D590–D598 *Nucleic Acids Research, 2020, Vol. 48, Database issue* *Published online 17 October 2019*
doi: 10.1093/nar/gkz916

PADS Arsenal: a database of prokaryotic defense systems related genes

Yadong Zhang[1,2,3,4], Zhewen Zhang[1,2,3,*], Hao Zhang[1,2,3,4], Yongbing Zhao[5], Zaichao Zhang[6] and Jingfa Xiao[1,2,3,4,*]

[1]National Genomics Data Center, Beijing 100101, China, [2]BIG Data Center, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China, [3]CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China, [4]College of Life Sciences, University of Chinese Academy of Sciences, Beijing 100049, China, [5]Department of Health Sciences Research, Mayo Clinic, Jacksonville, FL 32224, USA and [6]Department of Biology, The University of Western Ontario, London, Ontario N6A 5B7, Canada

Received August 12, 2019; Revised October 03, 2019; Editorial Decision October 03, 2019; Accepted October 04, 2019

ABSTRACT

Defense systems are vital weapons for prokaryotes to resist heterologous DNA and survive from the constant invasion of viruses, and they are widely used in biochemistry investigation and antimicrobial drug research. So far, numerous types of defense systems have been discovered, but there is no comprehensive defense systems database to organize prokaryotic defense gene datasets. To fill this gap, we unveil the prokaryotic antiviral defense system (PADS) Arsenal (https://bigd.big.ac.cn/padsarsenal), a public database dedicated to gathering, storing, analyzing and visualizing prokaryotic defense gene datasets. The initial version of PADS Arsenal integrates 18 distinctive categories of defense system with the annotation of 6 600 264 genes retrieved from 63,701 genomes across 33 390 species of archaea and bacteria. PADS Arsenal provides various ways to retrieve defense systems related genes information and visualize them with multifarious function modes. Moreover, an online analysis pipeline is integrated into PADS Arsenal to facilitate annotation and evolutionary analysis of defense genes. PADS Arsenal can also visualize the dynamic variation information of defense genes from pan-genome analysis. Overall, PADS Arsenal is a state-of-the-art open comprehensive resource to accelerate the research of prokaryotic defense systems.

INTRODUCTION

As mentioned in the Red Queen hypothesis, the ongoing and competitive arms race is one of the most powerful driving factors in co-evolution between prokaryotic organisms and viruses (1–3). As a consequence, prokaryotes have evolved numerous diverse and elaborate defense systems to protect themselves against viruses (4). Based on their action modes, the defense systems can be divided into two major groups, immunity and dormancy induction or programmed cell death (5,6). The immunity group contains restriction-modification (RM) system (7,8), DNA phosphorothioation system (known as DND system) (9–11), defense island system associated with restriction-modification (DISARM) system (12), bacteriophage exclusion (BREX) system (13), prokaryotic Argonautes (pAgos) system (14,15), and clustered regularly interspaced short palindromic repeats and adjacent to *cas* genes (CRISPR-Cas) system (16–19). The dormancy induction or programmed cell death by infection group includes toxin-antitoxin (TA) system (20–22) and abortive infection (ABI) system (23). Recently, several new types of defense systems have been discovered, such as DRUANTIA, GABIJA, and ZORYA (24). All of these defense systems not only prevent the introduction of heterologous DNA from plasmids or viruses, but also are widely applied in multiple fields, such as ABI system and RM system to avoid phage contamination in the fermentation industry (23,25,26), CRISPR-Cas system in precise genetic editing in biochemistry (27,28), TA system in picking cloning and living bacterial cellular single protein expression (29).

Several databases have been developed to integrate different defense systems. CRISPRdb and CRISPRone collect data of spacers and repeats, provide tools to search and display CRISPR-associated genes (30,31); REBASE is centered on RM system about restriction enzymes, methylases, and methylation specificity (32); TADB integrates information of type 2 toxin-antitoxin loci and genetic features and provides similarity search, genome context browse, and phylogenetic tools (33). However, all the databases or plat-

[*]To whom correspondence should be addressed. Tel: +86 10 84097443; Fax: +86 10 84097720; Email: xiaojingfa@big.ac.cn
Correspondence may also be addressed to Zhewen Zhang. Tel: +86 10 84097756; Fax: +86 10 84097720; Email: zhangzw@big.ac.cn

forms mentioned above are only focused on a single defense system or subtype. Confronting the ever-increasing prokaryotic genomic data and the fast-emerging newfound defense systems, an integrated database embedding an in-depth analysis platform for multiple defense systems is an urgent need. To fill this gap, here we present PADS Arsenal, a comprehensive database of prokaryotic defense systems related genes. With a large collection of prokaryotic genomic data from public databases, PADS Arsenal is dedicated to gathering, storing, analyzing and visualizing prokaryotic defense system gene datasets over 33 000 species.

DATABASE IMPLEMENTATION

In terms of data collecting, all prokaryotic genomic data were retrieved from NCBI (ftp://ftp.ncbi.nlm.nih.gov/genomes/all/) (34). For the identification of defense systems related genes, we first extracted some defense systems related genes as seed sequences from literature curation (12,13,19,24). In order to expand the seed dataset, we also downloaded protein families/sequences from COG (35), Pfam (36), REBASE (32), TIGRFAMs (37) and TADB (33) databases. Second, PSI-BLAST (38) was adopted to homology search of defense systems related genes. Sequences with identity value ≥30% were selected as putative defense systems related genes for further analyses (39). Third, all putative defense systems related genes were confirmed by checking conserved domains within the defense genes using InterProScan (40). In addition, we also randomly selected some strains from eight species (*Pseudomonas aeruginosa, Bacillus cytotoxicus, Listeria ivanovii, Listeria monocytogenes, Neisseria meningitides, Streptococcus pyogenes, Escherichia coli* and *Mycoplasma pneumoniae*) in PADS Arsenal for quality control. The identified CRISPR-Cas systems related genes in these strains were compared to the results of a well-known CRISPR-Cas systems identification tool CRISPRCasFinder (41). About 96% cas genes detected by CRISPRCasFinder were archived in PADS Arsenal. The reason for a small amount of gene missing was the slightly lower coverage of our seed datasets. We will integrate more seed sequences in the next version of PADS Arsenal for higher defense genes detection rate. Prokka was employed for genome annotation (42), Roary was applied for defense system gene orthologous clustering (43), ComplexHeatmap was used to construct the heatmap of defense system gene (44), and MAFFT was utilized for multiple sequences alignment (45). As for database construction, we used PHP, HTML5, CSS, Bootstrap, JQuery for front-end rendering and implementation of interactive events. Echarts, D3, circosJs, MSAviewer (46), phylotree.js (47) were adopted for building interactive graphs. DataTables and Bootstrap Table were used to render data tables. On the back-end, MySQL was employed to store data, and finally bioinformatics applications were achieved with PSI-BLAST (38), MAFFT (45), PhyML (48) and Python.

DATABASE CONTENT AND USAGE

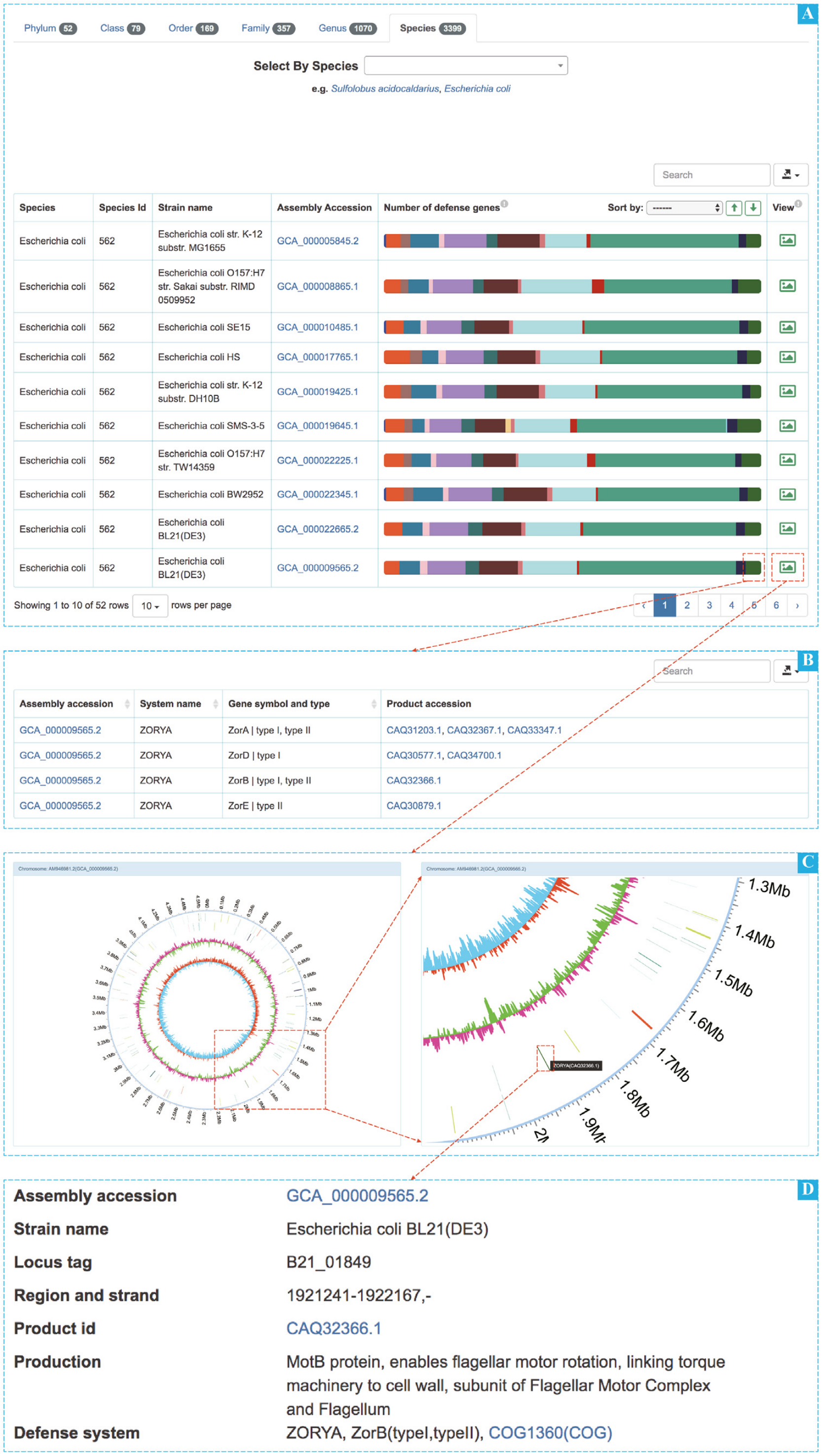
In current version 1.0, we have annotated 6 600 264 genes from 18 distinctive categories of defense systems. These genes were retrieved from 63 701 genomes, a total of 33 390 species across archaea and bacteria (Table 1). PADS Arsenal not only provides a user-friendly interface but also a rich analysis function, which offers flexible ways to retrieve and present a dynamic interactive defense systems related genes annotation pipeline.

In the browse module, all the completed prokaryotic genomes can be visualized by different taxonomic hierarchies. Users can select a taxonomy label and type some characters in the input box and click the corresponding taxonomic group. For example, searching for *E. coli* (Figure 1), a table of all related strains with a color bar will show up. Users can intuitively observe the composition of defense systems related genes and their corresponding strains and the composition variations of defense systems related genes between different strains. Each colored block can be clicked to show the details of all genes in that defense system. In addition, the last thumbnail click is used to display the information of the locus of the defense systems related genes, GC content and GC skew value of the genome by Circos graph. Strips of different colors represent different types of defense systems related genes, and each strip can be clicked for further information. Users can estimate the regions of defense island by combining all information of the arrangement of defense systems related genes across the genome, GC skew value, and the difference in GC content compared to the average of the genome.

To better search and explore the database, we provide four searching approaches (Figure 2). System-based and gene-based approaches can be applied when users are interested in a certain system or gene in a defense system, respectively. Species-based and assembly accession-based searches are also provided when users look for a species or an assembly accession ID. The results collected by the four searching approaches are identical, such as defense system category, defense system subtype, and gene symbol.

An interactive online pipeline of defense systems related gene annotation is integrated in the analysis module, combining the function of sequence homology search, multiple sequence alignment, and phylogenetic analysis. Users can upload a protein sequence for sequence similarity search. The targeting sequences will be further filtered by blast identity value and users can select seed sequences of interest for multiple sequence alignment. Users can also construct a phylogenetic tree to further annotate their uploaded sequence. For instance (Figure 3), we present the example sequence of DND and BREX systems and show the related results of homologous sequences search, multiple sequence alignment, and phylogenetic analysis in return.

Gene conservation is an important character for understanding the mechanism of defense system. To visualize the dynamic variation of defense systems related genes across species, a static presence-absence variation (PAV) analysis function is integrated in PADS Arsenal. In PAV analysis, users can select a species of interest to view the heatmap of PAV analysis result, by which users will choose a defense system to view the dynamic variation of defense systems related genes at the species-level from the insight of pan-genome. All defense system gene families (core, shared, unique) are listed in a table. For example, the results of searched *Chlamydia muridarum* and selected DISARM de-



Species	Species Id	Strain name	Assembly Accession	Number of defense genes	View
Escherichia coli	562	Escherichia coli str. K-12 substr. MG1655	GCA_000005845.2		
Escherichia coli	562	Escherichia coli O157:H7 str. Sakai substr. RIMD 0509952	GCA_000008865.1		
Escherichia coli	562	Escherichia coli SE15	GCA_000010485.1		
Escherichia coli	562	Escherichia coli HS	GCA_000017765.1		
Escherichia coli	562	Escherichia coli str. K-12 substr. DH10B	GCA_000019425.1		
Escherichia coli	562	Escherichia coli SMS-3-5	GCA_000019645.1		
Escherichia coli	562	Escherichia coli O157:H7 str. TW14359	GCA_000022225.1		
Escherichia coli	562	Escherichia coli BW2952	GCA_000022345.1		
Escherichia coli	562	Escherichia coli BL21(DE3)	GCA_000022665.2		
Escherichia coli	562	Escherichia coli BL21(DE3)	GCA_000009565.2		

Showing 1 to 10 of 52 rows 10 rows per page

Assembly accession	System name	Gene symbol and type	Product accession
GCA_000009565.2	ZORYA	ZorA | type I, type II	CAQ31203.1, CAQ32367.1, CAQ33347.1
GCA_000009565.2	ZORYA	ZorD | type I	CAQ30577.1, CAQ34700.1
GCA_000009565.2	ZORYA	ZorB | type I, type II	CAQ32366.1
GCA_000009565.2	ZORYA	ZorE | type II	CAQ30879.1

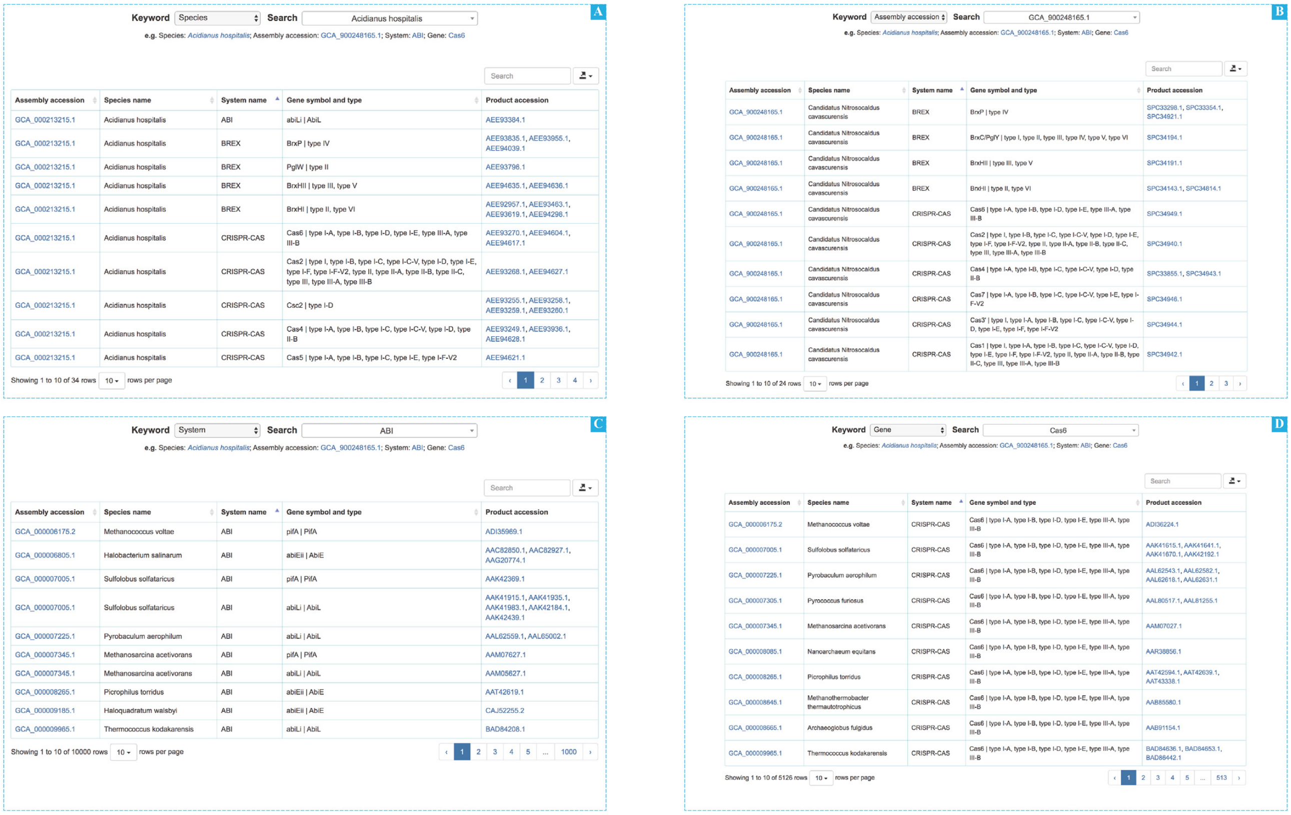
C
Chromosome: AM946981.2(GCA_000009565.2)
Chromosome: AM946981.2(GCA_000009565.2)
1.3Mb
1.4Mb
1.5Mb
1.6Mb
1.7Mb
1.8Mb
1.9Mb
ZORYA(CAQ32366.1)
D

Assembly accession	GCA_000009565.2
Strain name	Escherichia coli BL21(DE3)
Locus tag	B21_01849
Region and strand	1921241-1922167,-
Product id	CAQ32366.1
Production	MotB protein, enables flagellar motor rotation, linking torque machinery to cell wall, subunit of Flagellar Motor Complex and Flagellum
Defense system	ZORYA, ZorB(typeI,typeII), COG1360(COG)

Figure 1. Screenshots of browse page. (**A**) The *E. coli* search table based on species label at the browse page. (**B**) The ZORYA defense system gene table of *E. coli* BL21(DE3) by clicking the colored block. (**C**) Circos graphs of *E. coli* BL21(DE3) by clicking the shortcut link. (**D**) The detail information about a ZORYA defense system gene by clicking the strip (only partially shown).

Table 1. The statistics of annotated genes of each defense system in PADS Arsenal

Defense system	Archaea (1043 species)	Bacteria (32 347 species)
Abortive infection/phage exclusion systems (ABI)	909	70 595
Bacteriophage Exclusion (BREX)	9648	465 752
Clustered regularly interspaced short palindromic repeats with *cas* genes (CRISPR-CAS)	10 836	143 345
Defence island system associated with restriction–modification (DISARM)	6461	414 500
DNA phosphorothioation (DND)	1866	99 150
DRUANTIA	8642	726 041
GABIJA	2728	321 625
HACHIMAN	9675	713 810
KIWA	56	4780
LAMASSU	1026	127 721
Prokaryotic Argonautes (PAGOS)	90	1539
Restriction-Modification (RM)	13 276	1 016 565
SEPTU	990	151 706
SHEDU	16	2658
Toxin–Antitoxin (TA)	19 997	1 227 160
THOERIS	663	45 056
WADJET	2311	98 485
ZORYA	7456	873 130



A

Keyword Species Search Acidianus hospitalis

e.g. Species: *Acidianus hospitalis*; Assembly accession: GCA_900248165.1; System: ABI; Gene: Cas6

Assembly accession	Species name	System name	Gene symbol and type	Product accession
GCA_000213215.1	Acidianus hospitalis	ABI	abiLi | AbiL	AEE93384.1
GCA_000213215.1	Acidianus hospitalis	BREX	BrxP | type IV	AEE93835.1, AEE93955.1, AEE94039.1
GCA_000213215.1	Acidianus hospitalis	BREX	PglW | type II	AEE93796.1
GCA_000213215.1	Acidianus hospitalis	BREX	BrxHII | type III, type V	AEE94635.1, AEE94636.1
GCA_000213215.1	Acidianus hospitalis	BREX	BrxHI | type II, type VI	AEE92957.1, AEE93463.1, AEE93619.1, AEE94298.1
GCA_000213215.1	Acidianus hospitalis	CRISPR-CAS	Cas6 | type I-A, type I-B, type I-D, type I-E, type III-A, type III-B	AEE93270.1, AEE94604.1, AEE94617.1
GCA_000213215.1	Acidianus hospitalis	CRISPR-CAS	Cas2 | type I, type I-B, type I-C, type I-C-V, type I-D, type I-E, type I-F, type I-F-V2, type II, type II-A, type II-B, type II-C, type III, type III-A, type III-B	AEE93268.1, AEE94627.1
GCA_000213215.1	Acidianus hospitalis	CRISPR-CAS	Csc2 | type I-D	AEE93255.1, AEE93258.1, AEE93259.1, AEE93260.1
GCA_000213215.1	Acidianus hospitalis	CRISPR-CAS	Cas4 | type I-A, type I-B, type I-C, type I-C-V, type I-D, type II-B	AEE93249.1, AEE93936.1, AEE94628.1
GCA_000213215.1	Acidianus hospitalis	CRISPR-CAS	Cas5 | type I-A, type I-B, type I-C, type I-E, type I-F-V2	AEE94621.1

Showing 1 to 10 of 34 rows 10 rows per page ‹ 1 2 3 4 ›

B

Keyword Assembly accession Search GCA_900248165.1

e.g. Species: *Acidianus hospitalis*; Assembly accession: GCA_900248165.1; System: ABI; Gene: Cas6

Assembly accession	Species name	System name	Gene symbol and type	Product accession
GCA_900248165.1	Candidatus Nitrosocaldus cavascurensis	BREX	BrxP | type IV	SPC33298.1, SPC33354.1, SPC34921.1
GCA_900248165.1	Candidatus Nitrosocaldus cavascurensis	BREX	BrxC/PglY | type I, type II, type III, type IV, type V, type VI	SPC34194.1
GCA_900248165.1	Candidatus Nitrosocaldus cavascurensis	BREX	BrxHII | type III, type V	SPC34191.1
GCA_900248165.1	Candidatus Nitrosocaldus cavascurensis	BREX	BrxHI | type II, type VI	SPC34143.1, SPC34814.1
GCA_900248165.1	Candidatus Nitrosocaldus cavascurensis	CRISPR-CAS	Cas6 | type I-A, type I-B, type I-D, type I-E, type III-A, type III-B	SPC34949.1
GCA_900248165.1	Candidatus Nitrosocaldus cavascurensis	CRISPR-CAS	Cas2 | type I, type I-B, type I-C, type I-C-V, type I-D, type I-E, type I-F, type I-F-V2, type II, type II-A, type II-B, type II-C, type III, type III-A, type III-B	SPC34940.1
GCA_900248165.1	Candidatus Nitrosocaldus cavascurensis	CRISPR-CAS	Cas4 | type I-A, type I-B, type I-C, type I-C-V, type I-D, type II-B	SPC33855.1, SPC34943.1
GCA_900248165.1	Candidatus Nitrosocaldus cavascurensis	CRISPR-CAS	Cas7 | type I-A, type I-B, type I-C, type I-C-V, type I-E, type I-F-V2	SPC34946.1
GCA_900248165.1	Candidatus Nitrosocaldus cavascurensis	CRISPR-CAS	Cas3' | type I, type I-A, type I-B, type I-C, type I-C-V, type I-D, type I-E, type I-F, type I-F-V2	SPC34944.1
GCA_900248165.1	Candidatus Nitrosocaldus cavascurensis	CRISPR-CAS	Cas1 | type I, type I-A, type I-B, type I-C, type I-C-V, type I-D, type I-E, type I-F, type I-F-V2, type II, type II-A, type II-B, type II-C, type III, type III-A, type III-B	SPC34942.1

Showing 1 to 10 of 24 rows 10 rows per page ‹ 1 2 3 ›

C

Keyword System Search ABI

e.g. Species: *Acidianus hospitalis*; Assembly accession: GCA_900248165.1; System: ABI; Gene: Cas6

Assembly accession	Species name	System name	Gene symbol and type	Product accession
GCA_000006175.2	Methanococcus voltae	ABI	pifA | PifA	ADI35969.1
GCA_000006805.1	Halobacterium salinarum	ABI	abiEii | AbiE	AAC82850.1, AAC82927.1, AAG20774.1
GCA_000007005.1	Sulfolobus solfataricus	ABI	pifA | PifA	AAK42369.1
GCA_000007005.1	Sulfolobus solfataricus	ABI	abiLi | AbiL	AAK41915.1, AAK41935.1, AAK41983.1, AAK42184.1, AAK42439.1
GCA_000007225.1	Pyrobaculum aerophilum	ABI	abiLi | AbiL	AAL62559.1, AAL65002.1
GCA_000007345.1	Methanosarcina acetivorans	ABI	pifA | PifA	AAM07627.1
GCA_000007345.1	Methanosarcina acetivorans	ABI	abiLi | AbiL	AAM05627.1
GCA_000008265.1	Picrophilus torridus	ABI	abiEii | AbiE	AAT42619.1
GCA_000009185.1	Haloquadratum walsbyi	ABI	abiEii | AbiE	CAJ52255.2
GCA_000009965.1	Thermococcus kodakarensis	ABI	abiLi | AbiL	BAD84208.1

Showing 1 to 10 of 10000 rows 10 rows per page ‹ 1 2 3 4 5 ... 1000 ›

D

Keyword Gene Search Cas6

e.g. Species: *Acidianus hospitalis*; Assembly accession: GCA_900248165.1; System: ABI; Gene: Cas6

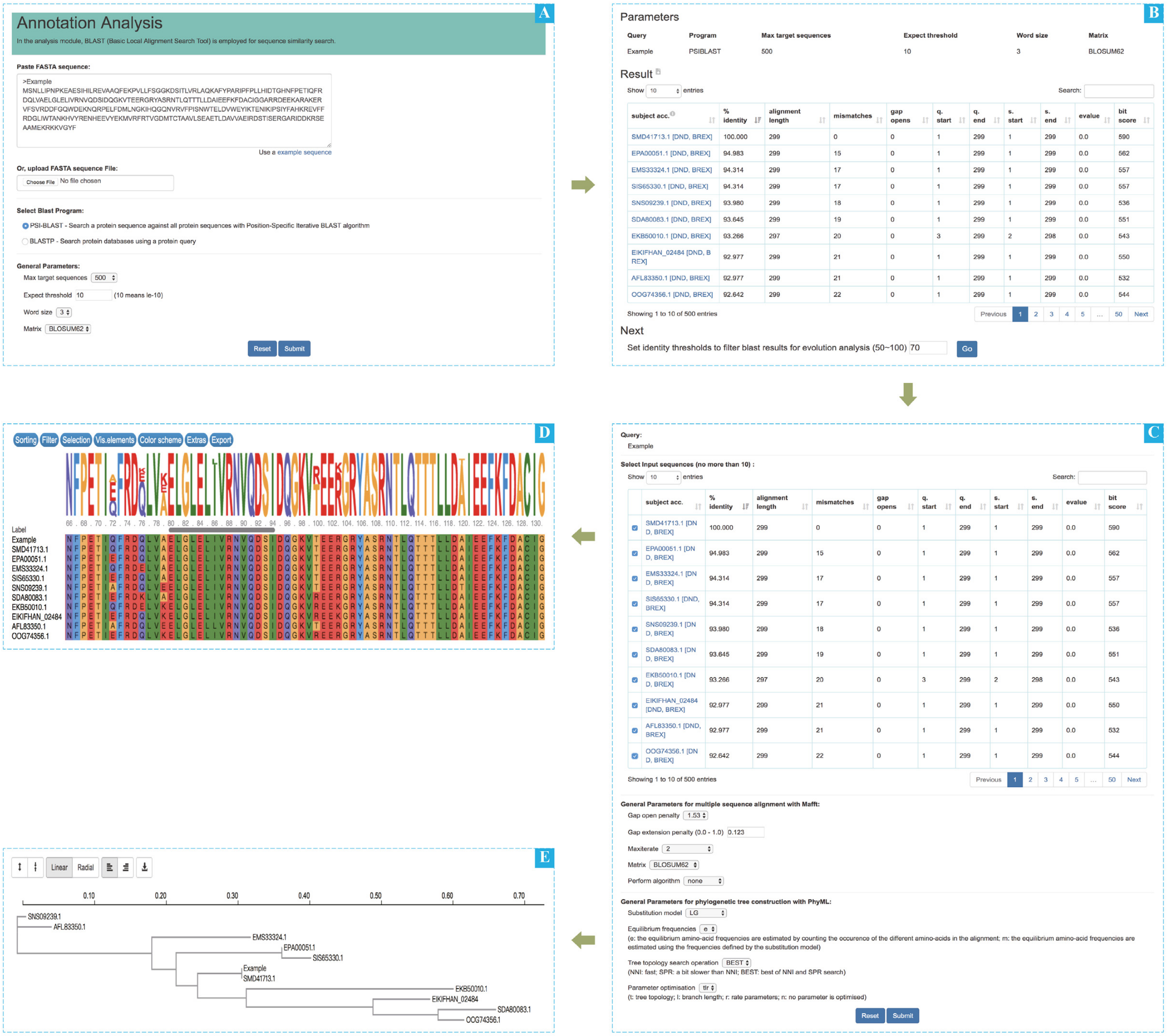
Assembly accession	Species name	System name	Gene symbol and type	Product accession
GCA_000006175.2	Methanococcus voltae	CRISPR-CAS	Cas6 | type I-A, type I-B, type I-D, type I-E, type III-A, type III-B	ADI36224.1
GCA_000007005.1	Sulfolobus solfataricus	CRISPR-CAS	Cas6 | type I-A, type I-B, type I-D, type I-E, type III-A, type III-B	AAK41615.1, AAK41641.1, AAK41670.1, AAK42192.1
GCA_000007225.1	Pyrobaculum aerophilum	CRISPR-CAS	Cas6 | type I-A, type I-B, type I-D, type I-E, type III-A, type III-B	AAL62543.1, AAL62582.1, AAL62618.1, AAL62631.1
GCA_000007305.1	Pyrococcus furiosus	CRISPR-CAS	Cas6 | type I-A, type I-B, type I-D, type I-E, type III-A, type III-B	AAL80517.1, AAL81255.1
GCA_000007345.1	Methanosarcina acetivorans	CRISPR-CAS	Cas6 | type I-A, type I-B, type I-D, type I-E, type III-A, type III-B	AAM07027.1
GCA_000008085.1	Nanoarchaeum equitans	CRISPR-CAS	Cas6 | type I-A, type I-B, type I-D, type I-E, type III-A, type III-B	AAR38856.1
GCA_000008265.1	Picrophilus torridus	CRISPR-CAS	Cas6 | type I-A, type I-B, type I-D, type I-E, type III-A, type III-B	AAT42594.1, AAT42639.1, AAT43338.1
GCA_000008645.1	Methanothermobacter thermautotrophicus	CRISPR-CAS	Cas6 | type I-A, type I-B, type I-D, type I-E, type III-A, type III-B	AAB85580.1
GCA_000008665.1	Archaeoglobus fulgidus	CRISPR-CAS	Cas6 | type I-A, type I-B, type I-D, type I-E, type III-A, type III-B	AAB91154.1
GCA_000009965.1	Thermococcus kodakarensis	CRISPR-CAS	Cas6 | type I-A, type I-B, type I-D, type I-E, type III-A, type III-B	BAD84636.1, BAD84653.1, BAD86442.1

Showing 1 to 10 of 5126 rows 10 rows per page ‹ 1 2 3 4 5 ... 513 ›

Figure 2. Screenshots of search page. (**A**) Species-based search results with *Acidianus hospitalis*. (**B**) Assembly accession-based search results for 'GCA_900248165.1'. (**C**) System-based search for ABI defense system. (**D**) Gene-based search for the *cas6* gene of CRISPR–Cas system.

fense system are shown in Figure 4. For further interpretation, the heatmap of *C. muridarum* suggests that genes associated with DISARM system are highly conserved. In addition, the orthologous clustering of defense system genes identified in PAV analysis also paves a way for downstream analyses.

In the statistic module, interactive charts are provided (Supplementary Figure S1). Users can get the overall distribution of defense systems related genes in archaea and bacteria kingdom through two pie charts. In the histogram, two browsing modes (single/multiple) are provided based on multiple taxonomic hierarchies (from phylum to genus).



Annotation Analysis

In the analysis module, BLAST (Basic Local Alignment Search Tool) is employed for sequence similarity search.

Paste FASTA sequence:

```
>Example
MSNLLIPNPKEAESIHILREVAAQFEKPVLLFSGGKDSITLVRLAQKAFYPARIPFPLLHIDTGHNFPETIQFR
DQLVAELGLELIVRNVQDSIDQGKVTEERGRYASRNTLQTTTLLDAIEEFKFDACIGGARRDEEKARAKER
VFSVRDDFGQWDEKNQRPELFDMLNGKIHQGQNVRVFPISNWTELDVWEYIKTENIKIPSIYFAHKREVFF
RDGLIWTANKHVYRENHEEVYEKMVRFRTVGDMTCTAAVLSEAETLDAVVAEIRDSTISERGARIDDKRSE
AAMEKRKKVGYF
```

Use a example sequence

Or, upload FASTA sequence File:

Choose File No file chosen

Select Blast Program:

● PSI-BLAST - Search a protein sequence against all protein sequences with Position-Specific Iterative BLAST algorithm

○ BLASTP - Search protein databases using a protein query

General Parameters:

Max target sequences 500

Expect threshold 10 (10 means le-10)

Word size 3

Matrix BLOSUM62

Reset Submit

Parameters

Query	Program	Max target sequences	Expect threshold	Word size	Matrix
Example	PSIBLAST	500	10	3	BLOSUM62

Result

Show 10 entries Search:

subject acc.	% identity	alignment length	mismatches	gap opens	q. start	q. end	s. start	s. end	evalue	bit score
SMD41713.1 [DND, BREX]	100.000	299	0	0	1	299	1	299	0.0	590
EPA00051.1 [DND, BREX]	94.983	299	15	0	1	299	1	299	0.0	562
EMS33324.1 [DND, BREX]	94.314	299	17	0	1	299	1	299	0.0	557
SIS65330.1 [DND, BREX]	94.314	299	17	0	1	299	1	299	0.0	557
SNS09239.1 [DND, BREX]	93.980	299	18	0	1	299	1	299	0.0	536
SDA80083.1 [DND, BREX]	93.645	299	19	0	1	299	1	299	0.0	551
EKB50010.1 [DND, BREX]	93.266	297	20	0	3	299	2	298	0.0	543
EIKIFHAN_02484 [DND, BREX]	92.977	299	21	0	1	299	1	299	0.0	550
AFL83350.1 [DND, BREX]	92.977	299	21	0	1	299	1	299	0.0	532
OOG74356.1 [DND, BREX]	92.642	299	22	0	1	299	1	299	0.0	544

Showing 1 to 10 of 500 entries Previous 1 2 3 4 5 … 50 Next

Next

Set identity thresholds to filter blast results for evolution analysis (50~100) 70 Go

Query:

Example

Select Input sequences (no more than 10) :

Show 10 entries Search:

	subject acc.	% identity	alignment length	mismatches	gap opens	q. start	q. end	s. start	s. end	evalue	bit score
☑	SMD41713.1 [DND, BREX]	100.000	299	0	0	1	299	1	299	0.0	590
☑	EPA00051.1 [DND, BREX]	94.983	299	15	0	1	299	1	299	0.0	562
☑	EMS33324.1 [DND, BREX]	94.314	299	17	0	1	299	1	299	0.0	557
☑	SIS65330.1 [DND, BREX]	94.314	299	17	0	1	299	1	299	0.0	557
☑	SNS09239.1 [DND, BREX]	93.980	299	18	0	1	299	1	299	0.0	536
☑	SDA80083.1 [DND, BREX]	93.645	299	19	0	1	299	1	299	0.0	551
☑	EKB50010.1 [DND, BREX]	93.266	297	20	0	3	299	2	298	0.0	543
☑	EIKIFHAN_02484 [DND, BREX]	92.977	299	21	0	1	299	1	299	0.0	550
☑	AFL83350.1 [DND, BREX]	92.977	299	21	0	1	299	1	299	0.0	532
☑	OOG74356.1 [DND, BREX]	92.642	299	22	0	1	299	1	299	0.0	544

Showing 1 to 10 of 500 entries Previous 1 2 3 4 5 … 50 Next

General Parameters for multiple sequence alignment with Mafft:

Gap open penalty 1.53

Gap extension penalty (0.0 - 1.0) 0.123

Maxiterate 2

Matrix BLOSUM62

Perform algorithm none

General Parameters for phylogenetic tree construction with PhyML:

Substitution model LG

Equilibrium frequencies e
(e: the equilibrium amino-acid frequencies are estimated by counting the occurence of the different amino-acids in the alignment; m: the equilibrium amino-acid frequencies are estimated using the frequencies defined by the substitution model)

Tree topology search operation BEST
(NNI: fast; SPR: a bit slower than NNI; BEST: best of NNI and SPR search)

Parameter optimisation tlr
(t: tree topology; l: branch length; r: rate parameters; n: no parameter is optimised)

Reset Submit

```
Label
Example         NFPETIQFRDQLVAELGLELIVRNVQDSIDQGKVTEERGRYASRNTLQTTTLLDAIEEFKFDACIG
SMD41713.1      NFPETIQFRDQLVAELGLELIVRNVQDSIDQGKVTEERGRYASRNTLQTTTLLDAIEEFKFDACIG
EPA00051.1      NFPETIEFRDQLVAELGLELIVRNVQDSIDQGKVTEERGRYASRNTLQTTTLLDAIEEFKFDACIG
EMS33324.1      NFPETIQFRDELVAELGLELIVRNVQDSIDQGKVTEERGRYASRNTLQTTTLLDAIEEFKFDACIG
SIS65330.1      NFPETIEFRDQLVAELGLELIVRNVQDSIDQGKVTEERGRYASRNTLQTTTLLDAIEEFKFDACIG
SNS09239.1      NFPETIAFRDQLVEELGLELIVRNVQDSIDQGKVTEERGRYASRNTLQTTTLLDTIEEFKFDACIG
SDA80083.1      NFPETIEFRDKLVAELGLELIVRNVQDSIDQGKVREERGRYASRNTLQTTTLLDAIEEFKFDACIG
EKB50010.1      NFPETIQFRDELVKELGLELIVRNVQDSIDQGKVREEKGRYASRNTLQTTTLLDAIEEFKFDACIG
EIKIFHAN_02484  NFPETIEFRDQLVKELGLELIVRNVQDSIDQGKVREEKGRYASRNTLQTTTLLDAIEEFKFDACIG
AFL83350.1      NFPETIAFRDQLVEELGLELLVRNVQDSIDQGKVTEERGRYASRNTLQTTTLLDAIEEFKFDACIG
OOG74356.1      NFPETIEFRDQLVKELGLELIVRNVQDSIDQGKVREERGRYASRNTLQTTTLLDAIEEFKFDACIG
```

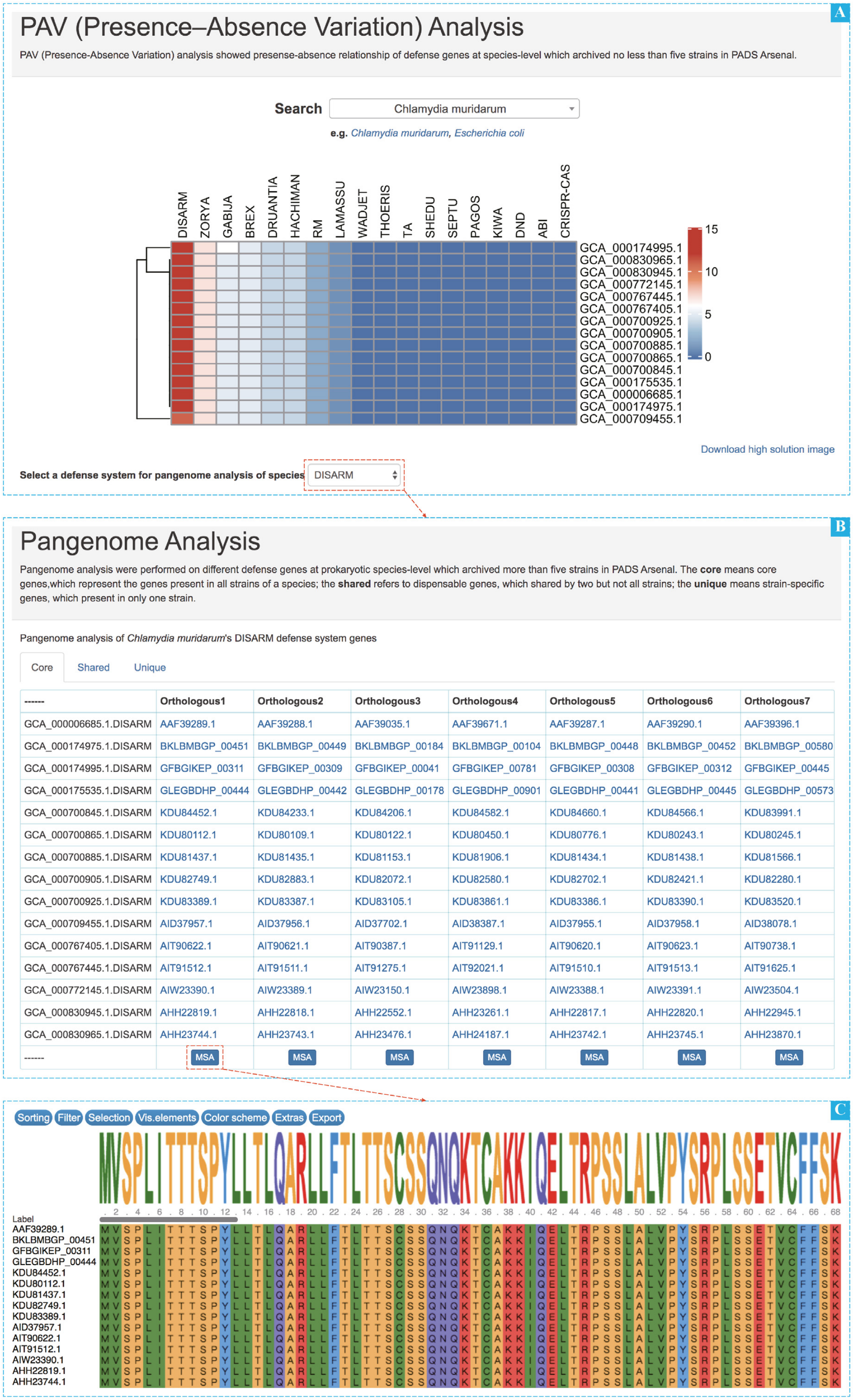
Figure 3. Screenshots of annotation page. (**A**) The upload of a sequence, the program selection and the parameters settings. (**B**) The preliminary results of the annotation and settings the filtering threshold. (**C**) Selected filtered results based on the threshold and parameters for multiple sequence alignment and building an evolutionary tree. (**D**) The result of multiple sequence alignment (only partially shown due to limited space). (**E**) The constructed evolutionary tree.

Users can recognize the presence-absence condition of different defense systems related genes at different taxonomic hierarchies by dynamic histograms. For instance, ZORYA defense systems related genes are widespread in phyla under archaea, while Abi genes are more specific and only observed in some archaeal genera. In addition, our statistics results for four species *E. coli, S. enterica, S. pyogenes* and *M. pneumoniae* show that some defense systems (TA, RM and ZORYA) might include different numbers of defense genes in different strains from the same species (Supplementary Figure S2). However, defense genes numbers in GABIJA, LAMASSU and WADJET defense systems are relatively stable.

All the processed results for these 6 600 264 defense systems related genes are publicly available at the download section. Besides, we also provide the data tables retrieved from the browse page and the search page, as well as the results of PAV analysis and online annotation.

FUTURE DIRECTIONS

Over the last several decades, defense systems related genes have been served as important editing, engineering and regulation tools due to their natural and powerful enzymatic activities, and the development of these tools has gone through two generations to date (6). RM enzymes



PAV (Presence–Absence Variation) Analysis

PAV (Presence-Absence Variation) analysis showed presense-absence relationship of defense genes at species-level which archived no less than five strains in PADS Arsenal.

Search: Chlamydia muridarum

e.g. *Chlamydia muridarum*, *Escherichia coli*

Select a defense system for pangenome analysis of species: DISARM

Pangenome Analysis

Pangenome analysis were performed on different defense genes at prokaryotic species-level which archived more than five strains in PADS Arsenal. The **core** means core genes,which represent the genes present in all strains of a species; the **shared** refers to dispensable genes, which shared by two but not all strains; the **unique** means strain-specific genes, which present in only one strain.

Pangenome analysis of *Chlamydia muridarum*'s DISARM defense system genes

Core | Shared | Unique

------	Orthologous1	Orthologous2	Orthologous3	Orthologous4	Orthologous5	Orthologous6	Orthologous7
GCA_000006685.1.DISARM	AAF39289.1	AAF39288.1	AAF39035.1	AAF39671.1	AAF39287.1	AAF39290.1	AAF39396.1
GCA_000174975.1.DISARM	BKLBMBGP_00451	BKLBMBGP_00449	BKLBMBGP_00184	BKLBMBGP_00104	BKLBMBGP_00448	BKLBMBGP_00452	BKLBMBGP_00580
GCA_000174995.1.DISARM	GFBGIKEP_00311	GFBGIKEP_00309	GFBGIKEP_00041	GFBGIKEP_00781	GFBGIKEP_00308	GFBGIKEP_00312	GFBGIKEP_00445
GCA_000175535.1.DISARM	GLEGBDHP_00444	GLEGBDHP_00442	GLEGBDHP_00178	GLEGBDHP_00901	GLEGBDHP_00441	GLEGBDHP_00445	GLEGBDHP_00573
GCA_000700845.1.DISARM	KDU84452.1	KDU84233.1	KDU84206.1	KDU84582.1	KDU84660.1	KDU84566.1	KDU83991.1
GCA_000700865.1.DISARM	KDU80112.1	KDU80109.1	KDU80122.1	KDU80450.1	KDU80776.1	KDU80243.1	KDU80245.1
GCA_000700885.1.DISARM	KDU81437.1	KDU81435.1	KDU81153.1	KDU81906.1	KDU81434.1	KDU81438.1	KDU81566.1
GCA_000700905.1.DISARM	KDU82749.1	KDU82883.1	KDU82072.1	KDU82580.1	KDU82702.1	KDU82421.1	KDU82280.1
GCA_000700925.1.DISARM	KDU83389.1	KDU83387.1	KDU83105.1	KDU83861.1	KDU83386.1	KDU83390.1	KDU83520.1
GCA_000709455.1.DISARM	AID37957.1	AID37956.1	AID37702.1	AID38387.1	AID37955.1	AID37958.1	AID38078.1
GCA_000767405.1.DISARM	AIT90622.1	AIT90621.1	AIT90387.1	AIT91129.1	AIT90620.1	AIT90623.1	AIT90738.1
GCA_000767445.1.DISARM	AIT91512.1	AIT91511.1	AIT91275.1	AIT92021.1	AIT91510.1	AIT91513.1	AIT91625.1
GCA_000772145.1.DISARM	AIW23390.1	AIW23389.1	AIW23150.1	AIW23898.1	AIW23388.1	AIW23391.1	AIW23504.1
GCA_000830945.1.DISARM	AHH22819.1	AHH22818.1	AHH22552.1	AHH23261.1	AHH22817.1	AHH22820.1	AHH22945.1
GCA_000830965.1.DISARM	AHH23744.1	AHH23743.1	AHH23476.1	AHH24187.1	AHH23742.1	AHH23745.1	AHH23870.1
------	MSA	MSA	MSA	MSA	MSA	MSA	MSA

```
               MVSPLITTTSPYLLTLQARLLFTLTTSCSSQNQKTCAKKIQELTRPSSLALVPYSRPLSSETVCFFSK
               . 2 . 4 . 6 . 8 . 10 . 12 . 14 . 16 . 18 . 20 . 22 . 24 . 26 . 28 . 30 . 32 . 34 . 36 . 38 . 40 . 42 . 44 . 46 . 48 . 50 . 52 . 54 . 56 . 58 . 60 . 62 . 64 . 66 . 68
Label
AAF39289.1     MVSPLITTTSPYLLTLQARLLFTLTTSCSSQNQKTCAKKIQELTRPSSLALVPYSRPLSSETVCFFSK
BKLBMBGP_00451 MVSPLITTTSPYLLTLQARLLFTLTTSCSSQNQKTCAKKIQELTRPSSLALVPYSRPLSSETVCFFSK
GFBGIKEP_00311 MVSPLITTTSPYLLTLQARLLFTLTTSCSSQNQKTCAKKIQELTRPSSLALVPYSRPLSSETVCFFSK
GLEGBDHP_00444 MVSPLITTTSPYLLTLQARLLFTLTTSCSSQNQKTCAKKIQELTRPSSLALVPYSRPLSSETVCFFSK
KDU84452.1     MVSPLITTTSPYLLTLQARLLFTLTTSCSSQNQKTCAKKIQELTRPSSLALVPYSRPLSSETVCFFSK
KDU80112.1     MVSPLITTTSPYLLTLQARLLFTLTTSCSSQNQKTCAKKIQELTRPSSLALVPYSRPLSSETVCFFSK
KDU81437.1     MVSPLITTTSPYLLTLQARLLFTLTTSCSSQNQKTCAKKIQELTRPSSLALVPYSRPLSSETVCFFSK
KDU82749.1     MVSPLITTTSPYLLTLQARLLFTLTTSCSSQNQKTCAKKIQELTRPSSLALVPYSRPLSSETVCFFSK
KDU83389.1     MVSPLITTTSPYLLTLQARLLFTLTTSCSSQNQKTCAKKIQELTRPSSLALVPYSRPLSSETVCFFSK
AID37957.1     MVSPLITTTSPYLLTLQARLLFTLTTSCSSQNQKTCAKKIQELTRPSSLALVPYSRPLSSETVCFFSK
AIT90622.1     MVSPLITTTSPYLLTLQARLLFTLTTSCSSQNQKTCAKKIQELTRPSSLALVPYSRPLSSETVCFFSK
AIT91512.1     MVSPLITTTSPYLLTLQARLLFTLTTSCSSQNQKTCAKKIQELTRPSSLALVPYSRPLSSETVCFFSK
AIW23390.1     MVSPLITTTSPYLLTLQARLLFTLTTSCSSQNQKTCAKKIQELTRPSSLALVPYSRPLSSETVCFFSK
AHH22819.1     MVSPLITTTSPYLLTLQARLLFTLTTSCSSQNQKTCAKKIQELTRPSSLALVPYSRPLSSETVCFFSK
AHH23744.1     MVSPLITTTSPYLLTLQARLLFTLTTSCSSQNQKTCAKKIQELTRPSSLALVPYSRPLSSETVCFFSK
```

Figure 4. Screenshots of PAV analysis page. (**A**) The heatmap of defense system genes distribution for *C. muridarum*. (**B**) The detailed information of DISARM defense system orthologous gene clusters based on the heatmap. (**C**) The results of multiple sequence alignment of an orthologous gene cluster by clicking the 'MSA' button.

were used as key genetic engineering tools in the early stage (49–51). Recently, CRISPR–Cas systems have been widely used as genetic editing tools with its functional diversity, which includes versatile mechanisms of crRNA guide processing, self/non-self discrimination, and target cleavage (48). Moreover, prokaryotic Argonaute proteins have been reported to mediate nucleic acid-guided cleavage of cognate DNA targets (52,53) or RNA targets (54,55) *in vitro*. This might lead to a new generation of genome-editing tools (56,57). In this study, we construct PADS Arsenal in a wide variety of application, including displaying defense systems related genes in a complete genome-scale at different taxonomic hierarchies, searching defense systems related genes, annotating and analyzing specific sequences with multiple tools and depicting dynamic variation of defense systems related genes across species. PADS Arsenal archives defense systems related genes rather than indicating complete defense systems. This is mainly because there are no definite descriptions of complete system or active defense system for some multiple gene systems (more than three genes in a system), such as DISARM, DND and Druantia. The integrity identification of all the 18 defense systems or their subtypes is a great challenge and it is also the future development direction for PADS Arsenal. In current version, PADS Arsenal will help users to detect potential defense systems related genes as engineering tools, but none of these systems can be functional if they are not complete. For defense systems integrity, we count the number of strains with or without complete systems, the results presented that RM and TA defense systems are complete in all analyzed strains of *E. coli, S. enterica* and *S. pyogenes* (Supplementary Figure S3). This implies that the complete RM and TA defense systems might be essential for these species. However, the integrity of HACHIMAN, KIWA and SEPTU defense systems shows dynamic changes in different strains of the same species (*E. coli* and *S. enterica*). Some recent studies indicate that the defense genes are the most evolutionarily dynamic functional class of genes and the gene loss is about three times more than gene gain (57,58).

There will be many new defense systems that have yet to be discovered (2,5). In future, PADS Arsenal, as one of the important database resources in BIG Data Center (59), will continuously collect and organize more types of defense systems and prokaryotic genomic data. Defense islands, formed by many physically clustered genes that are involved in archaeal and bacterial defense functions, provide a shortcut for discovering new defense systems (4,6,60). We will develop and integrate novel prediction methods to facilitate the identification of defense islands. In some defense systems, genomic modification plays a key role in self/non-self discrimination, for instance, in the BREX system, methylation on the fifth locus of non-palindromic TAGGAG motifs to guide self/non-self discrimination (13); and in the DISARM system, methylation on the second locus of CCWGG motifs as a marker of self DNA (12). And with a greater integration with motif and gene modification site information of self/non-self discrimination through literature curation and deep mining of genome modification information will be a welcome improvement.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Dr Jun Zhong and Members of the BIG Data Center for reporting bugs and sending comments. We also thank Dr Jiayan Wu for the proofreading.

FUNDING

National Key Research Program of China [2016YFB0201702 to J.X.]; National Natural Science Foundation of China [31771465 and 31970634 to J.X.]; Promoting Big Data Development Project, the National Development and Reform Commission of China [2016-999999-65-01-000696-07 to J.X.]; International Partnership Program of the Chinese Academy of Sciences [153F11KYSB20160008]; The 13th Five-year Informatization Plan of Chinese Academy of Sciences [XXH13505-05 to J.X.]. Funding for open access charge: National Key Research Program of China [2016YFB0201702].

Conflict of interest statement. None declared.

REFERENCES

1. Valen,L.V. (1973) A new evolutionary law. *Evol. Theory*, **1**, 1–30.
2. Stern,A. and Sorek,R. (2011) The phage-host arms race: shaping the evolution of microbes. *Bioessays*, **33**, 43–51.
3. Koonin,E.V. and Wolf,Y.I. (2012) Evolution of microbes and viruses: a paradigm shift in evolutionary biology? *Front. Cell Infect. Microbiol.*, **2**, 119.
4. Makarova,K.S., Wolf,Y.I., Snir,S. and Koonin,E.V. (2011) Defense islands in bacterial and archaeal genomes and prediction of novel defense systems. *J. Bacteriol.*, **193**, 6039–6056.
5. Makarova,K.S., Wolf,Y.I. and Koonin,E.V. (2013) Comparative genomics of defense systems in archaea and bacteria. *Nucleic Acids Res.*, **41**, 4360–4377.
6. Koonin,E.V., Makarova,K.S. and Wolf,Y.I. (2017) Evolutionary Genomics of Defense Systems in Archaea and Bacteria. *Annu. Rev. Microbiol.*, **71**, 233–261.
7. Arber,W. and Linn,S. (1969) DNA modification and restriction. *Annu. Rev. Biochem.*, **38**, 467–500.
8. Ershova,A.S., Rusinov,I.S., Spirin,S.A., Karyagina,A.S. and Alexeevski,A.V. (2015) Role of restriction-modification systems in prokaryotic evolution and ecology. *Biochemistry (Mosc.)*, **80**, 1373–1386.
9. Zhou,X., Deng,Z., Firmin,J.L., Hopwood,D.A. and Kieser,T. (1988) Site-specific degradation of *Streptomyces lividans* DNA during electrophoresis in buffers contaminated with ferrous iron. *Nucleic Acids Res.*, **16**, 4341–4352.
10. Wang,L., Chen,S., Xu,T., Taghizadeh,K., Wishnok,J.S., Zhou,X., You,D., Deng,Z. and Dedon,P.C. (2007) Phosphorothioation of DNA in bacteria by dnd genes. *Nat. Chem. Biol.*, **3**, 709–710.
11. Wang,L., Chen,S., Vergin,K.L., Giovannoni,S.J., Chan,S.W., DeMott,M.S., Taghizadeh,K., Cordero,O.X., Cutler,M., Timberlake,S. *et al.* (2011) DNA phosphorothioation is widespread and quantized in bacterial genomes. *Proc. Natl. Acad. Sci. U.S.A.*, **108**, 2963–2968.
12. Ofir,G., Melamed,S., Sberro,H., Mukamel,Z., Silverman,S., Yaakov,G., Doron,S. and Sorek,R. (2018) DISARM is a widespread bacterial defence system with broad anti-phage activities. *Nat. Microbiol.*, **3**, 90–98.
13. Goldfarb,T., Sberro,H., Weinstock,E., Cohen,O., Doron,S., Charpak-Amikam,Y., Afik,S., Ofir,G. and Sorek,R. (2015) BREX is a novel phage resistance system widespread in microbial genomes. *EMBO J.*, **34**, 169–183.

14. Hur,J.K., Olovnikov,I. and Aravin,A.A. (2014) Prokaryotic Argonautes defend genomes against invasive DNA. *Trends Biochem. Sci.*, **39**, 257–259.
15. Swarts,D.C., Makarova,K., Wang,Y., Nakanishi,K., Ketting,R.F., Koonin,E.V., Patel,D.J. and van der Oost,J. (2014) The evolutionary journey of Argonaute proteins. *Nat. Struct. Mol. Biol.*, **21**, 743–753.
16. van der Oost,J., Jore,M.M., Westra,E.R., Lundgren,M. and Brouns,S.J. (2009) CRISPR-based adaptive and heritable immunity in prokaryotes. *Trends Biochem. Sci.*, **34**, 401–407.
17. Garneau,J.E., Dupuis,M.E., Villion,M., Romero,D.A., Barrangou,R., Boyaval,P., Fremaux,C., Horvath,P., Magadan,A.H. and Moineau,S. (2010) The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. *Nature*, **468**, 67–71.
18. Horvath,P. and Barrangou,R. (2010) CRISPR/Cas, the immune system of bacteria and archaea. *Science*, **327**, 167–170.
19. Makarova,K.S., Haft,D.H., Barrangou,R., Brouns,S.J., Charpentier,E., Horvath,P., Moineau,S., Mojica,F.J., Wolf,Y.I., Yakunin,A.F. *et al.* (2011) Evolution and classification of the CRISPR-Cas systems. *Nat. Rev. Microbiol.*, **9**, 467–477.
20. Gerdes,K., Christensen,S.K. and Lobner-Olesen,A. (2005) Prokaryotic toxin-antitoxin stress response loci. *Nat. Rev. Microbiol.*, **3**, 371–382.
21. Yamaguchi,Y., Park,J.H. and Inouye,M. (2011) Toxin-antitoxin systems in bacteria and archaea. *Annu. Rev. Genet.*, **45**, 61–79.
22. Page,R. and Peti,W. (2016) Toxin-antitoxin systems in bacterial growth arrest and persistence. *Nat. Chem. Biol.*, **12**, 208–214.
23. Chopin,M.C., Chopin,A. and Bidnenko,E. (2005) Phage abortive infection in lactococci: variations on a theme. *Curr. Opin. Microbiol.*, **8**, 473–479.
24. Doron,S., Melamed,S., Ofir,G., Leavitt,A., Lopatina,A., Keren,M., Amitai,G. and Sorek,R. (2018) Systematic discovery of antiphage defense systems in the microbial pangenome. *Science*, **359**, eaar4120.
25. Allison,G.E. and Klaenhammer,T.R. (1998) Phage resistance mechanisms in lactic acid bacteria. *Int. Dairy J.*, **8**, 207–226.
26. Hansen,E.B. (2002) Commercial bacterial starter cultures for fermented foods of the future. *Int. J. Food Microbiol.*, **78**, 119–131.
27. Cong,L., Ran,F.A., Cox,D., Lin,S., Barretto,R., Habib,N., Hsu,P.D., Wu,X., Jiang,W., Marraffini,L.A. *et al.* (2013) Multiplex genome engineering using CRISPR/Cas systems. *Science*, **339**, 819–823.
28. Hwang,W.Y., Fu,Y., Reyon,D., Maeder,M.L., Tsai,S.Q., Sander,J.D., Peterson,R.T., Yeh,J.R. and Joung,J.K. (2013) Efficient genome editing in zebrafish using a CRISPR-Cas system. *Nat. Biotechnol.*, **31**, 227–229.
29. Unterholzner,S.J., Poppenberger,B. and Rozhon,W. (2013) Toxin-antitoxin systems: biology, identification, and application. *Mob Genet. Elements*, **3**, e26219.
30. 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.
31. Zhang,Q. and Ye,Y. (2017) Not all predicted CRISPR-Cas systems are equal: isolated cas genes and classes of CRISPR like elements. *BMC Bioinformatics*, **18**, 92.
32. Roberts,R.J., Vincze,T., Posfai,J. and Macelis,D. (2015) REBASE–a database for DNA restriction and modification: enzymes, genes and genomes. *Nucleic Acids Res.*, **43**, D298–D299.
33. Xie,Y., Wei,Y., Shen,Y., Li,X., Zhou,H., Tai,C., Deng,Z. and Ou,H.Y. (2018) TADB 2.0: an updated database of bacterial type II toxin-antitoxin loci. *Nucleic Acids Res.*, **46**, D749–D753.
34. Sayers,E.W., Agarwala,R., Bolton,E.E., Brister,J.R., Canese,K., Clark,K., Connor,R., Fiorini,N., Funk,K., Hefferon,T. *et al.* (2019) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **47**, D23–D28.
35. Galperin,M.Y., Makarova,K.S., Wolf,Y.I. and Koonin,E.V. (2015) Expanded microbial genome coverage and improved protein family annotation in the COG database. *Nucleic Acids Res.*, **43**, D261–D269.
36. El-Gebali,S., Mistry,J., Bateman,A., Eddy,S.R., Luciani,A., Potter,S.C., Qureshi,M., Richardson,L.J., Salazar,G.A., Smart,A. *et al.* (2019) The Pfam protein families database in 2019. *Nucleic Acids Res.*, **47**, D427–D432.
37. Haft,D.H., Loftus,B.J., Richardson,D.L., Yang,F., Eisen,J.A., Paulsen,I.T. and White,O. (2001) TIGRFAMs: a protein family resource for the functional identification of proteins. *Nucleic Acids Res.*, **29**, 41–43.
38. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
39. Kuzniar,A., van Ham,R.C., Pongor,S. and Leunissen,J.A. (2008) The quest for orthologs: finding the corresponding gene across genomes. *Trends Genet.: TIG*, **24**, 539–551.
40. Hunter,S., Jones,P., Mitchell,A., Apweiler,R., Attwood,T.K., Bateman,A., Bernard,T., Binns,D., Bork,P., Burge,S. *et al.* (2012) InterPro in 2011: new developments in the family and domain prediction database. *Nucleic Acids Res.*, **40**, D306–D312.
41. Couvin,D., Bernheim,A., Toffano-Nioche,C., Touchon,M., Michalik,J., Neron,B., Rocha,E.P.C., Vergnaud,G., Gautheret,D. and Pourcel,C. (2018) CRISPRCasFinder, an update of CRISRFinder, includes a portable version, enhanced performance and integrates search for Cas proteins. *Nucleic Acids Res.*, **46**, W246–W251.
42. Seemann,T. (2014) Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, **30**, 2068–2069.
43. Page,A.J., Cummins,C.A., Hunt,M., Wong,V.K., Reuter,S., Holden,M.T., Fookes,M., Falush,D., Keane,J.A. and Parkhill,J. (2015) Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics*, **31**, 3691–3693.
44. Gu,Z., Eils,R. and Schlesner,M. (2016) Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics*, **32**, 2847–2849.
45. Katoh,K., Misawa,K., Kuma,K. and Miyata,T. (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.*, **30**, 3059–3066.
46. Yachdav,G., Wilzbach,S., Rauscher,B., Sheridan,R., Sillitoe,I., Procter,J., Lewis,S.E., Rost,B. and Goldberg,T. (2016) MSAViewer: interactive JavaScript visualization of multiple sequence alignments. *Bioinformatics*, **32**, 3501–3503.
47. Shank,S.D., Weaver,S. and Kosakovsky Pond,S.L. (2018) phylotree.js - a JavaScript library for application development and interactive data visualization in phylogenetics. *BMC Bioinformatics*, **19**, 276.
48. Guindon,S., Dufayard,J.F., Lefort,V., Anisimova,M., Hordijk,W. and Gascuel,O. (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic Biol.*, **59**, 307–321.
49. Vasu,K. and Nagaraja,V. (2013) Diverse functions of restriction-modification systems in addition to cellular defense. *Microbiol. Mol. Biol. Rev.: MMBR*, **77**, 53–72.
50. Roberts,R.J. (1976) Restriction endonucleases. *CRC Crit. Rev. Biochem.*, **4**, 123–164.
51. Williams,R.J. (2003) Restriction endonucleases: classification, properties, and applications. *Mol. Biotechnol.*, **23**, 225–243.
52. Swarts,D.C., Hegge,J.W., Hinojo,I., Shiimori,M., Ellis,M.A., Dumrongkulraksa,J., Terns,R.M., Terns,M.P. and van der Oost,J. (2015) Argonaute of the archaeon *Pyrococcus furiosus* is a DNA-guided nuclease that targets cognate DNA. *Nucleic Acids Res.*, **43**, 5120–5129.
53. Swarts,D.C., Jore,M.M., Westra,E.R., Zhu,Y., Janssen,J.H., Snijders,A.P., Wang,Y., Patel,D.J., Berenguer,J., Brouns,S.J.J. *et al.* (2014) DNA-guided DNA interference by a prokaryotic Argonaute. *Nature*, **507**, 258–261.
54. Yuan,Y.R., Pei,Y., Ma,J.B., Kuryavyi,V., Zhadina,M., Meister,G., Chen,H.Y., Dauter,Z., Tuschl,T. and Patel,D.J. (2005) Crystal structure of A. aeolicus argonaute, a site-specific DNA-guided endoribonuclease, provides insights into RISC-mediated mRNA cleavage. *Mol. Cell*, **19**, 405–419.
55. Kaya,E., Doxzen,K.W., Knoll,K.R., Wilson,R.C., Strutt,S.C., Kranzusch,P.J. and Doudna,J.A. (2016) A bacterial Argonaute with noncanonical guide RNA specificity. *Proc. Natl. Acad. Sci. U.S.A.*, **113**, 4057–4062.
56. Hegge,J.W., Swarts,D.C. and van der Oost,J. (2018) Prokaryotic Argonaute proteins: novel genome-editing tools? *Nat. Rev. Microbiol.*, **16**, 5–11.

57. Puigbo,P., Makarova,K.S., Kristensen,D.M., Wolf,Y.I. and Koonin,E.V. (2017) Reconstruction of the evolution of microbial defense systems. *BMC Evol. Biol.*, **17**, 94.
58. Puigbo,P., Lobkovsky,A.E., Kristensen,D.M., Wolf,Y.I. and Koonin,E.V. (2014) Genomes in turmoil: quantification of genome dynamics in prokaryote supergenomes. *BMC Biol.*, **12**, 66.
59. Members of BIG Data Center (2019) Database Resources of the BIG Data Center in 2019. *Nucleic Acids Res.*, **47**, D8–D14.
60. Makarova,K.S., Wolf,Y.I., Forterre,P., Prangishvili,D., Krupovic,M. and Koonin,E.V. (2014) Dark matter in archaeal genomes: a rich source of novel mobile elements, defense systems and secretory complexes. *Extremophiles*, **18**, 877–893.