PADS Arsenal: a database of prokaryotic defense systems related genes

Yadong Zhang1,2,3,4, Zhewen Zhang1,2,3,*, Hao Zhang1,2,3,4, Yongbing Zhao5, Zaichao Zhang6 and Jingfa Xiao1,2,3,4,*

1National Genomics Data Center, Beijing 100101, China, 2BIG Data Center, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China, 3CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China, 4College of Life Sciences, University of Chinese Academy of Sciences, Beijing 100049, China, 5Department of Health Sciences Research, Mayo Clinic, Jacksonville, FL 32224, USA and 6Department 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

© The Author(s) 2019. Published by Oxford University Press on behalf of Nucleic Acids Research.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com
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, circoJs, MSAvieviewer (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-
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                    |
| GABIA                                              | 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                    |

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.

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). 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.
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 6,602,264 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
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.
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 31907634 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 Alexeevskii, 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) Phosphorothioylation 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 phosphorothioate 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., Alil, S., Ofir, G. and Sorek, R. (2015) BREX is a novel phage resistance system widespread in microbial genomes. EMBO J., 34, 169–183.
28. Hwang, W.Y., Fu, Y., Reyon, D., Maeder, M.L., Tsai, S.Q., Sander, J.D., Hansen, E.B. (2002) Commercial bacterial starter cultures for fermented foods of the future. BMC Bioinformatics, 3, 401–407.

26. Allison, G.E. and Klaenhammer, T.R. (1998) Phage resistance defense systems in the microbial pangenome. Trends Biochem. Sci., 23, 711–717.

25. 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.

24. Garneau, J.E., Dupuis, M.E., Villion, M., Romero, D.A., Barrangou, R., Boyaval, P., Fremeaux, B., Horvath, P., Magadan, A.H. and Moineau, S. (2010) The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. Nature, 468, 67–71.

23. Horvath, P. and Barrangou, R. (2010) CRISPR/Cas, the immune system of bacteria and archaea. Science, 327, 167–170.

22. 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.

19. Yamaguchi, Y., Park, J.H. and Inouye, M. (2011) Toxin-antitoxin systems in bacteria and archaea. Annu. Rev. Genet., 45, 61–97.

18. Page, R. and Peti, W. (2016) Toxin-antitoxin systems in bacterial growth arrest and persistence. Nat. Chem. Biol., 12, 208–214.

17. Chopin, M.C., Chopin, A. and Bidonko, E. (2005) Phage abortive infection in lactococci: variations on a theme. Curr. Opin. Microbiol., 8, 473–479.

16. Doron, S., Melamed, S., Ofir, G., Leavitt, A., Lopatina, A., Keren, M., Amioti, G. and Sorek, R. (2018) Systematic discovery of antiphage defense systems in the microbial pangenome. Science, 359, eaar4120.

15. Allison, G.E. and Klaenhammer, T.R. (1998) Phage resistance mechanisms in lactic acid bacteria. Int. Dairy J., 8, 207–226.

14. Hansel, E.B. (2002) Commercial bacterial starter cultures for fermented foods of the future. Int. J. Food Microbiol., 78, 119–131.

13. 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.

12. 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.

11. Unterholzner, S.J., Poppenberger, B. and Rzeho, W. (2013) Toxin-antitoxin systems: biology, identification, and application. Mob Genet. Elements, 3, c26219.

10. 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.

9. 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.

8. Roberts, R.J. (1976) Restriction endonucleases. Nature, 207, 258–261.

7. Williams, R.J. (2003) Restriction endonucleases: classification, properties, and applications. Mol. Biotechnol., 23, 225–243.

6. Swarts, D.C., Hegge, J.W., Hinojo, I., Shilmori, M., Ellis, M.A., Dumrongkulraksa, J., Terns, R.M., Terns, M.P. and van der Oost, J. (2015) CRISPR CasFinder, an update of CRISPRFinder, includes a portable version, enhanced performance and integrates search for Cas proteins. Nucleic Acids Res., 43, W246–W251.

5. Seemann, T. (2014) Prokka: rapid prokaryotic genome annotation. Bioinformatics, 30, 2060–2061.

4. Page, A.J., Cummins, C.A., Hunt, M., Wong, V.K., Reuter, S., Holden, M.T., Fooke, M., Falush, D., Keane, J.A. and Parkhill, J. (2015) Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics, 31, 3691–3693.

3. Gu, Z., Eils, R. and Schlesmer, M. (2016) Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics, 32, 2847–2849.

2. 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.
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.