i~~~~~~~~~~~~~l~~~l~~~~~~~~~lG~~  190 (215)
T cd09684         128 KFENAIDRIT----------------SKANPRQIERVPAGAKFKFELIYSVYEGDDEEEDFKNILEGLKLLEDD--YLGGS  190 (215)
T ss_pred             EeeeccCCCC----------------CCcCCceeEEeCCCCEEEEEEEEEecCCCCcHHHHHHHHHHHHHhhcC--CCCCC


Q ss_pred            CCCCcceEEEEEEEEEEcccce
Q Q_3743303_3     673 KSMGFGSVEIDVESVRVKHRSGE  695 (782)
Q Consensus       673 KPLGlGSVkI~I~~l~l~~~~~  695 (782)
                      +.-|+|.|++++..+...+...+
T Consensus       191 ~s~G~G~~~~~~~~~~~~~~~~  213 (215)
T cd09684         191 GSRGYGKVKFRDIKLITKPGEYY  213 (215)
T ss_pred             CCCCcceEEEEEEEEEeccCcc
```

```
TIGR03986 TIGR03986; CRISPR-associated protein. Members of this protein family, part of the larger RAMP family, are found exclusively in species with CRISPR systems, in local
contexts containing other RAMP (Repeat-Associated Mystery Proteins).
Probab=100.00  E-value=2.5e-57  Score=568.13  Aligned_cols=500  Identities=26%  Similarity=0.416  Sum_probs=0.0  Template_Neff=8.200

Q ss_pred            ccccCCCcccCCC----CCCCCcchhccCCceeEEEEEEEEcCCEEccCCCCChhhchhhhcCCcccccccccccceee
Q Q_4784969       896 PYYFLEPEKSVHRD----PVPPGHEEAFRGGLLTGRITCRLTVRTPLIVPNTETDDAFNMKEKAGKKKDAYHKSYRFFTL  971 (1659)
Q Consensus       896 PY~Fv~~~~v~~d------p~~h~~~~~~~~SG~I~ctLtt~TPL~I~d~~~~~~~~~~~~~~~~~~hk~y~ff~~  971 (1659)
                      ||||||+++.+...    ..|.+|+ .+..+++||+|+|+|++.|||+|+++.....           ...+..+.||..
```



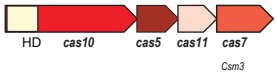
```
Q ss_pred               cCCCCCCCCCCCCCcceEEEEEccCCCeEEEEEEEeCCCHHHHHHHHHhccCCCCceeECCCCCCCccceEEEEEEeeEE
Q Q_4784969        1487 EGIHPISGENIEPDENNRTVEVLDKGNRFVFELSFENLEPRELGLLIHSLQLEKGLAHKLGMAKSMGFGSVEIDVESVRV 1566 (1659)
Q Consensus        1487 ~~~~~~~~~~~~~~~~n~tI~~l~pG~~F~F~I~FENLS~~ELGaLLW~L~l~~~~~HkLGmGKPLGfGSVkI~i~~l~l 1566 (1659)
                        ..   ........+++.+++++|.+|++|+|+|+|+||+++|||+|||+|++.+.++|+||+|||+|||+|+|++..+.+
T Consensus         432 ~~---~~~~~~~~~~~~~~~~~~~v~~g~~F~~~I~f~nl~~~ELg~Ll~aL~~~~~~~~~~~lG~gKp~G~G~v~i~~~~~~~~~  508 (562)
T TIGR03986        432 EK---NSENPNKNEKLNTVIKPVKEGTTFRFRIRFENLSPVELGALLWALDLGEGWYHKIGMGKPLGFGSVKITVDGLFL  508 (562)
T ss_pred               hh---cCCCCCccccccEEEEEeCCCCEEEEEEEEeCCCHHHHHHHHHHhccCCCccceecCCCCCccceEEEEEEeeEEE


Q ss_pred               EecCCCceeecCCCchH---------HHHHHHHH----HHhhhhhHHHHHHHhcC
Q Q_4784969        1567 KHRSGEWDYKDGETVD---------GWIEEGKR----GVAAKGKANDLRKLLYL 1607 (1659)
Q Consensus        1567 ~~~~~~w~~~~~~~~~~~~~~---------~~~~~~~~~~----~~~~~~~~~i~~L~~il~~ 1607 (1659)
                        ......+..-......         +|.++|..    .+.+.+.+++|++|+++
T Consensus         509 ~~~~~~y~~~~~~~~~~~~~~~~~~~~~~~~~f~~~~~~~~~~~~~~~~~~~~~~~~~i~~l~~~~~~  562 (562)
T TIGR03986        509 NRKERYSSLFASDDATEKVDEYIEAFKKYMEEKLGSSLKETAQIEDLKALLSW  562 (562)
T ss_pred               EchHHhhhhccCCCchhHHHHHHHHHHHHHHHHHHhccCHHhhHHHHHHHHHhcC
```

Supplementary Figure 3

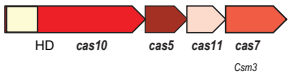
Selected gene neighbourhoods of subtype III-F CRISPR-Cas systems

Designations and abbreviations are the same as in Supplementary Figure 1. See also the system description in the text.

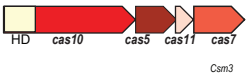
Pseudothermotoga lettingae TMO NC_009828 98639..102659



Thermotoga sp. KOL6 NZ_LNDE01000001 206594..202369

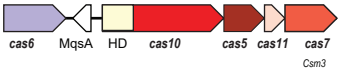


Ferroglobus placidus DSM 10642 CP001899 252831..249216

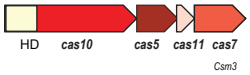


Anoxybacter fermentans NZ_CP016379 2618433..2614060

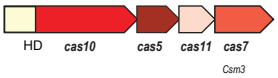
Nitrospinae bacterium RIFCSPLOWO2_02_39_17 MHDL01000098 6188..1191



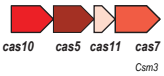
Ferroglobus placidus DSM 10642 NC_013849 252831..249216



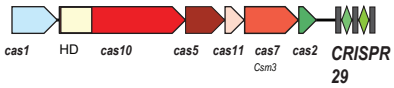
Candidatus Verstraetearchaeota archaeon QMRB01000063 4094..56



Deltaproteobacteria bacterium QMLH01000866 1..2230

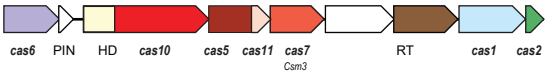


Candidatus Lokiarchaeota archaeon QMYW01000240 5719..1154

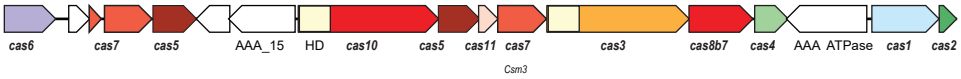


Thermoprotei archaeon QMSH01000092 4208..493

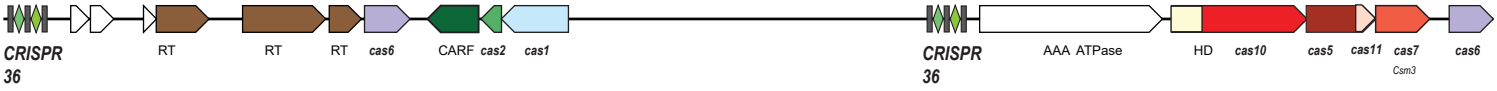
Deltaproteobacteria bacterium CG06_land_8_20_14_3_00_44_19 PEVP01000099 8676..576



Candidatus Pacearchaeota archaeon QMUZ01000004 16459..2173



Methanosarcina lacustris Z-7289 NZ_CP009515 2718430..2697091



Supplementary Figure 4

Selected gene neighbourhoods of subtype IV-C CRISPR-Cas systems

Designations and abbreviations are the same as in t Supplementary Figure 1. See also the system description in the text.

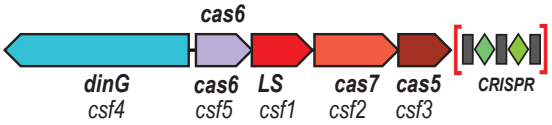
A. Two representatives of the known type IV subtypes.

B. Selected gene neighborhoods of subtype IV-C CRISPR-Cas system. A subtype I-A system that is inserted into the subtype IV-C locus of *Thermococcus sp* is shaded.

Additional abbreviations: PD-DExK, nuclease of the PD-DExK family (also known as RecB or restriction endonuclease family).

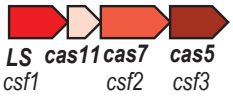
A

Type IV-A



Thioalkalivibrio_K90mix
NC_013930 61135..66712

Type IV-B

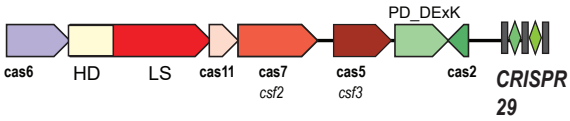


Rhodococcus jostii RHA1
NC_008270 70536..73274

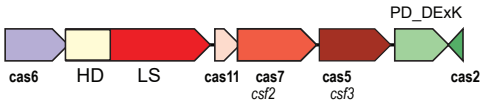
B

Type IV-C

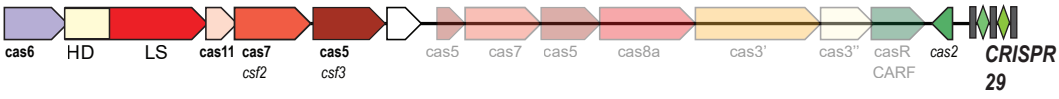
Thermococcus onnurineus NA1 NC_011529 302053..296280



Thermococcus pacificus NZ_CP015102 830663..836385

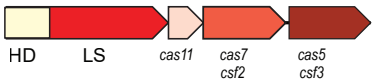


Thermococcus sp. 2319x1 NZ_CP012200 1116628..1128479

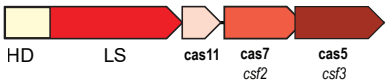


Pyrodictium delaneyi NZ_NCQP01000002 29461..33678

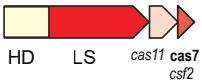
Thermoflexia bacterium RFKE01000401 4661..105



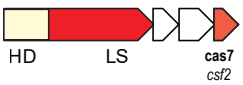
Oscillochloris trichoides DG-6 NZ_GL501404 258926..254179



Thermoprotei archaeon QMSI01000139 645..3017



Thermoplasmata archaeon QMSU01000223 275..3207



Supplementary Figure 5 – Derived Class 1 systems lacking CRIPSR arrays.

Designations and abbreviations are the same as in the Supplementary Figure 1.

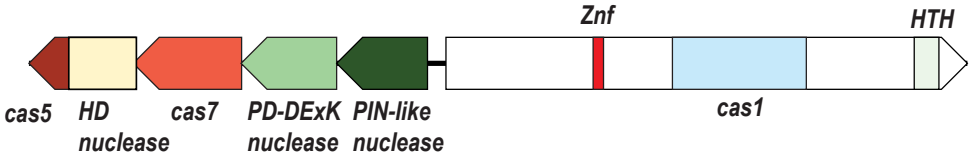
A. HRAMP (Halobacterial RAMP) system representative. The HRAMP system has been described in detail[22]. XXX, a signature protein of the HRAMP system that it not similar to any other protein family. DEDDy, a nuclease of the DnaQ family of 3'-5' exonucleases; HNH, DNase of the HNH family.

B. Derived CRISPR-Cas systems from metagenomic sequences of Asgard archaea.

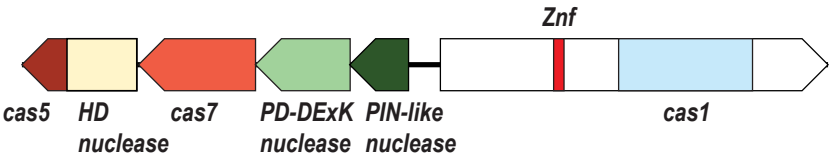
A

Halogeometricum borinquense DSM 11551 PR3
NC_014729.1 1241331..1246082

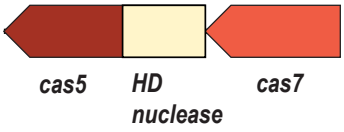
B



Candidatus Heimdallarchaeota archaeon B3_Heim
NJBF01000020 7671..453



Candidatus Heimdallarchaeota archaeon B3_Heim
NJBF01000013 72134..78334

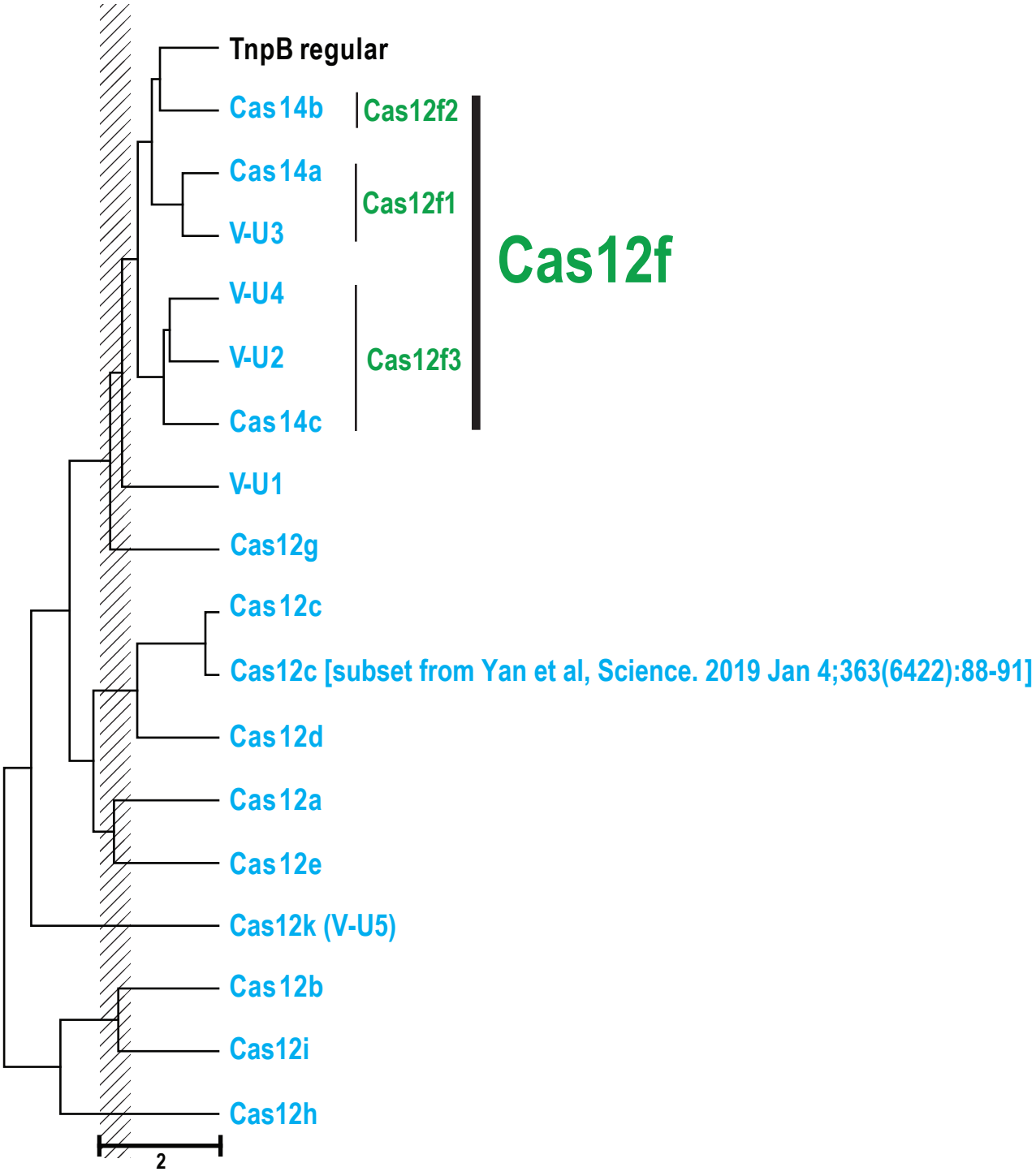


Candidatus Lokiarchaeota archaeon,
SDMZ01000003 1947..4534

Supplementary Figure 6

Deep relationships between Type V effector families.

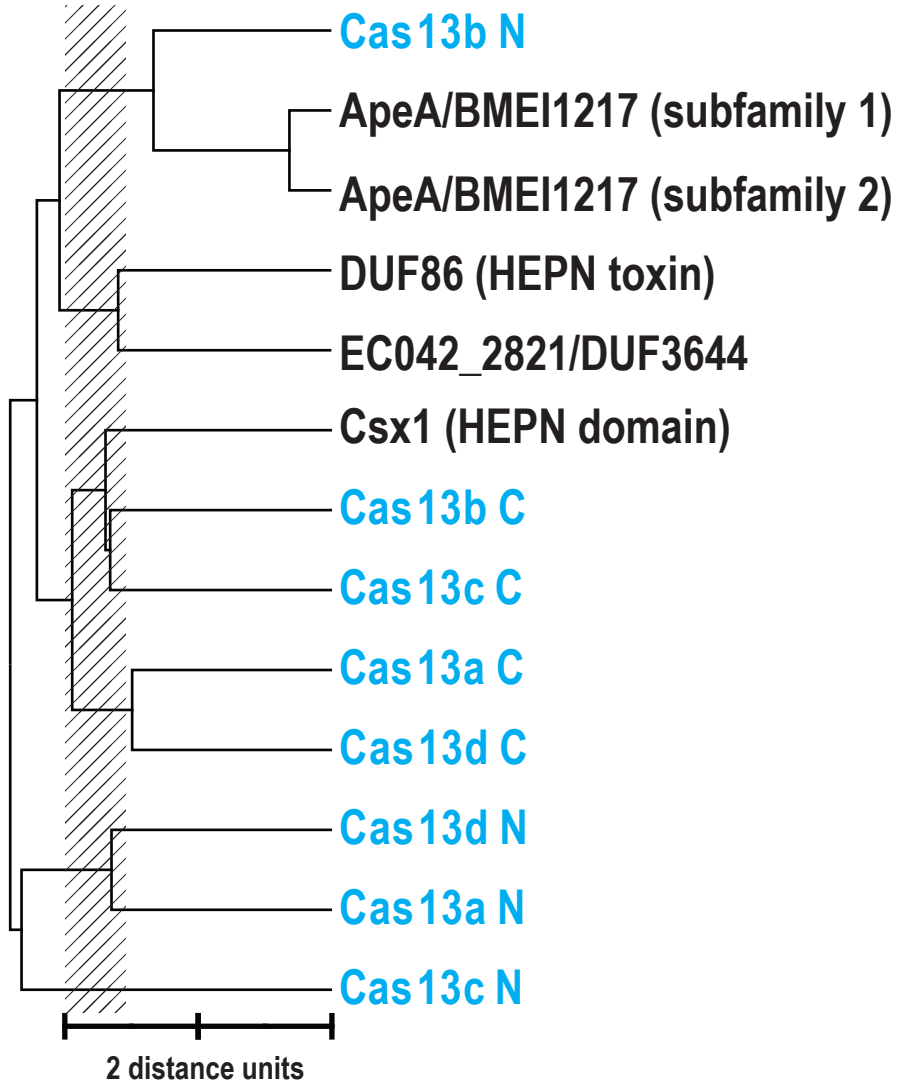
Profile-profile comparisons were performed and the UPGMA dendrogram was constructed as described in Supplementary Methods. Multiple alignments used for this analysis are available in Supplementary Dataset 2 (see prefix "Type_V"). The dashed rectangle corresponds to the tree depth D between 1.5 and 2 ($D = 2$ roughly corresponds to the pairwise HHsearch similarity score of $exp(-2D) \approx 0.02$ relative to the self-score). This region typically reflects sequence relationships that could be reproduced using other methods, such as PSI-BLAST. Generally, if profiles are grouped between 1 and 1.5 distance units they are assigned to one subtype. The proposed renaming scheme for smaller type V effector is shown in green. These relationships were further confirmed using phylogenetic analysis in which the respective type V effectors were combined with TnpB sequences (Supplementary Dataset 4_trees).


TnpB regular
Cas14b
Cas14a
V-U3
V-U4
V-U2
Cas14c
Cas12f2
Cas12f1
Cas12f3
Cas12f
V-U1
Cas12g
Cas12c
Cas12c [subset from Yan et al, Science. 2019 Jan 4;363(6422):88-91]
Cas12d
Cas12a
Cas12e
Cas12k (V-U5)
Cas12b
Cas12i
Cas12h
2

Supplementary Figure 7

Deep relationships between Type VI effector families and several most similar HEPN domain families.

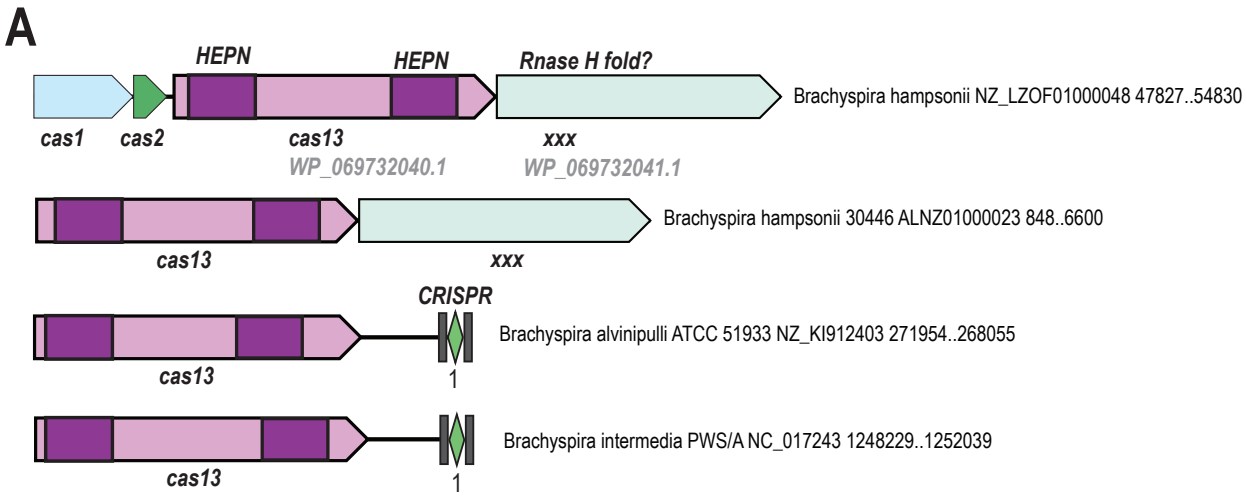
To identify the families of HEPN domain-containing proteins with the highest similarity to Cas13, a PSI-BLAST search against the NCBI NR database was performed with the N-terminal domain of Cas13b (WP_004919755.1, *Riemerella anatipestifer*, 1-195 aa) as the query, with the following parameters: no low complexity filtering, E-value = 0.01, and eukaryotic sequences excluded. After 3 iterations, 5 distinct best scoring HEPN proteins were manually selected. Each of these proteins was used as a query for a PSI-BLAST search, and 8 to 10 diverse sequences from the 100 best hits were selected and aligned using MUSCLE. These families were designated according to the previously published description of the HEPN domain families [23]. These alignments (profiles) were combined with profiles for N- and C-terminal domains of known Cas13 families. Profile-profile comparisons and the UPGMA dendrogram were performed as described in Supplementary Methods. Multiple alignments used for this analysis are available in Supplementary Dataset 2 (see prefix "Type_VI"). The dashed rectangle corresponds to the tree depth between 1.5 and 2 distance units (see also **Supplementary figure 6).**


Cas13b N
ApeA/BMEI1217 (subfamily 1)
ApeA/BMEI1217 (subfamily 2)
DUF86 (HEPN toxin)
EC042_2821/DUF3644
Csx1 (HEPN domain)
Cas13b C
Cas13c C
Cas13a C
Cas13d C
Cas13d N
Cas13a N
Cas13c N
2 distance units

Supplementary Figure 8.

Identification and characterization of a distinct variant of type VI in several *Brachyspira* species.

A. **Selected gene neighbourhoods of the type VI variant from *Brachyspira* species.** Designations are the same as in the Supplementary Figure 1. XXX, an uncharacterized protein with weak but reliable sequence similarity to DUF3800 (see panel B).
B. **HHpred output for protein XXX encoded in some of the *Brachyspira* type VI loci.** The DUF3800 family belongs to RNase H fold with the highest similarity to RNase H (eg. HHpred search initiated with a DUF3800 representative WP_041466413.1 identifies profile TIGR00716, ribonuclease HIII, with probability 92%). The predicted catalytic aspartate is shown by the red arrow.
C. **Alignment of CRISPR miniarrays identified in two *Brachyspira* species.** Identical nucleotide positions are shown in blue for repeats and orange for spacer regions.
D. **UPGMA dendrogram for the N- and C-HEPN domains of the Cas13 proteins from the *Brachyspira* type VI variant and other type VI systems.** The tree was built as described for Supplementary Figure 7. Two additional Brachyspira alignments are available in Supplementary Dataset 2 (see prefix "Brachyspira _Type_VI").
E. **Multiple alignment of N and C- terminal HEPN motifs of the Cas13 proteins from the *Brachyspira* type VI variant.** The catalytic motifs of the HEPN domains are highlighted in red.



B

PF12686.7 ; DUF3800 ; Protein of unknown function (DUF3800)

Probabilitv: 93.9 E-value: 0.73 Score: 42.74 Aligned Cols: 115 Identities: 14% Similarity: 0.086

```
Q ss_pred          EEeCCCCCccCCe---eEEeeeEEeccCCcchhHHHHHHHHHHHhcCCccccceeeEehhHccCchhh
Q Q_8757787_7  195 IYIDESGKIINNE---EFLVGGYFIEDINFRCWNEKARRLFYNILGKYSNNISNKKIIYHRNEIKNKKLK  261 (918)
Q Consensus    195 iYiDEsGnI~~n~~~~F~vGGYf~kdi~~~dW~~~~~a~l~~l~~kY~~n~~~~~iiyHR~eI~n~~~k  261 (918)
                   +||||||....+.   .|++||..+.+..   +....-..+....+++          ++..+++-.+..
T Consensus      2 ~yiDESG~~~~~~~~~~v~a~v~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~e~~~~~~~  58 (174)
T PF12686.7      2 AFIDESGTVGIPGGTHFLVVALLSTGQPR---EIELPIRHAMKKYGPS----------LSRGEIKAANFE  58 (174)
T ss_pred          eEEeCCCCCCCCCCeEEEEEEEeCChH---HhHHHHHHHHHHHCcc----------cCCCCccccccC
```

potential first catalytic aspartate of RNase H fold

```
Q ss_pred          ccccHHHHHHHHhcCCeEEEEEecCCCCceechhHHHHHHHHHHHHHHHHHHHHhcCccccCCeeEEEE
Q Q_8757787_7  262 KNITYDVFNFVKINNGKFISICENKVKNNLFSRRYYISLLSELIMYTVRTIIVENKINLIDDKLLLII  329 (918)
Q Consensus    262 K~ITkdiFnFik~N~gkfvcIyEn~pngieFtr~YY~dL~S~LIv~Tl~kIi~e~~idl~~~kiLLII  329 (918)
                  ...-..+++.+.+.+..++++.-++ ......+...-.....++..+|..+..       ....+.+++
T Consensus     59 ~~~~~~~~~~i~~~~~~~~~~i~k~~~~~~~~~~~~~y~~~~~~ll~~~~~~~~~~~~i~~  119 (174)
T PF12686.7     59 EKAITRLLTEIAKENVSIFATIVDQ-SVIVKPPAEMEEIYRRAVSKTVYKLVE------QCPRVNICL  119 (174)
T ss_pred          HHHHHHHHHHHHhcCCcEEEEEEec-cccCCCCcchhHHHHHHHHHHHHHHHh------hCCCEEEEE
```

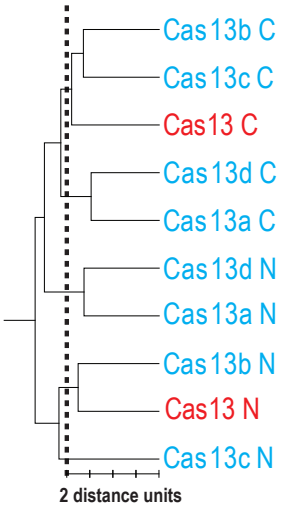
C

B.intermedia vs B.alvinipulli

```
-TTAAAGAGCAGTTTTAATAAACTGTTTCTGGATATACAGGAAAGCGGGATTA--TACGACATTCCATCGGTGCTTTGTGTAGTCTCGATACCCTGTTG--CTTATAGCTTAAAGAGCAGTTTTAATAAACTGTTTCTGGATATACAGGAAAGCGGGATTA

CTTAAAGAGTGGTTTTAATAAACTGTTTCTGGATATACAGGAAAGCGGGATTT--CCACCAGCTACTGCAATAGTTTTTTAGGTCTCGATACCCTGTTG--CTTATGGCTTAAAGAGAGGTTTTAATAAACTGTTTCTGGATATACAGGAAAGCGGGATTT
```

D

UPGMA dendrogram of profile-profile similarity scores



E

N-terminal HEPN domain

```
WP_069732040.1 VIDASYKQAIDFIMDFRNILIHKNVRDDH
CCY77332.1     KIEKSFEDAMIFLRDFRNIMIHKNIKEDT
WP_028329891.1 KINNAFTEAAVFLRDFRNIMMHKNVKEDT
WP_107926119.1 KIDSAFNQAAKFLHSFKNIMMHKNIKDKT
```

C-terminal HEPN domain

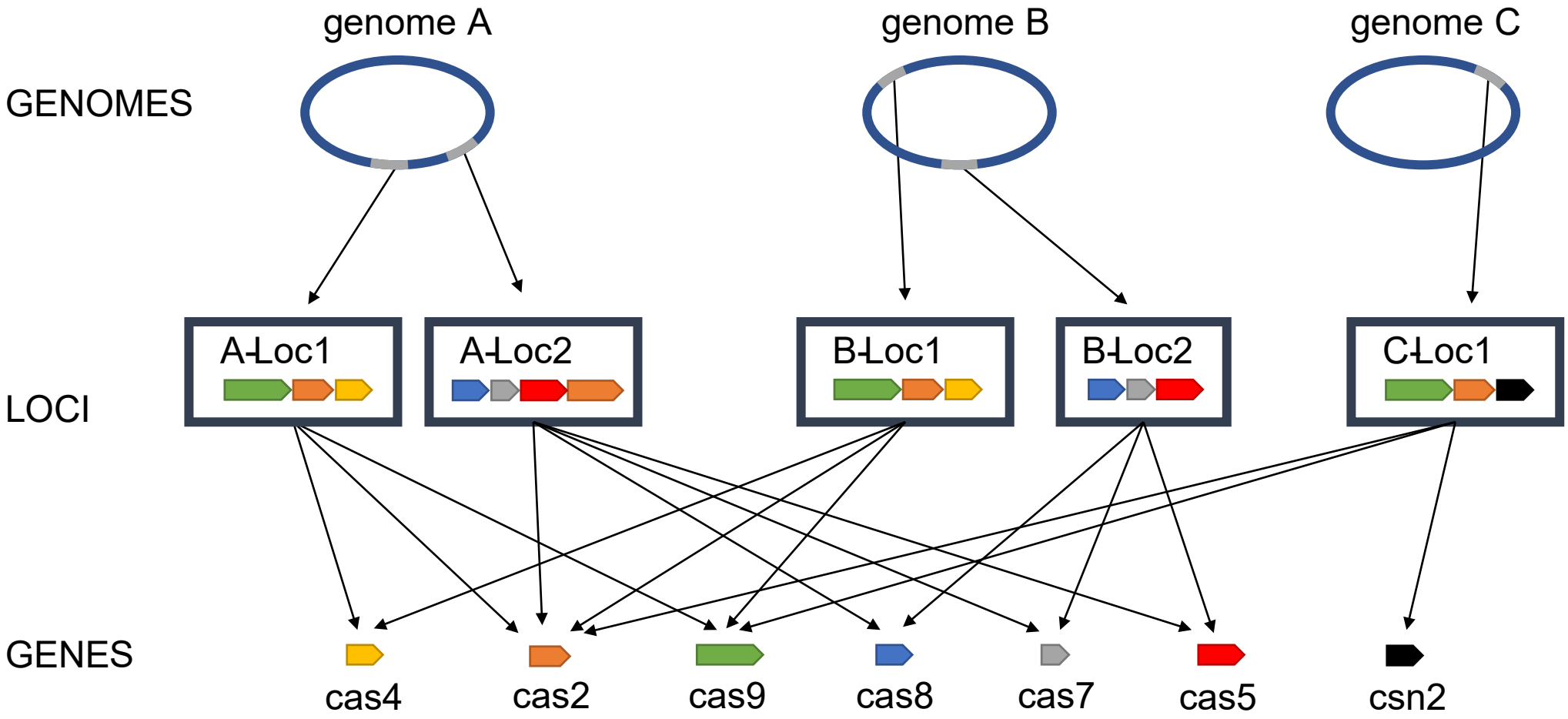
```
WP_069732040.1 NNDKMTPKISETIKDVRNMICHMRVKD--
CCY77332.1     YNKSTDKKLEGKLSVVKNILAHGRLSD--
ELV06347.1     NNDKMTPKISETIKDVRNMICHMRVKDGT
WP_028329891.1 NNSAFDDSETEKIINVRNMLAHGRLFDES
WP_107926119.1 ANYKYNKSLPNEIKTCRIMLSHGRLFDSL
```

Supplementary Figure 9

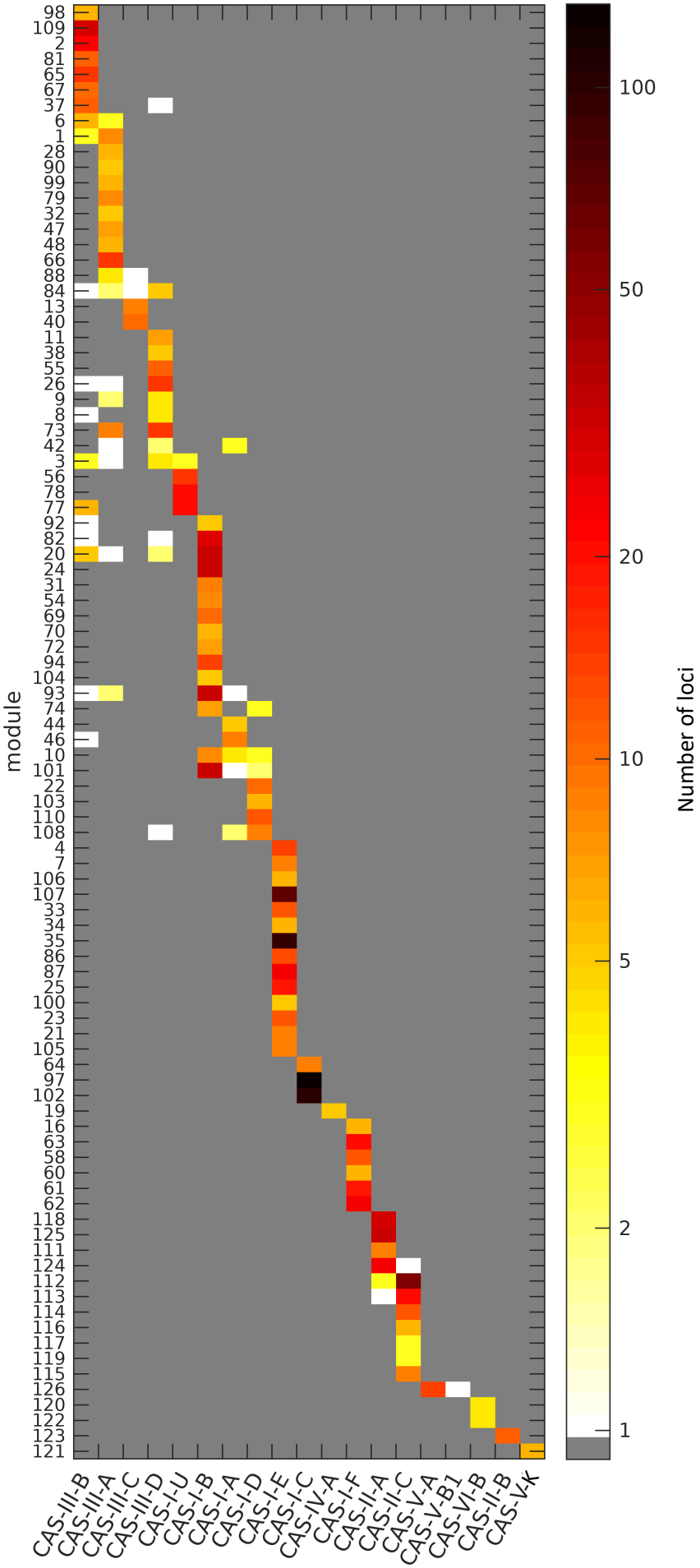
Bipartite network analysis of CRISPR–Cas systems.

A. Bipartite network architecture: genomes, CRISPR–*cas* loci and *cas* genes. The network contains two categories of nodes, CRISPR–*cas* genomic loci and *cas* genes that are connected by an edge when a given locus includes a particular gene. This approach yields modules that combine loci and genes of which they are comprised. The figure shows a specific example, with five CRISPR–Cas loci that can be split into two modules, one characterized by the presence of *cas9*, i.e type II (A-Loc1, B-Loc1, and C-Loc1), and the other by the combination of *cas8*, *cas7*, and *cas5*, i.e. type I (A-Loc2 and B-Loc2). For clarity, only some representative *cas* genes are included.
B. Modules identified in the bipartite networks of CRISPR–*cas* loci and *cas* genes. The vertical axis shows module numbers, and the horizontal axis shows CRISPR–Cas subtypes. The modules are color-coded according to the number of CRISPR–*cas* loci they include (see the color gradient to the right of the maps).

A


genome A
genome B
genome C
GENOMES
LOCI
A-Loc1
A-Loc2
B-Loc1
B-Loc2
C-Loc1
GENES
cas4
cas2
cas9
cas8
cas7
cas5
csn2

B


module
98
109
2
81
65
67
37
6
1
28
90
99
79
32
47
48
66
88
84
13
40
11
38
55
26
9
8
73
42
3
56
78
77
92
82
20
24
31
54
69
70
72
94
104
93
74
44
46
10
101
22
103
110
108
4
7
106
107
33
34
35
86
87
25
100
23
21
105
64
97
102
19
16
63
58
60
61
62
118
125
111
124
112
113
114
116
117
119
115
126
120
122
123
121
CAS-III-B
CAS-III-A
CAS-III-C
CAS-III-D
CAS-I-U
CAS-I-B
CAS-I-A
CAS-I-D
CAS-I-E
CAS-I-C
CAS-IV-A
CAS-I-F
CAS-II-A
CAS-II-C
CAS-V-A
CAS-V-B1
CAS-VI-B
CAS-II-B
CAS-V-K
100
50
20
10
5
2
1
Number of loci

Supplementary Tables

Supplementary Table 1. The core proteins of CRISPR-Cas systems

Family	Biochemical evidence/*in silico* prediction	Examples of available structures and structural features
Cas1	Metal-dependent deoxyribonuclease [24,25] that functions as the integrase during adaptation [26]; deletion of Cas1 in *E. coli* results in increased sensitivity to DNA damage and impaired chromosomal segregation [27]. Typically, forms complex with Cas2, but at least one subfamily does not require Cas2 for spacer integration [28].	PDB: 3GOD, 3LFX, 2YZS Unique fold with two domains: N-terminal β stranded domain and catalytic C-terminal α-helical domain.
Cas2	RNase specific to U-rich regions [29], double-stranded DNAse; forms a tight complex with Cas1 and appears to perform a structural role during adaptation. Cas2 proteins are predicted to be active nucleases in many CRISPR-Cas systems but appear to be inactivated in others. The role of the nuclease activity of Cas2 in CRISPR-Cas function (if any) remains unclear.	PDB: 2IVY, 2I8E, 3EXC, 4P6I RRM (ferredoxin) fold.
Cas3 (helicase and HD domain)	Single-stranded DNA nuclease (HD domain) and helicase [30]; required for interference [31].	PDB: 4QQW, 4QQX, 4QQZ, 4QQY
Cas3'' (stand alone HD nuclease)	Metal-dependent deoxyribonuclease specific for double-stranded oligonucleotides [32].	PDB: 3S4L, 3SKD
Cas4	PD-(DE)xK superfamily nuclease with four conserved cysteines coordinating one [4Fe-4S] or [2Fe-2S] cluster [33-35]; cleaves ssDNA in the 5' to 3' or both directions [34-36]. A component of the adaptation complexes in many subtypes, assisting in precise protospacer processing and PAM selection [37,38].	PDB: 4IC1

Cas5	Subunit of Cascade complex interacting with large subunit and Cas7 subunit and binding the 5'-handle of crRNA [31,39-44]. In subtype I-C, Cas5 is the ribonuclease that replaces the Cas6 function [45].	PDB: 3KG4; 3VZI; 3VZH Two domains of RRM (ferredoxin) fold, the C-terminal domain is deteriorated in many Cas5 proteins of Type I systems.
Cas6	Metal-independent endoribonuclease that generates crRNAs [31,39,46-50].	PDB: 2XLJ, 1WJ9,3I4H, 4C8Z, 4DZD Two domains of RRM (ferredoxin) fold, RAMP superfamily.
Cas7	Subunit of Cascade complexes binding crRNA [31,41-44]; often present in Cascade complexes in several copies.	PDB: 3PS0, 4N0L and many others. RRM (ferredoxin) fold with subdomains, RAMP superfamily.
Cas8abcefu, (large subunit)	Subunit of Class 1 Cascade complex, involved in PAM recognition [31,44,51-53].	PDB: 4AN8
Cas9	Type II effector protein. In Type II CRISPR-Cas systems, Cas9 is essential for pre-crRNA processing and to cleave the target DNA [54,55], although it requires help of the house-keeping RNase III and a dedicate trans-activating (tracr) RNA encoded in the respective CRISPR-*cas* locus [56]. Both the RuvC and HNH nuclease domains of Cas9 are involved in the cleavage of the target DNA [57,58]. Additionally, Cas9 contributes to adaptation, in particular, by recognizing the PAM motif [59,60].	PDB: 4OGC, 4OO8, 4CMP Cas9 contains several subdomains, including RuvC and HNH nuclease domains and adopt a bi-lobed general structure [61,62].
Cas10 (large subunit)	Subunit of Cascade (Cmr and Csm) complex [41-43,47].	PDB: 3UNG, 4DOZ Two domains homologous to Palm domain polymerases and cyclases, both belong to RRM (ferredoxin) fold; Zn finger containing domain and C-terminal alpha helical domain [63]; Fusion: HD nuclease domain.
Cas11 **(small subunit)**	Small, mostly alpha helical protein, subunit of Class 1 Cascade complexes [31,39,41-43,47,50,64].	PDB: 2ZCA (Cse2); 2ZOP, 2OEB (Cmr5); 3ZC4 (Csa5);

		Cse2 has two alpha helical bundle-like domains; Cmr5 has a domain matching N-terminal domain of Cse1 and Csa5 has a domain matching C-terminal domain of Cse2.
Cas12	Type V effector protein. In Type V CRISPR-Cas systems, the respective Cas12 variants show highly diverse properties [9,65-69]. Most of the Cas 12 proteins contain an active RuvC-like nuclease domain that is typically responsible for the cleavage of both strands of the target DNA. Several Cas12 also contain a subdomain involved in processing of pre-crRNA, but other type V loci, similarly to type II, encode a tracrRNA that is involved in the processing of pre-crRNA along with the house-keeping bacterial RNase III.	PDB: 5NFV and many others (Cas12a); 5WQE and others (Cas12b); 6NY1 (Cas12e); Cas12 adopts a bilobed shape and contains several subdomains, including RuvC nuclease domain, and often, an OB-fold domain involved in pre-RNA cleavage [70-72].
Cas13	Type VI effectors. All Cas13 proteins contain two HEPN superfamily RNase domains. Accordingly, type VI systems target exclusively RNA [73-76]. Cas13a and Cas13d are active in processing of pre-crRNA [69,75,77,78]. Once activated by the RNA target recognition, Cas13 becomes a non-specific RNase that appears to be toxic for the cell, inducing dormancy [73,75,79].	PDB: 5XWP, 5W1I (Cas13a); 6DTD, 6AAY (Cas13b); 6E9E (Cas13d)[80,81] [76,82,83] Cas13 proteins adopt a bilobed shape and contain two HEPN superfamily RNase domains.

Supplementary Table 2.

The classes, types and subtypes of CRISPR-Cas systems, their signature proteins and key features

subtype	Mono-phyletic in Cas1 tree	Signature proteins: **Strong**/weak* (other name)	Comment
Class 1: multisubunit effector complexes			
Type I: Cascade effector complexes			
I-A	No	Cas8a, Csa5 (small subunit)	Cas3 is often split into the helicase Cas3' and HD nuclease Cas3'' and a separate gene for small subunit *csa5* is often present. Several distinct subfamilies of Cas8a exist.
I-B	No	Cas8b	I-B systems belong to several distinct clades on the Cas1 tree. Characterized only by gene composition: all loci have *cas5*, *cas7*, *cas8* and *cas6* genes. Usually the *cas3* gene is not split. Several distinct subfamilies of Cas8b exist.
I-C	No	Cas8c	These systems usually do not have a *cas6* gene. Cas5 is catalytically active and replaces Cas6 function.
I-D	No	**Cas10d (large subunit)**	The HD domain is associated with the large subunit rather than with Cas3 but lacks the circular permutation of the motifs like the HD domain fused with Cas10 in type III systems.
I-E	Yes	**Cse1 (Cas8e), Cse2 (small subunit)**	The *cas4* gene is not associated with this system.
I-F1	Yes	**Cas8f (Csy1), Cas5f1 (Csy2), Cas7f1 (Csy3), Cas6f**	The *cas4* gene is not associated with this system, *cas2* is fused to *cas3*. There is no separate gene for a small subunit, which is either missing or fused to the large subunit.
I-F3	N/A	**Cas8f/Cas5f** **(Csy1/Csy2) fusion**	The *cas1-cas2-cas3* genes are not present. Usually three genes (*csy1/csy2* fusion, *csy3* and *cas6f*) are present in an operon, Effector complex interacts with the transposase subunits (TniQ/TnsD family proteins) and guides the insertion of the Tn7-like transposon to the vicinity of the protospacer.
I-F2	Yes	**Cas5f2 (PBPRB1993) Cas7f2 (PBPRB1992)**	A derived variant of I-F with two distinct genes of group Cas5 (PBPRB1992) and group Cas7 (PBPRB1993) RAMPs. Large subunit is missing.
I-G	No	**GSU0054** (Cas5 group RAMP)	These systems usually do not have identifiable *cas6*. Cas5 has several specific insertions or fusions, but is likely to be catalytically active. There are systems with different subfamilies of the large subunit, which are often severely

			deteriorated and sometimes even missing. Cas3 contains a C-terminal HD domain.
Type III: Csm (III-A,D,F,E) and Cmr (III-B,C) effector complexes			
III-A	No	Csm2 (small subunit)	Also known as the Csm module and Cas10 usually has active catalytic motifs, which are involved in synthesis of a signaling molecule, the cyclic oligoadenylate. The III-A loci typically contain several *cas7* group genes and is often linked to *csm6* which has CARF and C-terminal HEPN domain. Might be associated with *cas1-cas2* gene pairs of different origin.
III-B	No	Cmr5 (small subunit)	Also known as the Cmr (or RAMP) module. Cas10 often has active catalytic motifs, which are involved in synthesis of a signaling molecule, the cyclic oligoadenylate. These systems are usually associated with several Cas7 group RAMPs and are rarely present in a genome as a stand-alone system They are usually not linked to *cas1-cas2* gene pair. Cmr1 has a duplication of RAMP domains both from the Cas7 group.
III-C	No	Cas10 (MTH326 or Csx11)	Have a small subunit and several Cas5 and Cas7 RAMP protein shared with Type III-B. The large subunit is often inactivated and some Cmr1 family proteins possess only one RAMP domain.
III-D	No	Csx10 (Cas5 group RAMP) Csx19 (uncharacterized component)	Have small subunit and several Cas5 and Cas7 RAMP protein shared with Type III-A. The signature gene is *csx10* which is related to Cas5 group RAMPs. Another specific gene csx19(*all1473)* is likely to be a component of effector complex but is not similar to any known Cas proteins. The large subunit is often lacking the HD domain, but has active catalytic motifs, which are involved in synthesis of a signaling molecule, the cyclic oligoadenylate. Csx10 could be fused to the small subunit in some systems and Cas7 group RAMPs are often fused and have large insertions.
III-E	N/A	**gRAMP (SCABRO_02597)**	The gRAMP protein contains three Cas7 group RAMP superfamily domain and a putative small subunit domain. Typically, it is associated with CHAT (caspase) domain containing protein.
III-F	Yes	Small subunit (TLET_RS00525), Large subunit (TLET_RS00515)	The large subunit, Cas7 and Cas5 group RAMP proteins of the system shows remote but significant similarity with type III system components, whereas the putative small subunit does not. III-F systems contain only one *cas7*-like gene. Cas10-like subunit has inactivated cyclase/polymerase domain.
Type IV: Csf effector complexes and IV-C effector complex			
IV-A	No	Large subunit (Csf1)	These systems possess Cas5, Cas7 and a gene for a very reduced large subunit which typically contains Zn finger domain. Occasionally associated with adaptation module, Cas6 and CRISPR array. DinG-like helicase is an ancillary gene that

			is often present in the loci. Mostly found on plasmids or other integrated elements.
IV-B	N/A	Small subunit (RHA1_ro10070)	Large subunit, Cas5 and Cas7 are homologous to respective components of subtype IV-A. Additionally a gene for putative small subunit is present in the same loci (eg. RHA1_ro10070). The system is linked to CysH. Most systems are encoded on plasmids or prophages.
IV-C	N/A	**LS (D6793_05715)** SS (D6793_05710)	Csf2-like protein (Cas7 group RAMP) and Csf3 (Cas5 group RAMP) resemble respective homologs from subtypes IV-A and IV-B. Putative large and small subunit lack any similarity with respective proteins from any other CRISPR-Cas systems. The putative large subunit contains HD nuclease domain with the same order of conserved motifs as in HD domain of Cas3 .
Class 2: Single protein (multidomain) effector complexes			
Type II: Cas9 effector protein			
II-A	Yes	Csn2	Monophyletic group on Cas9 and Cas1 tree. There are four genes in these operons with *csn2* gene in addition to *cas1_2_9*. Typically, Csn2 is inactivated ATPase involved in spacer acquisition. There are at least 7 distinct families of Csn2.
II-B	Yes	Cas9 (Csx12 subfamily)	Monophyletic group on Cas9 tree with four gene operons containing *cas4* in addition to *cas1_2_9*.
II-C	No	N/A	Only three genes are present in the II-C operon - *cas1_2_9*.
Type V: Cas12 effector protein			
V-A	Yes	Cas12a (Cpf1)	A founding member of type V. Cas12a is a large protein which C-terminal region shares a significant similarity with TnpB the ORF encoded by transposable element IS605. Contains RuvC-like nuclease which is sufficient to cleave both DNA strands. Does not require tracrRNA; contains catalytic subdomain for pre-crRNA cleavage.
V-B	Yes	Cas12b (C2c1)	Cas12b2 is much smaller than Cas12b1 (originally characterized Cas12b).
V-C	Yes (C and D clade)	Cas12c (C2c3)	Cas12c has been shown to demonstrate strong interference activity in E. coli indicative of dsDNA targeting. The minimal active system consists of a RuvC-containing effector, crRNA, and tracrRNA showing only a short anti-repeat complementary to the crRNA direct repeat. The system is linked to with an adaptation module containing only Cas1.
V-D	Yes (C and D clade)	Cas12d (CasY)	Distant homolog of Cas12c. The system is linked to with an adaptation module containing only Cas1.

V-E	Yes	Cas12e (CasX)	A distinct type V effector with limited structural similarity with Cas12a. Contains RuvC-like nuclease domain in the C-terminal region, which is sufficient for dsDNA cleavage. Requires tracrRNA for activity. Adaptation module contains cas4 in addition to cas1_2.
V-F	N/A	Cas12f (c2c4,c2c8,c2c9,c2c10 ,Cas14a,Cas14b,Cas14 c)	Cas12f proteins contain RuvC-like nuclease domain in the C-terminal region, which displays high similarity with TnpB. Each subfamily forms a distinct branch in the TnpB tree, but the branches are paraphyletic. N-terminal region is unique for each variant. The length of effectors varies from 600 to 800 aa. Cas12f1 shows single-stranded DNA targeting resulting in collateral ssDNA cleavage and require tracrRNA for activity. Some of V-F loci contain adaptation module consisting of Cas1_2_4 genes
V-G	N/A	Cas12g	Cas12g proteins contain RuvC-like nuclease domain in the C-terminal region. Cas12g is differentiated from other characterized type V systems by RNA-targeting that triggers collateral cleavage of both ssRNA and ssDNA. This mechanism of interference is abolished by mutation of RuvC catalytic residues. The minimal Cas12g system consists of the compact (~770aa) RuvC-containing effector, crRNA, and a tracrRNA.
V-H	N/A	Cas12h	Cas12h proteins contain RuvC-like nuclease domain in the C-terminal region. Compact (~870aa) has been shown to demonstrate strong interference activity in *E. coli* indicative of dsDNA targeting. Does not require tracrRNA. The type V-H system does not appear to contain an adaptation module.
V-I	N/A	Cas12i	Cas12i proteins contain RuvC-like nuclease domain in the C-terminal region. Showing distant similarity to Cas12b, the Cas12i effectors are differentiated in-part by significant N-terminal truncations. The minimal Cas12i system, consisting of the effector and no tracrRNA, is capable of pre-crRNA maturation, dsDNA cleavage with predominant nicking, and ssDNA collateral cleavage.
V-K	N/A	**Cas12k**	Cas12k is a component of CRISPR-associated transposase (CAST) interacting with Tn7-like transposase subunits TnsB, TnsC and TniQ. CAST catalyzes RNA-guided DNA transposition by unidirectionally inserting segments of DNA 60-66 bp downstream of the protospacer. Typically Cas12k and cognate CRISPR array are encoded at the left end of the Tn7-like transposon.
Type VI: Cas13 effector protein			
VI-A	no	Cas13a	Cas13a is the original CRISPR effector characterized as having RNA target-activated collateral RNA cleavage. The active Cas13a complex consists of the dual HEPN domain-containing

			effector, and processes its own pre-crRNA. The system contains no tracrRNA and preferentially target sites with a PFS.
VI-B1	N/A	Cas13b1, Csx28	The type VI-B system does not appear to contain an adaptation module. The type VI-B CRISPR systems consist of the dual HEPN domain-containing Cas13b effector. Cas13b1 proteins form a distinct branch in the Cas13b tree and specifically associated with Csx28 accessory protein, which have four predicted transmembrane domains. Csx27 represses Cas13b RNA nuclease activity.
VI-B2	N/A	Cas13b2, Csx27	The Cas13b2 effectors has the same features as Cas13b. Cas13b2 proteins form a distinct branch in the Cas13b tree and specifically associated with Csx28 accessory proteins, wich has a predicted transmembrane domain. Csx28 enhances Cas13b RNA nuclease activity.
VI-C	N/A	Cas13c	The type VI-C system does not appear to contain an adaptation module. Cas13c effector family has two HEPN domains, like other Cas13 effector, but otherwise it is not similar to other Cas13 proteins.
VI-D	Yes	Cas13d	The smallest of the type VI RNA-targeting CRISPR effectors, Cas13d (~928aa) shows strong RNA target-activated collateral RNA cleavage with no requirement for a protospacer-flanking motif (PFS). Cas13d is co-occurs with polyphyletic WYL domain-containing accessory protein families. It has been shown that Cas13d RNA cleavage is positively regulated by WYL1. Type VI-D contains only limited co-occurrence with Cas1/2 adaptation modules.

Note: * **Strong**/weak – is the characteristic of the signature protein family with respect to subtype recognition/classification ability using the respective profile. Strong means that it has a relatively high specificity and high selectivity, i.e. is a reliable signature, whereas weak means that search for this family yields either a high fraction of false positives or false negatives, but nevertheless, the family remains the best available signature for a particular subtype.

References

1 Coordinators, N. R. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* **44**, D7-19, doi:10.1093/nar/gkv1290 (2016).

2 Steinegger, M. & Soding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat Biotechnol* **35**, 1026-1028, doi:10.1038/nbt.3988 (2017).

3 Makarova, K. S., Wolf, Y. I. & Koonin, E. V. Archaeal Clusters of Orthologous Genes (arCOGs): An Update and Application for Analysis of Shared Features between Thermococcales, Methanococcales, and Methanobacteriales. *Life (Basel)* **5**, 818-840, doi:10.3390/life5010818 life5010818 [pii] (2015).

4 Bland, C. *et al.* CRISPR recognition tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. *BMC Bioinformatics* **8**, 209, doi:1471-2105-8-209 [pii] 10.1186/1471-2105-8-209 (2007).

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

6 Marchler-Bauer, A. *et al.* CDD: conserved domains and protein three-dimensional structure. *Nucleic Acids Res* **41**, D348-352, doi:10.1093/nar/gks1243 gks1243 [pii] (2013).

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

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

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

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

11 Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**, 1792-1797 (2004).

12 Soding, J. Protein homology detection by HMM-HMM comparison. *Bioinformatics* **21**, 951-960, doi:bti125 [pii] 10.1093/bioinformatics/bti125 (2005).

13 Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2--approximately maximum-likelihood trees for large alignments. *PLoS One* **5**, e9490, doi:10.1371/journal.pone.0009490 (2010).

14 Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* **30**, 772-780, doi:10.1093/molbev/mst010 mst010 [pii] (2013).

15 Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* **46**, D8-D13, doi:10.1093/nar/gkx1095 4621330 [pii] (2018).

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

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

18 Marquitti, F. M. D., Giuimaraes, P. R., Pires, M. M. & Bittencourt, L. F. MODULAR: software for the autonomous computation of modularity in large network sets. *Ecography* **37**, 221-224 (2014).
19 Barber, M. J. Modularity and community detection in bipartite networks. *Phys Rev E Stat Nonlin Soft Matter Phys* **76**, 066102, doi:10.1103/PhysRevE.76.066102 (2007).
20 Lancichinetti, A. & Fortunato, S. Consensus clustering in complex networks. *Sci Rep* **2**, 336, doi:10.1038/srep00336 (2012).
21 Lancichinetti, A., Radicchi, F., Ramasco, J. J. & Fortunato, S. Finding statistically significant communities in networks. *PLoS One* **6**, e18961, doi:10.1371/journal.pone.0018961 (2011).
22 Makarova, K. S. *et al.* Predicted highly derived class 1 CRISPR-Cas system in Haloarchaea containing diverged Cas5 and Cas7 homologs but no CRISPR array. *FEMS Microbiol Lett* **366**, doi:10.1093/femsle/fnz079 (2019).
23 Anantharaman, V., Makarova, K. S., Burroughs, A. M., Koonin, E. V. & Aravind, L. Comprehensive analysis of the HEPN superfamily: identification of novel roles in intra-genomic conflicts, defense, pathogenesis and RNA processing. *Biol Direct* **8**, 15, doi:10.1186/1745-6150-8-15 1745-6150-8-15 [pii] (2013).

24 Han, D., Lehmann, K. & Krauss, G. SSO1450--a CAS1 protein from Sulfolobus solfataricus P2 with high affinity for RNA and DNA. *FEBS Lett* **583**, 1928-1932 (2009).
25 Wiedenheft, B. *et al.* Structural basis for DNase activity of a conserved protein implicated in CRISPR-mediated genome defense. *Structure* **17**, 904-912 (2009).
26 Nunez, J. K., Lee, A. S., Engelman, A. & Doudna, J. A. Integrase-mediated spacer acquisition during CRISPR-Cas adaptive immunity. *Nature*, doi:10.1038/nature14237 nature14237 [pii] (2015).

27 Babu, M. *et al.* A dual function of the CRISPR-Cas system in bacterial antivirus immunity and DNA repair. *Mol Microbiol* **79**, 484-502, doi:10.1111/j.1365-2958.2010.07465.x (2011).
28 Wright, A. V. *et al.* A Functional Mini-Integrase in a Two-Protein-type V-C CRISPR System. *Mol Cell* **73**, 727-737 e723, doi:10.1016/j.molcel.2018.12.015 (2019).
29 Beloglazova, N. *et al.* A novel family of sequence-specific endoribonucleases associated with the clustered regularly interspaced short palindromic repeats. *J Biol Chem* **283**, 20361-20371 (2008).
30 Sinkunas, T. *et al.* Cas3 is a single-stranded DNA nuclease and ATP-dependent helicase in the CRISPR/Cas immune system. *EMBO J*, doi:emboj201141 [pii] 10.1038/emboj.2011.41 (2011).

31 Brouns, S. J. *et al.* Small CRISPR RNAs guide antiviral defense in prokaryotes. *Science* **321**, 960-964 (2008).
32 Han, D. & Krauss, G. Characterization of the endonuclease SSO2001 from Sulfolobus solfataricus P2. *FEBS Lett* **583**, 771-776 (2009).
33 Makarova, K. S., Grishin, N. V., Shabalina, S. A., Wolf, Y. I. & Koonin, E. V. A putative RNA-interference-based immune system in prokaryotes: computational analysis of the predicted enzymatic machinery, functional analogies with eukaryotic RNAi, and hypothetical mechanisms of action. *Biol Direct* **1**, 7 (2006).
34 Lemak, S. *et al.* Toroidal structure and DNA cleavage by the CRISPR-associated [4Fe-4S] cluster containing Cas4 nuclease SSO0001 from Sulfolobus solfataricus. *J Am Chem Soc* **135**, 17476-17487, doi:10.1021/ja408729b (2013).
35 Lemak, S. *et al.* The CRISPR-associated Cas4 protein Pcal_0546 from Pyrobaculum calidifontis contains a [2Fe-2S] cluster: crystal structure and nuclease activity. *Nucleic Acids Res* **42**, 11144-11155, doi:10.1093/nar/gku797 gku797 [pii] (2014).

36 Zhang, J., Kasciukovic, T. & White, M. F. The CRISPR associated protein Cas4 Is a 5' to 3' DNA exonuclease with an iron-sulfur cluster. *PLoS One* **7**, e47232, doi:10.1371/journal.pone.0047232 PONE-D-12-23364 [pii] (2012).

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

38 Lee, H., Dhingra, Y. & Sashital, D. G. The Cas4-Cas1-Cas2 complex mediates precise prespacer processing during CRISPR adaptation. *Elife* **8**, doi:10.7554/eLife.44248 (2019).

39 Jore, M. M. *et al.* Structural basis for CRISPR RNA-guided DNA recognition by Cascade. *Nat Struct Mol Biol*, doi:nsmb.2019 [pii] 10.1038/nsmb.2019 (2011).

40 Jore, M. M., Brouns, S. J. & van der Oost, J. RNA in Defense: CRISPRs Protect Prokaryotes against Mobile Genetic Elements. *Cold Spring Harb Perspect Biol*, doi:cshperspect.a003657 [pii] 10.1101/cshperspect.a003657 (2011).

41 Staals, R. H. *et al.* Structure and activity of the RNA-targeting Type III-B CRISPR-Cas complex of Thermus thermophilus. *Mol Cell* **52**, 135-145, doi:10.1016/j.molcel.2013.09.013 S1097-2765(13)00683-7 [pii] (2013).

42 Spilman, M. *et al.* Structure of an RNA silencing complex of the CRISPR-Cas immune system. *Mol Cell* **52**, 146-152, doi:10.1016/j.molcel.2013.09.008 S1097-2765(13)00678-3 [pii] (2013).

43 Rouillon, C. *et al.* Structure of the CRISPR interference complex CSM reveals key similarities with cascade. *Mol Cell* **52**, 124-134, doi:10.1016/j.molcel.2013.08.020 S1097-2765(13)00593-5 [pii] (2013).

44 Wiedenheft, B. *et al.* Structures of the RNA-guided surveillance complex from a bacterial immune system. *Nature* **477**, 486-489, doi:10.1038/nature10402 nature10402 [pii] (2011).

45 Koo, Y., Jung, D. K. & Bae, E. Crystal structure of Streptococcus pyogenes Csn2 reveals calcium-dependent conformational changes in its tertiary and quaternary structure. *PLoS One* **7**, e33401, doi:10.1371/journal.pone.0033401 PONE-D-11-18800 [pii] (2012).

46 Carte, J., Wang, R., Li, H., Terns, R. M. & Terns, M. P. Cas6 is an endoribonuclease that generates guide RNAs for invader defense in prokaryotes. *Genes Dev* **22**, 3489-3496 (2008).

47 Hale, C. R. *et al.* RNA-guided RNA cleavage by a CRISPR RNA-Cas protein complex. *Cell* **139**, 945-956 (2009).

48 Haurwitz, R. E., Jinek, M., Wiedenheft, B., Zhou, K. & Doudna, J. A. Sequence- and structure-specific RNA processing by a CRISPR endonuclease. *Science* **329**, 1355-1358 (2010).

49 Niewoehner, O., Jinek, M. & Doudna, J. A. Evolution of CRISPR RNA recognition and processing by Cas6 endonucleases. *Nucleic Acids Res* **42**, 1341-1353, doi:10.1093/nar/gkt922 gkt922 [pii] (2014).

50 Reeks, J. *et al.* Structure of the archaeal Cascade subunit Csa5: relating the small subunits of CRISPR effector complexes. *RNA Biol* **10**, 762-769, doi:10.4161/rna.23854 23854 [pii] (2013).

51 Sashital, D. G., Wiedenheft, B. & Doudna, J. A. Mechanism of foreign DNA selection in a bacterial adaptive immune system. *Mol Cell* **46**, 606-615, doi:10.1016/j.molcel.2012.03.020 S1097-2765(12)00227-4 [pii] (2012).

52 van Duijn, E. *et al.* Native tandem and ion mobility mass spectrometry highlight structural and modular similarities in clustered-regularly-interspaced shot-palindromic-repeats (CRISPR)-associated protein complexes from Escherichia coli and Pseudomonas aeruginosa. *Mol Cell Proteomics* **11**, 1430-1441, doi:10.1074/mcp.M112.020263 M112.020263 [pii] (2012).

53 Zhang, J. *et al.* Structure and mechanism of the CMR complex for CRISPR-mediated antiviral immunity. *Mol Cell* **45**, 303-313, doi:10.1016/j.molcel.2011.12.013 S1097-2765(11)00957-9 [pii] (2012).

54 Barrangou, R. *et al.* CRISPR provides acquired resistance against viruses in prokaryotes. *Science* **315**, 1709-1712, doi:315/5819/1709 [pii] 10.1126/science.1138140 (2007).

55 Garneau, J. E. *et al.* The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. *Nature* **468**, 67-71, doi:nature09523 [pii] 10.1038/nature09523 (2010).

56 Deltcheva, E. *et al.* CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. *Nature* **471**, 602-607, doi:10.1038/nature09886 nature09886 [pii] (2011).

57 Jinek, M. *et al.* A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **337**, 816-821, doi:10.1126/science.1225829 science.1225829 [pii] (2012).

58 Sapranauskas, R. *et al.* The Streptococcus thermophilus CRISPR/Cas system provides immunity in Escherichia coli. *Nucleic Acids Res* **39**, 9275-9282, doi:10.1093/nar/gkr606 gkr606 [pii] (2011).

59 Heler, R. *et al.* Cas9 specifies functional viral targets during CRISPR-Cas adaptation. *Nature* **519**, 199-202, doi:10.1038/nature14245 nature14245 [pii] (2015).

60 Wei, Y., Terns, R. M. & Terns, M. P. Cas9 function and host genome sampling in Type II-A CRISPR-Cas adaptation. *Genes Dev* **29**, 356-361, doi:10.1101/gad.257550.114 29/4/356 [pii] (2015).

61 Nishimasu, H. *et al.* Crystal structure of Cas9 in complex with guide RNA and target DNA. *Cell* **156**, 935-949, doi:10.1016/j.cell.2014.02.001 S0092-8674(14)00156-1 [pii] (2014).

62 Jinek, M. *et al.* Structures of Cas9 endonucleases reveal RNA-mediated conformational activation. *Science* **343**, 1247997, doi:10.1126/science.1247997 science.1247997 [pii] (2014).

63 Cocozaki, A. I. *et al.* Structure of the Cmr2 subunit of the CRISPR-Cas RNA silencing complex. *Structure* **20**, 545-553, doi:10.1016/j.str.2012.01.018 S0969-2126(12)00049-4 [pii] (2012).

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

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

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

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

68 Zetsche, B. *et al.* Cpf1 Is a Single RNA-Guided Endonuclease of a Class 2 CRISPR-Cas System. *Cell* **163**, 759-771, doi:10.1016/j.cell.2015.09.038 S0092-8674(15)01200-3 [pii] (2015).

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

70 Liu, L. *et al.* C2c1-sgRNA Complex Structure Reveals RNA-Guided DNA Cleavage Mechanism. *Mol Cell* **65**, 310-322, doi:S1097-2765(16)30813-9 [pii] 10.1016/j.molcel.2016.11.040 (2017).

71 Swarts, D. C., van der Oost, J. & Jinek, M. Structural Basis for Guide RNA Processing and Seed-Dependent DNA Targeting by CRISPR-Cas12a. *Mol Cell* **66**, 221-233 e224, doi:S1097-2765(17)30206-X [pii] 10.1016/j.molcel.2017.03.016 (2017).

72 Liu, J. J. *et al.* CasX enzymes comprise a distinct family of RNA-guided genome editors. *Nature* **566**, 218-223, doi:10.1038/s41586-019-0908-x (2019).

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

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

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

76 Zhang, C. *et al.* Structural Basis for the RNA-Guided Ribonuclease Activity of CRISPR-Cas13d. *Cell* **175**, 212-223 e217, doi:10.1016/j.cell.2018.09.001 (2018).

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

78 Zhang, B. *et al.* Two HEPN domains dictate CRISPR RNA maturation and target cleavage in Cas13d. *Nat Commun* **10**, 2544, doi:10.1038/s41467-019-10507-3 (2019).

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

80 Liu, L. *et al.* The Molecular Architecture for RNA-Guided RNA Cleavage by Cas13a. *Cell* **170**, 714-726 e710, doi:10.1016/j.cell.2017.06.050 (2017).

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

82 Knott, G. J. *et al.* Guide-bound structures of an RNA-targeting A-cleaving CRISPR-Cas13a enzyme. *Nat Struct Mol Biol* **24**, 825-833, doi:10.1038/nsmb.3466 (2017).

83 Zhang, B. *et al.* Structural insights into Cas13b-guided CRISPR RNA maturation and recognition. *Cell Res* **28**, 1198-1201, doi:10.1038/s41422-018-0109-4 (2018).

Supplementary Datasets

All datasets are available at ftp://ftp.ncbi.nih.gov/pub/wolf/_suppl/CRISPRclass19/

Supplementary Dataset 1. Detailed annotation of CRISPR-Cas loci

Supplementary Dataset 2. Multiple alignments (profiles) used to search for Cas and ancillary proteins.

Supplementary Dataset 3. Detailed information and comments for 126 modules derived from network analysis

Supplementary Dataset 4. Phylogenetic trees for the following protein families:

1. **Cas1**
2. **Cas9**
3. **TnpB and Cas12F variants**

Supplementary Dataset 5. Detailed annotation of CRISPR-Cas loci corresponding to new subtypes and yet unclassified systems.

Supplementary Dataset 6. Distribution of CRISPR-Cas systems across prokaryotic diversity