Database, 2019, 1–8
doi: 10.1093/database/baz093
Database tool

Database tool

CasPDB: an integrated and annotated database for Cas proteins from bacteria and archaea

Zhongjie Tang[1,†], ShaoQi Chen[1,†], Ang Chen[1], Bifang He[1,2], Yuwei Zhou[1], Guoshi Chai[1], FengBiao Guo[1,*] and Jian Huang[1,*]

[1]Center for Informational Biology, University of Electronic Science and Technology of China, Chengdu 611731, China and [2]School of Medicine, Guizhou University, Guiyang 550025, China

*Corresponding author: Tel: +86-28-8320-2351; Fax: +86-28-8320-8238; Email: fbguo@uestc.edu.cn

Correspondence may also be addressed to Jian Huang. Tel: +86-28-6183-0670; Fax: +86-28-8320-8238; Email: hj@uestc.edu.cn

[†]These authors contributed equally to the work.

Citation details: Tang,Z., Chen,S.Q., Chen,A. *et al.* CasPDB: an integrated and annotated database for Cas proteins from bacteria and archaea. *Database* (2019) Vol. 2019: article ID baz093; doi:10.1093/database/baz093

Received 26 November 2018; Revised 1 May 2019; Accepted 21 June 2019

Abstract

Clustered regularly interspaced short palindromic repeats (CRISPR) and associated proteins (Cas) constitute CRISPR–Cas systems, which are antiphage immune systems present in numerous bacterial and most archaeal species. In recent years, CRISPR–Cas systems have been developed into reliable and powerful genome editing tools. Nevertheless, finding similar or better tools from bacteria or archaea remains crucial. This requires the exploration of different CRISPR systems, identification and characterization new Cas proteins. Archives tailored for Cas proteins are urgently needed and necessitate the prediction and grouping of Cas proteins into an information center with all available experimental evidence. Here, we constructed Cas Protein Data Bank (CasPDB), an integrated and annotated online database for Cas proteins from bacteria and archaea. The CasPDB database contains 287 reviewed Cas proteins, 257 745 putative Cas proteins and 3593 Cas operons from 32 023 bacteria species and 1802 archaea species. The database can be freely browsed and searched. The CasPDB web interface also represents all the 3593 putative Cas operons and its components. Among these operons, 328 are members of the type II CRISPR–Cas system.

Database URL: http://immunet.cn/caspdb/

Introduction

Clustered regularly interspaced short palindromic repeats (CRISPR) and their associated proteins (Cas) constitute CRISPR–Cas systems (1). Since the first biological evidence for the participation of CRISPR–Cas systems in adaptive immunity was reported in 2007 (2, 3), these systems have gradually become a research hotspot (4, 5), further expanding biotechnological toolkits, and especially revolutionizing the genome editing technology (6). As the adaptive immunity system of archaea and bacteria, CRISPR–Cas systems imprint exogenic elements as memories by inserting them into an array of CRISPR repeats, and then the inserted fragment can be used in the form of guide CRISPR RNAs (crRNAs) that cooperate with Cas proteins to recognize and shear the cognate viral genome or plasmid upon new infection (7). For example, the CRISPR–Cas9 technology derived from the *Streptococcus pyogenes* engineers the dual tracrRNA:crRNA into a single guide RNA (sgRNA) that retains two critical features: a 20-nucleotide sequence at the 5′ end of the sgRNA that undergoes Watson–Crick base pairing with any DNA sequence of interest and a double-stranded structure at the 3′ side of the guide sequence that binds to Cas9 (8, 9). Therefore, the exploration of the mechanisms of CRISPR–Cas systems requires understanding and characterization of new Cas proteins, and this may also be a rational start point for the discovery of novel scissors for genome-editing technology.

A wealth of structural and functional information on core Cas proteins has accumulated through the continuous implementation of large-scale research (10–13) and has facilitated the classification of CRISPR–Cas systems. At present, class 1 CRISPR–Cas systems include the CRISPR–Cas type I, type III and type IV systems. Their effector modules are composed of several Cas proteins (10, 14). The class 2 CRISPR–Cas systems mainly include CRISPR–Cas type II, type V and type VI systems. Their effector module consists of a single, multidomain Cas protein. The common class 2 type effector includes Cas9 (Csn1, II), Cas12a (Cpf1, VA), Cas12b (C2c1, VB), Cas13a (C2c2, VI-A), Cas13b (C2c6, VI-B) and Cas13c (C2c7, VI-C). As for Cas9, it contains two unrelated nuclease domains, RuvC and HNH, responsible for cleavage of the displaced (non-target) and target DNA strands, respectively (10, 14–16). Totally, the annotation and recognition of Cas proteins in bacteria and archaea are necessary not only for correctly inferring the type of CRISPR–Cas systems but also for supporting the development of novel scissors for gene-editing technology.

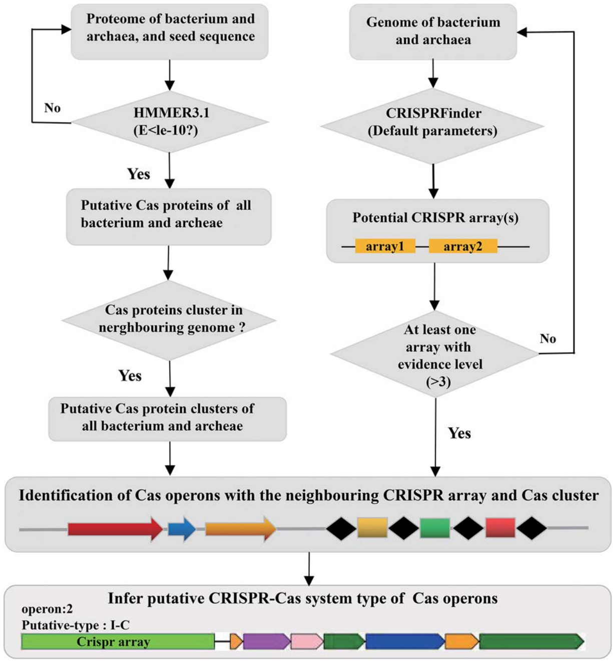

Proteome of bacterium and archaea, and seed sequence
HMMER3.1 (E<le-10?)
No
Yes
Putative Cas proteins of all bacterium and archeae
Cas proteins cluster in nerghbouring genome ?
Yes
Putative Cas protein clusters of all bacterium and archeae
Genome of bacterium and archaea
CRISPRFinder (Default parameters)
Potential CRISPR array(s)
array1
array2
At least one array with evidence level (>3)
No
Yes
Identification of Cas operons with the neighbouring CRISPR array and Cas cluster
Infer putative CRISPR-Cas system type of Cas operons
operon:2
Putative-type : I-C
Crispr array

Figure 1. The construction of CasPDB.

Unfortunately, the UniProt protein resource contains only 287 bacterial and archaeal Cas proteins that have been manually reviewed (17). Many novel Cas proteins await discovery. Thus, predicting Cas proteins from bacteria and archaea proteomic data and integrating them into a specific database is highly important. Some of the web services and databases about the CRISPR–Cas system have been developed. For example, the well-known CRISPRFinder (18) is a web tool for searching CRISPRs, and CRISPRCasFinder (19) supports to search CRISPRs and the Cas gene. Relevant databases include CRISPRdb (20), a data resource of CRISPRs, and CrisprGE (21), a useful web resource for genome editing that provides comprehensive information of the genome editing approach, including target sequence, modification length and CrisprID. At present, no specific database is available for annotating and housing Cas proteins. Therefore, a comprehensive data resource for potential Cas proteins and Cas operons from bacteria and archaea is still necessary. Some computational methods have been used for the rapid annotation of protein sequences. The Hidden Markov Model (HMM) technique (22) is one of the methods used to identify potential Cas proteins from proteomic data. We built the HMMCAS website (http://i.uestc.edu.cn/hmmcas/index.html) to predict Cas proteins (23). Herein, we identified proteome-wide putative Cas proteins across bacteria and archaea using the HMM technique, screened Cas protein clusters and the neighboring CRISPR array through CRISPRFinder and inferred the potential CRISPR–Cas system type. All the information was stored in a database called CasPDB, which is short for Cas Protein Data Bank. It can be conveniently and flexibly browsed, searched and downloaded. The putative Cas proteins and operons can be easily visualized.

Materials and Methods

The construction of CasPDB is illustrated in Figure 1.

Data Collection and Model Construction

Firstly, we downloaded 101 and 38 CRISPR-associated multiple sequence alignments or seed alignments of protein families from the TIGRFAMs (version 15.0) (24) and Pfam (version 30.0) databases (25), respectively. We also retrieved 24 novel Cas seed alignments collected from http://omics.informatics.indiana.edu/mg/CAS (26). Secondly, we used these CRISPR-associated seed alignments to construct the HMMs by the hmmbuild program with default parameters. We then employed the HMMs to predict all candidate Cas proteins from the whole proteome of different bacterial and archaeal species collected from GenBank (all data for archaea is downloaded from https://ftp.ncbi.nih.gov/genomes/genbank/archaea/,\ignorespacesand that of bacteria is from https://ftp.ncbi.nih.gov/genomes/genbank/bacteria). The hmmbuild program (using the gathering cutoff) in HMMER3.1 (22,27) is implemented by Python scripts (https://www.python.org/).

Identification of Putative Cas Proteins and Cas Operons

Using the hmmscan homology search program to compare the HMM modules (HMMs) against the complete proteomic sequence data of 32 023 and 18 021 bacterial and archaeal species, we obtained 257 745 putative Cas proteins. If two or more Cas proteins in a proteome are not spaced by three or more non-Cas proteins, these Cas proteins form a Cas protein cluster. We got a total of 30 495 Cas protein clusters. We also collected 287 Cas proteins from UniProt which are manually reviewed (17). The genome data of bacteria and archaea is from the same URL of proteome data. To further identify potential Cas operons in bacteria and archaea, we used CRISPRFinder (18) to genome-wide screen potential CRISPR arrays on the genome adjacent to the Cas protein clusters. If a CRISPR array locates to a Cas protein cluster within 5000 bp, we define the Cas protein cluster and the neighboring CRISPR array a Cas operon. Finally, based on the composition of the Cas proteins in the cluster and their relative genomic location, the types of possible CRISPR–Cas systems can be judged. For example, the cluster with Cas9–Cas1–Cas2–Csn2 in the Cas operon will be defined as the type II-A system, clusters with Cas3 belong to type I and Cas9 type II, Cas10 type III, etc. All putative Cas proteins were mapped back to the genome, and the corresponding protein accession number was provided (Figure 1).

Nomenclature Standardization and Classification

We standardized the species name and taxonomy ID in accordance with the NCBI Taxonomy (https://www.ncbi.nlm.nih.gov/Taxonomy/). Protein names and corresponding accession numbers were also standardized with NCBI GenBank. The putative CRISPR–Cas system type was identified on the basis of the core putative Cas proteins and their relatively genomic location in the Cas operon.

Maintenance and Quality Assurance

As described above, we developed a pipeline to identify putative Cas proteins and Cas operons according to the genome and proteome data stored at NCBI. The Cas proteins were determined using the HMMER program based on the Cas seed files collected from multiple resources. The species name, taxonomy ID, protein names and relevant accession numbers were in accordance with the corresponding databases of NCBI. The putative type of Cas operons was named according to the newest nomenclature of CRISPR–Cas systems. As our pipeline strictly agrees with the data standards of NCBI, the data quality of CasPDB can be assured. We plan to run the pipeline and update the CasPDB database annually to keep up with the update of the genome and proteome data of archaea and bacteria, the new collections of the Cas seed files, the new version of relevant NCBI databases and the new rules for the classification of CRISPR–Cas systems. The online version of the CasPDB database is always the newest release. All previous releases will be archived and can be downloaded from the download page.

Implementation of Web Interface

All metadata in CasPDB were stored in a MySQL database. The CasPDB website is implemented with the HyperText Markup Language (HTML, https://www.w3.org/), Cascading Style Sheets (CSS), and JavaScript. ECharts 3 (http://echarts.baidu.com/) were used for the web-based visualization of Cas operons.

Results

Data Statistics

At present, the CasPDB database contains 287 reviewed Cas proteins, 257 745 putative Cas proteins, 30 495 Cas protein clusters and 3593 Cas operons from 32 023 bacterial species and 1802 archaeal species (Figure 2A). In all candidate Cas proteins, Cas1 accounts for 20.11%, Cas2 accounts for 17.46% and Cas3, Cas5 and Cas6 accounts for more than 10%. One of the major editing proteins, namely Cas9, only accounts for 2.59% of the candidate proteins (Figure 2B). No putative Cas12 or Cas13 proteins were identified because these types of Cas proteins lack seed alignments at the time when we ran the pipeline. Moreover, we identified 328 putative operons of the type II system. These CRISPR–Cas9 systems can be used for further analysis on genome engineering.

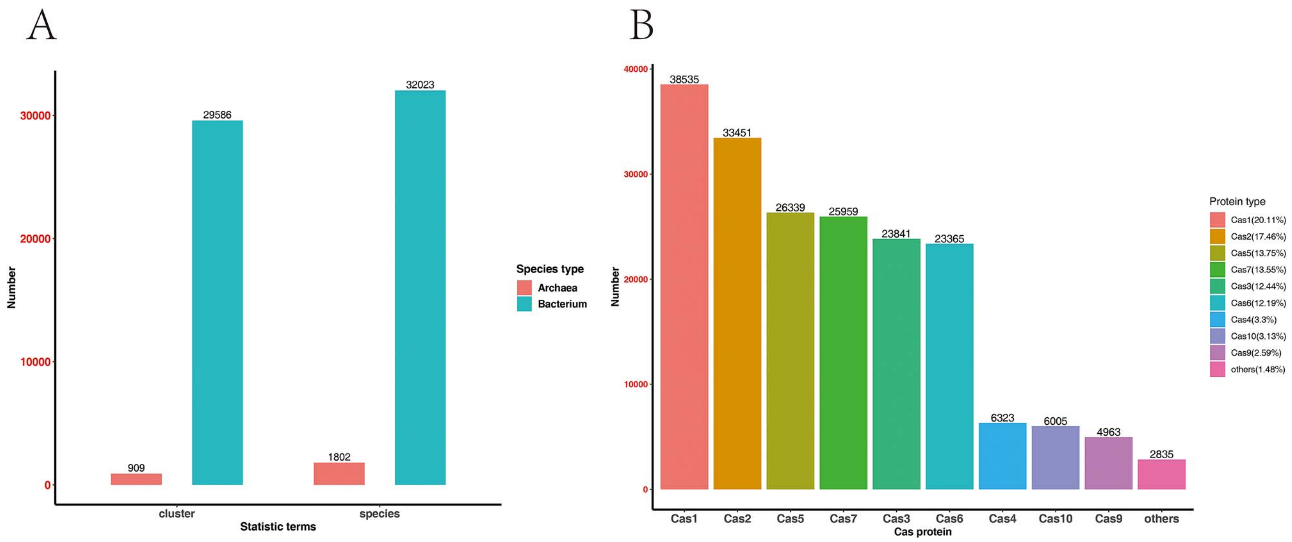


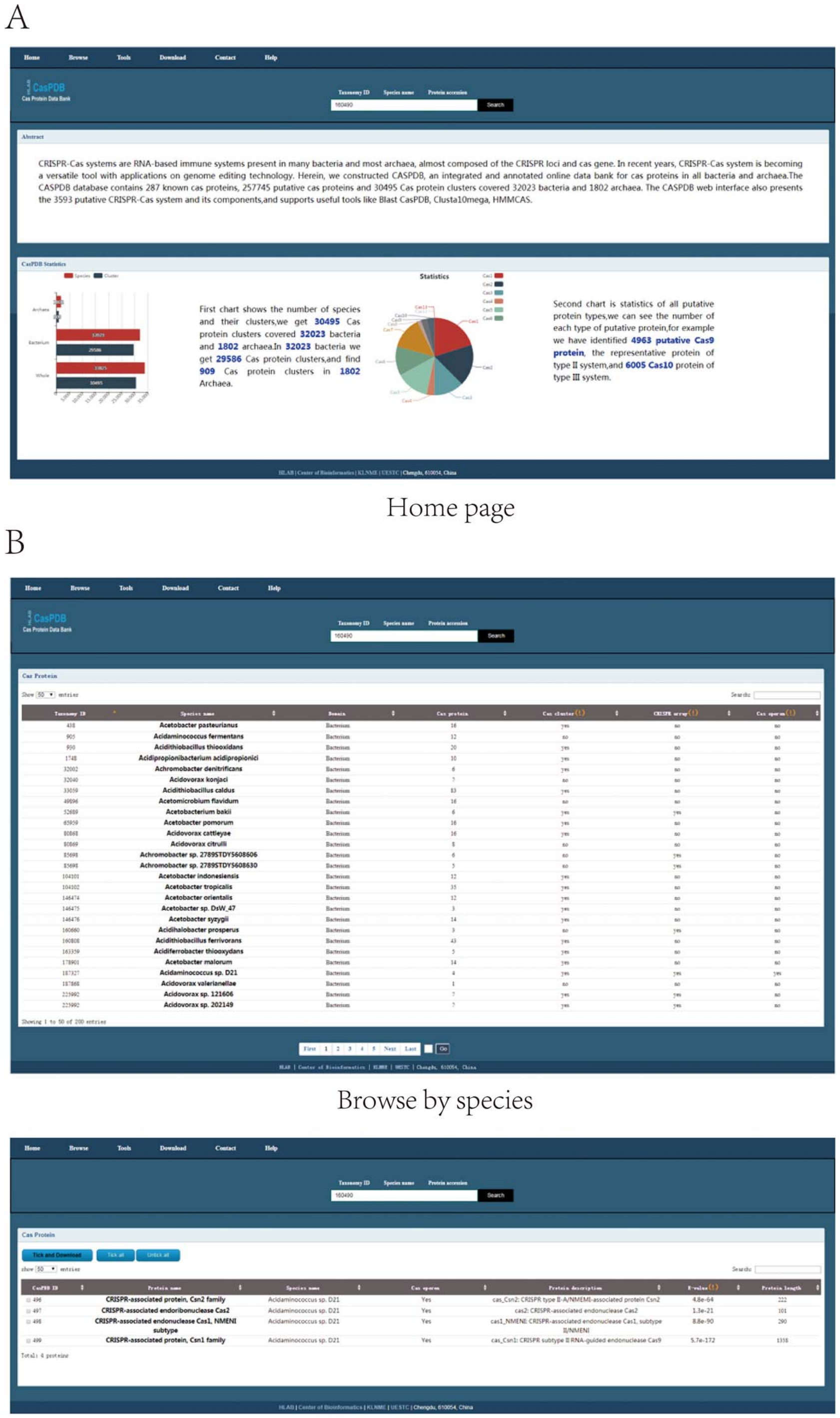
Figure 2. Statistics of CasPDB. The number of bacterial and archaeal species and their putative Cas operons **(A)**. Statistical distribution of each type of Cas proteins **(B)**.

Guidance of CasPDB

The CasPDB web interface enables the searching, browsing and downloading of putative Cas proteins and operons from bacterial and archaeal species. Firstly, the search panel accepts three types of key words: Taxonomy ID, Species name and Protein accession. Users can input the taxonomy ID or scientific species name of the NCBI Taxonomy database, or protein accession of NCBI GenBank to find Cas proteins. The home page describes the basic information of CasPDB and shows statistics of the Cas proteins, Cas protein clusters and species (see in Figure 3A). As shown on the browse page, all putative Cas protein numbers and putative CRISPR–Cas system distributions across all bacteria and archaea have been listed with the following information: Taxonomy ID, Species name (scientific name of bacteria or archaea), Domain (Cas proteins from Bacteria or Archaea), Cas protein (number of putative Cas proteins in target species), Cas cluster (includes putative Cas protein clusters or not), CRISPR array (includes putative CRISPR array or not) and Cas operon (includes Cas operon or not). Upon further clicking the species name in the browse page, all putative Cas proteins of target species will be listed with the following information: CasPDB ID (CasPDB ID of putative Cas protein), protein name (putative Cas protein name), species name, Cas operon, protein description (detailed annotation information of putative Cas protein), E-value (the statistical significance of the possibility that a candidate protein is a Cas protein) and protein length (see Figure 3B). Secondly, the detail page of putative Cas protein will display protein information, cluster of Cas proteins if any, protein sequence and putative CRISPR–Cas system, if any. Protein information includes the basic protein information and protein accession number link to the NCBI GenBank database. A cluster of Cas proteins figures genomic neighboring Cas proteins, labeled with different colors and detail protein information. The putative CRISPR–Cas system displays the element of the Cas operon, includes the Cas protein cluster and CRISPR array and is labeled with corresponding putative Cas protein basic information and hyperlinked to the corresponding CasPDB page (see Figure 4A). Finally, on the download page, users can download all putative Cas proteins, species information and known Cas proteins from UniProt across all bacteria and archaea with their basic information in tab-delimited format or FASTA format. Besides, users can also download only the items of interest through checking the box in front of the CasPDB ID when browsing a species or on the search result page (Figure 4B). The CasPDB database also provides additional information and an external interface of other prediction tools for putative Cas proteins, such as Clustal Omega, Blast CasPDB and HMMCAS (23). The tool Blast CasPDB can blast the user's query protein sequences against the CasPDB database. By comprehensively using these tools, users can conveniently obtain more complete characteristics of Cas proteins.

Identification and Presentation of Putative Cas Operons

In CasPDB, we identified 3593 Cas operons, among which 328 belongs to the type II CRISPR–Cas systems. In the detail page, the picture of the putative CRISPR–Cas system shows operon (operon ordinal in target species), operon-putative type (putative CRISPR–Cas system type of



CRISPR-Cas systems are RNA-based immune systems present in many bacteria and most archaea, almost composed of the CRISPR loci and cas gene. In recent years, CRISPR-Cas system is becoming a versatile tool with applications on genome editing technology. Herein, we constructed CASPDB, an integrated and annotated online data bank for cas proteins in all bacteria and archaea.The CASPDB database contains 287 known cas proteins, 257745 putative cas proteins and 30495 Cas protein clusters covered 32023 bacteria and 1802 archaea. The CASPDB web interface also presents the 3593 putative CRISPR-Cas system and its components,and supports useful tools like Blast CasPDB, Clusta10mega, HMMCAS.

First chart shows the number of species and their clusters,we get **30495** Cas protein clusters covered **32023** bacteria and **1802** archaea.In **32023** bacteria we get **29586** Cas protein clusters,and find **909** Cas protein clusters in **1802** Archaea.

Second chart is statistics of all putative protein types,we can see the number of each type of putative protein,for example we have identified **4963 putative Cas9 protein**, the representative protein of type II system,and **6005 Cas10** protein of type III system.

Home page

Show 50 ▾ entries

Taxonomy ID	Species name	Domain	Cas protein	Cas cluster(!)	CRISPR array(!)	Cas operon(!)
438	Acetobacter pasteurianus	Bacterium	16	yes	no	no
905	Acidaminococcus fermentans	Bacterium	12	no	no	no
930	Acidithiobacillus thiooxidans	Bacterium	20	yes	no	no
1748	Acidipropionibacterium acidipropionici	Bacterium	10	yes	no	no
32002	Achromobacter denitrificans	Bacterium	6	yes	no	no
32040	Acidovorax konjaci	Bacterium	7	no	no	no
33059	Acidithiobacillus caldus	Bacterium	83	yes	no	no
49896	Acetomicrobium flavidum	Bacterium	16	no	no	no
52689	Acetobacterium bakii	Bacterium	6	yes	yes	no
65959	Acetobacter pomorum	Bacterium	16	yes	no	no
80868	Acidovorax cattleyae	Bacterium	16	yes	no	no
80869	Acidovorax citrulli	Bacterium	8	no	no	no
85698	Achromobacter sp. 2789STDY5608606	Bacterium	6	no	yes	no
85698	Achromobacter sp. 2789STDY5608630	Bacterium	5	no	yes	no
104101	Acetobacter indonesiensis	Bacterium	12	yes	no	no
104102	Acetobacter tropicalis	Bacterium	35	yes	no	no
146474	Acetobacter orientalis	Bacterium	12	yes	no	no
146475	Acetobacter sp. DsW_47	Bacterium	3	yes	no	no
146476	Acetobacter syzygii	Bacterium	14	yes	no	no
160660	Acidihalobacter prosperus	Bacterium	3	no	yes	no
160808	Acidithiobacillus ferrivorans	Bacterium	43	yes	no	no
163359	Acidiferrobacter thiooxydans	Bacterium	5	yes	no	no
178901	Acetobacter malorum	Bacterium	14	yes	no	no
187327	Acidaminococcus sp. D21	Bacterium	4	yes	yes	yes
187868	Acidovorax valerianellae	Bacterium	1	no	no	no
225992	Acidovorax sp. 121606	Bacterium	7	yes	yes	no
225992	Acidovorax sp. 202149	Bacterium	7	yes	yes	no

Showing 1 to 50 of 200 entries

Search:

Browse by species

show 50 ▾ entries

Search:

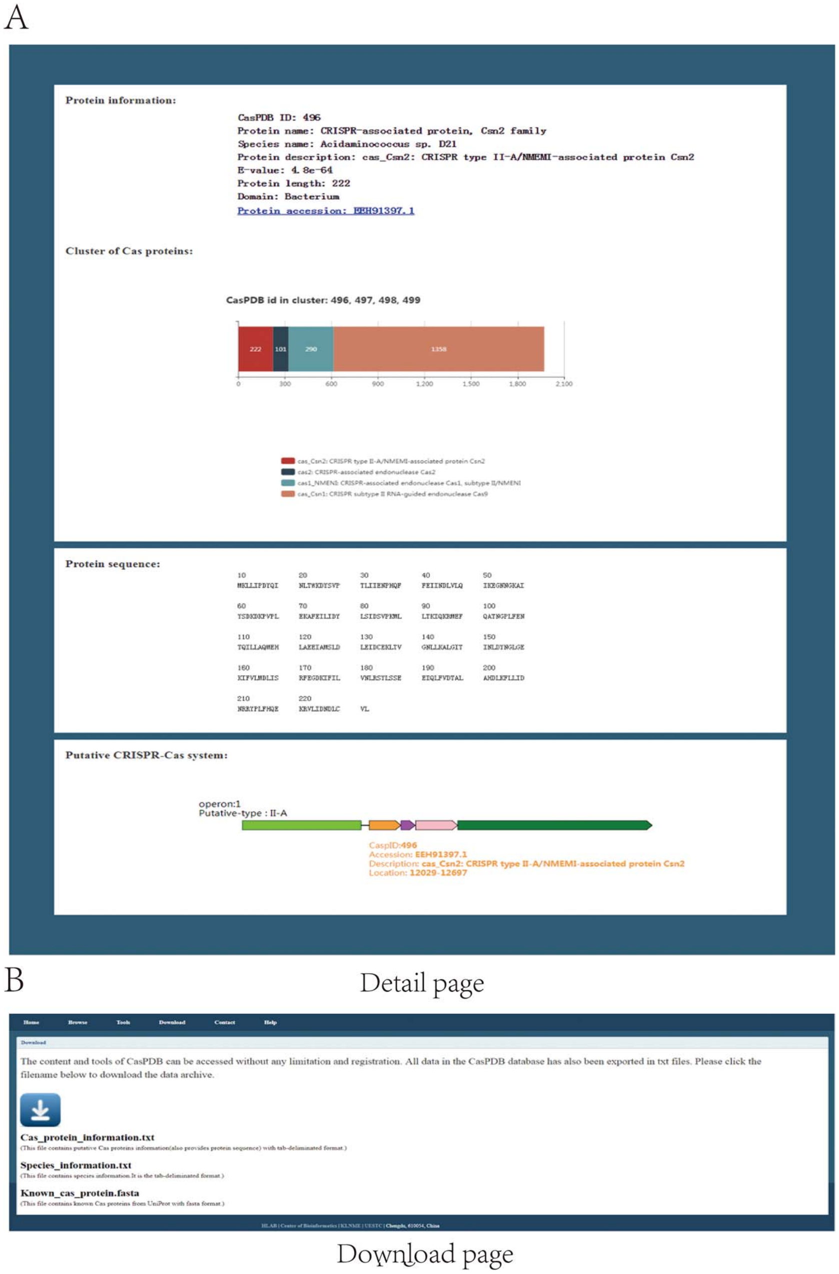
CasPDB ID	Protein name	Species name	Cas operon	Protein description	E-value(!)	Protein length
□ 496	CRISPR-associated protein, Csn2 family	Acidaminococcus sp. D21	Yes	cas_Csn2: CRISPR type II-A/NMEMI-associated protein Csn2	4.8e-64	222
□ 497	CRISPR-associated endoribonuclease Cas2	Acidaminococcus sp. D21	Yes	cas2: CRISPR-associated endonuclease Cas2	1.3e-21	101
□ 498	CRISPR-associated endonuclease Cas1, NMENI subtype	Acidaminococcus sp. D21	Yes	cas1_NMENI: CRISPR-associated endonuclease Cas1, subtype II/NMENI	8.8e-90	290
□ 499	CRISPR-associated protein, Csn1 family	Acidaminococcus sp. D21	Yes	cas_Csn1: CRISPR subtype II RNA-guided endonuclease Cas9	5.7e-172	1338

Total: 4 proteins

Browse by proteins of target species

Figure 3. Home and browse pages. Search option and home page **(A)**. Browse page with all Cas proteins. The bottom section of the page shows protein distribution in bacteria **(B)**.

Cas operon), Cas proteins (different color arrows) and CRISPR array (green rectangle). When the mouse moves over any Cas protein icon, the detail information including CasPDB ID, Accession, Description, Location (genomic location of putative Cas protein) will be represented. Besides, clicking the protein icon in the operon figure, the page will skip to the detail page of the corresponding protein in CasPDB. When the mouse moves over the CRISPR array, the detail information with Element (type of element in operon), Spacer_num (number of



Protein information:

CasPDB ID: 496
Protein name: CRISPR-associated protein, Csn2 family
Species name: Acidaminococcus sp. D21
Protein description: cas_Csn2: CRISPR type II-A/NMEMI-associated protein Csn2
E-value: 4.8e-64
Protein length: 222
Domain: Bacterium
Protein accession: EEH91397.1

Cluster of Cas proteins:

CasPDB id in cluster: 496, 497, 498, 499

🟥 cas_Csn2: CRISPR type II-A/NMEMI-associated protein Csn2
■ cas2: CRISPR-associated endonuclease Cas2
■ cas1_NMENI: CRISPR-associated endonuclease Cas1, subtype II/NMENI
■ cas_Csn1: CRISPR subtype II RNA-guided endonuclease Cas9

Protein sequence:

```
10          20          30          40          50
MKLLIPDYQI  NLTWKDYSVP  TLIIENPHQF  FEIINDLVLQ  IKEGNNGKAI

60          70          80          90          100
YSDKDKPVPL  EKAFEILIDY  LSIDSVPKML  LTKIQKRMEF  QATNGPLFEN

110         120         130         140         150
TQILLAQMEM  LAEEIAMSLD  LEIDCEKLTV  GNLLKALGIT  INLDYNGLGE

160         170         180         190         200
KIFVLMDLIS  RFEGDKIFIL  VNLRSYLSSE  EIQLFVDTAL  AMDLKFLLID

210         220
NRRYPLFHQE  KRVLIDNDLC  VL
```

Putative CRISPR-Cas system:

operon:1
Putative-type : II-A

CaspID:496
Accession: EEH91397.1
Description: cas_Csn2: CRISPR type II-A/NMEMI-associated protein Csn2
Location: 12029-12697

Detail page

The content and tools of CasPDB can be accessed without any limitation and registration. All data in the CasPDB database has also been exported in txt files. Please click the filename below to download the data archive.

Cas_protein_information.txt
(This file contains putative Cas proteins information(also provides protein sequence) with tab-deliminated format.)

Species_information.txt
(This file contains species information.It is the tab-deliminated format.)

Known_cas_protein.fasta
(This file contains known Cas proteins from UniProt with fasta format.)

HLAB | Center of Bioinformatics | KLNME | UESTC | Chengdu, 610054, China

Download page

Figure 4. Detail and download pages. The detail page shows basic information of protein and operons **(A)**. Download option for all Cas proteins and proteins of selected items **(B)**.

spacer), Description (repeat sequence in CRISPR), Location (genomic location of CRISPR array) will be displayed (see in Figure 4A).

Discussion

Over the past 10 years, with the deep understanding of the CRISPR–Cas systems, RNA-guided DNA endonucleases, such as Cas9 and c2c2, have been programmed to be genome editing tools (9, 28). The continuous expansion of CRISPR–Cas systems has provided prospective novel genome-editing tools (29). Unfortunately, the numbers of experimentally validated and annotated Cas proteins remain insufficient for meeting the need of researchers, and no specialized database for Cas proteins has been designed. Hence, putative Cas proteins from the whole proteome of bacterial and archaeal species should be studied systematically. In our work, we identified putative Cas proteins from proteomic data based on the Cas-related seed alignments of experimentally characterized Cas proteins, and annotation information of basic protein information, amino acid sequence and genomic location for each putative Cas protein. We further identified the genome-wide CRISPR array and screened their genomic neighboring (<5000 bp) Cas protein clusters, and formed putative Cas operons. According to the core Cas proteins and their relative order in Cas operons, we classified the potential CRISPR–Cas system type of Cas operons. The CasPDB database will help users to find the potential CRISPR–Cas systems in bacteria and archaea and further infer other elements in the CRISPR–Cas systems, such as tracrRNA and PAM. The comprehensive analysis of these elements in the same CRISPR–Cas system and the optimization of their annotated information may be a highly effective approach for identifying new scissors for genome editing (30, 31). We believe that CasPDB is a valuable data resource for bacterial and archaeal Cas proteins and can aid in the discovery of new scissors for genome editing.

In conclusion, we provided a user-friendly website with data retrieval capabilities and browsing, searching and downloading options to facilitate data access. In the future, we will continue to collect new CRISPR-associated multiple sequence alignments and seed alignments and integrate proteome and genome data of new bacteria and archaea species to screen new potential Cas proteins, further combining the genomic and proteomic data to identity the putative Cas proteins and other elements of CRISPR–Cas systems, to provide a complete perspective for understanding what role CRISPR–Cas systems play and to support the use of valuable CRISPR–Cas systems or Cas proteins as new scissors for genome-editing technology.

Acknowledgments

The authors are grateful to the anonymous reviewers for their valuable suggestion and comments, which have led to the improvement of this paper.

Funding

National Natural Science Foundation of China (grant number 61571095); Fundamental Research Funds for the Central Universities of China (ZYGX2005Z006).

Conflict of interest. None declared.

References

1. Barrangou,R. and Marraffini,L.A. (2014) CRISPR-Cas systems: prokaryotes upgrade to adaptive immunity. *Mol. Cell*, **54**, 234–244.
2. Marraffini,L.A. (2015) CRISPR-Cas immunity in prokaryotes. *Nature*, **526**, 55–61.
3. Barrangou,R., Fremaux,C., Deveau,H. *et al.* (2007) CRISPR provides acquired resistance against viruses in prokaryotes. *Science*, **315**, 1709–1712.
4. Ju,X.D., Xu,J. and Sun,Z.S. (2018) CRISPR editing in biological and biomedical investigation. *J. Cell. Biochem.*, **119**, 52–61.
5. Richter,H., Randau,L. and Plagens,A. (2013) Exploiting CRISPR/Cas: interference mechanisms and applications. *Int. J. Mol. Sci.*, **14**, 14518–14531.
6. Choi,K.R. and Lee,S.Y. (2016) CRISPR technologies for bacterial systems: current achievements and future directions. *Biotechnol. Adv.*, **34**, 1180–1209.
7. Horvath,P. and Barrangou,R. (2010) CRISPR/Cas, the immune system of bacteria and archaea. *Science*, **327**, 167–170.
8. Doudna,J.A. and Charpentier,E. (2014) Genome editing. The new frontier of genome engineering with CRISPR-Cas9. *Science*, **346**, 1258096.
9. Wang,H., La Russa,M. and Qi,L.S. (2016) CRISPR/Cas9 in genome editing and beyond. *Annu. Rev. Biochem.*, **85**,227–264.
10. Makarova,K.S., Wolf,Y.I., Alkhnbashi,O.S. *et al.* (2015) An updated evolutionary classification of CRISPR-Cas systems. *Nat. Rev. Microbiol.*, **13**, 722–736.
11. Nunez,J.K., Kranzusch,P.J., Noeske,J. *et al.* (2014) Cas1-Cas2 complex formation mediates spacer acquisition during CRISPR-Cas adaptive immunity. *Nat. Struct. Mol. Biol.*, **21**, 528–534.
12. Westra,E.R., Swarts,D.C., Staals,R.H. *et al.* (2012) The CRISPRs, they are a-changin': how prokaryotes generate adaptive immunity. *Annu. Rev. Genet.*, **46**, 311–339.
13. Makarova,K.S., Wolf,Y.I. and Koonin,E.V. (2013) The basic building blocks and evolution of CRISPR-CAS systems. *Biochem. Soc. Trans.*, **41**, 1392–1400.
14. Shmakov,S., Smargon,A., Scott,D. *et al.* (2017) Diversity and evolution of class 2 CRISPR-Cas systems. *Nat. Rev. Microbiol.*, **15**, 169–182.
15. Tang,Y. and Fu,Y. (2018) Class 2 CRISPR/Cas: an expanding biotechnology toolbox for and beyond genome editing. *Cell Biosci.*, **8**, 59.
16. Makarova,K.S., Haft,D.H., Barrangou,R. *et al.* (2011) Evolution and classification of the CRISPR-Cas systems. *Nat. Rev. Microbiol.*, **9**, 467–477.
17. UniProt,C. (2015) UniProt: a hub for protein information. *Nucleic Acids Res.*, **43**, D204–D212.
18. Grissa,I., Vergnaud,G. and Pourcel,C. (2007) CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. *Nucleic Acids Res.*, **35**, W52–W57.
19. Couvin,D., Bernheim,A., Toffano-Nioche,C. *et al.* (2018) CRISPRCasFinder, an update of CRISRFinder, includes a portable version, enhanced performance and integrates search for Cas proteins. *Nucleic Acids Res.*, **46**, W246–W251.
20. Grissa,I., Vergnaud,G. and Pourcel,C. (2007) The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats. *BMC bioinformatics*, **8**, 172.
21. Kaur,K., Tandon,H., Gupta,A.K. *et al.* (2015) CrisprGE: a central hub of CRISPR/Cas-based genome editing. *Database (Oxford)*, **2015**, bav055.
22. Finn,R.D., Clements,J., Arndt,W. *et al.* (2015) HMMER web server: 2015 update. *Nucleic Acids Res.*, **43**, W30–W38.
23. Chai,G., Yu,M., Jiang,L. *et al.* (2017) HMMCAS: a web tool for the identification and domain annotations of Cas proteins. *IEEE/ACM Trans. Comput. Biol. Bioinform.*

24. Haft,D.H., Selengut,J.D. and White,O. (2003) The TIGRFAMs database of protein families. *Nucleic Acids Res.*, **31**, 371–373.
25. Punta,M., Coggill,P.C., Eberhardt,R.Y. *et al.* (2012) The Pfam protein families database. *Nucleic Acids Res.*, **40**, D290–D301.
26. Zhang,Q., Doak,T.G. and Ye,Y. (2014) Expanding the catalog of cas genes with metagenomes. *Nucleic Acids Res.*, **42**, 2448–2459.
27. Finn,R.D., Clements,J. and Eddy,S.R. (2011) HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.*, **39**, W29–W37.
28. Senthong,V., Li,X.S., Hudec,T. *et al.* (2016) Plasma trimethylamine N-oxide, a gut microbe-generated phosphatidylcholine metabolite, is associated with atherosclerotic burden. *J. Am. Coll. Cardiol.*, **67**, 2620–2628.
29. Murugan,K., Babu,K., Sundaresan,R. *et al.* (2017) The revolution continues: newly discovered systems expand the CRISPR-Cas toolkit. *Mol. Cell*, **68**, 15–25.
30. Chyou,T.Y. and Brown,C.M. (2018) Prediction and diversity of tracrRNAs from type II CRISPR-Cas systems. *RNA Biol.*, 1–12.
31. Mendoza,B.J. and Trinh,C.T. (2018) In silico processing of the complete CRISPR-Cas spacer space for identification of PAM sequences. *Biotechnol. J.*, **13**, e1700595.