Published online 25 May 2022 *Nucleic Acids Research, 2022, Vol. 50, Web Server issue* ***W541–W550***
https://doi.org/10.1093/nar/gkac400

PADLOC: a web server for the identification of antiviral defence systems in microbial genomes

Leighton J. Payne [1], Sean Meaden[2], Mario R. Mestre[3], Chris Palmer[4], Nicolás Toro [5], Peter C. Fineran [1,6,7,8] and Simon A. Jackson [1,6,7,8,*]

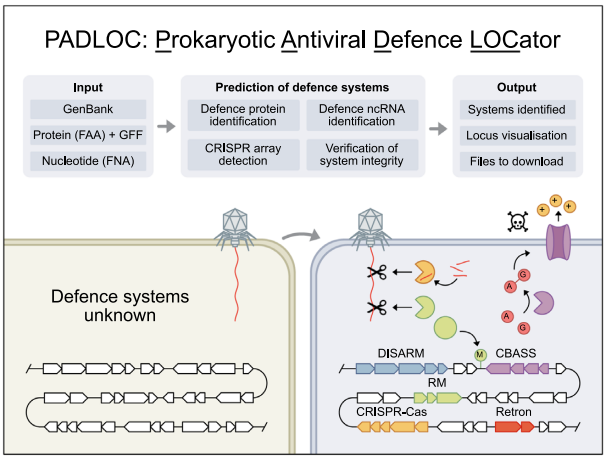
[1]Department of Microbiology and Immunology, University of Otago, Dunedin, New Zealand, [2]Biosciences, University of Exeter, Penryn, UK, [3]Independent Researcher, Spain, [4]Information Technology Services Research and Teaching Group, University of Otago, Dunedin, New Zealand, [5]Department of Soil Microbiology and Symbiotic Systems, Estación Experimental del Zaidín, Consejo Superior de Investigaciones Científicas, Structure, Dynamics and Function of Rhizobacterial Genomes, Grupo de Ecología Genética de la Rizosfera, Granada, Spain, [6]Genetics Otago, University of Otago, Dunedin, New Zealand, [7]Bioprotection Aotearoa, University of Otago, Dunedin, New Zealand and [8]Maurice Wilkins Centre for Molecular Biodiscovery, University of Otago, Dunedin, New Zealand

Received April 03, 2022; Revised April 24, 2022; Editorial Decision May 03, 2022; Accepted May 05, 2022

ABSTRACT

Most bacteria and archaea possess multiple antiviral defence systems that protect against infection by phages, archaeal viruses and mobile genetic elements. Our understanding of the diversity of defence systems has increased greatly in the last few years, and many more systems likely await discovery. To identify defence-related genes, we recently developed the Prokaryotic Antiviral Defence LOCator (PADLOC) bioinformatics tool. To increase the accessibility of PADLOC, we describe here the PADLOC web server (freely available at https://padloc.otago.ac.nz), allowing users to analyse whole genomes, metagenomic contigs, plasmids, phages and archaeal viruses. The web server includes a more than 5-fold increase in defence system types detected (since the first release) and expanded functionality enabling detection of CRISPR arrays and retron ncRNAs. Here, we provide user information such as input options, description of the multiple outputs, limitations and considerations for interpretation of the results, and guidance for subsequent analyses. The PADLOC web server also houses a precomputed database of the defence systems in > 230,000 RefSeq genomes. These data reveal two taxa, Campylobacterota and Spriochaetota, with unusual defence system diversity and abundance. Overall, the PADLOC web server provides a convenient and accessible resource for the detection of antiviral defence systems.

GRAPHICAL ABSTRACT



INTRODUCTION

Diverse antiviral defence systems have evolved in bacteria and archaea that defend against infection by their viruses and mobile genetic elements. There are over 60 known broad families of defence systems, with more than 20 distinct system types discovered in the past five years (precise tallies are difficult since classification schemes, such as class, type and subtype, vary between families of systems and the mechanisms of many systems remain unknown) (1–5). As such, system discovery has greatly outpaced the development of tools that make use of these new insights. To provide widespread accessibility to known and newly discovered system types, and ensure consistency between system annotations, we recently developed the Prokaryotic Antiviral Defence Locator (PADLOC) tool as a framework to systematically identify antiviral defence systems (6). To sim-

*To whom correspondence should be addressed. Tel: +64 3 479 8428; Email: simon.jackson@otago.ac.nz

plify the use of PADLOC, we have developed the PADLOC web server, which expands the functionality of PADLOC, serves as a convenient and accessible interface for using the tool, and provides an extensive database of precomputed results – currently for more than 230,000 bacterial and archaeal genomes.

Details regarding operation and benchmarking of the PADLOC tool itself have been described elsewhere (6). However, there are some important aspects of system detection that users of the PADLOC web server should understand when interpreting results. Genes encoding defence system proteins are identified by searching their protein sequences with a curated database of profile Hidden Markov Models (HMMs), currently representing > 700 families of defence-related proteins. Many of the protein families are represented by multiple HMMs (for example, there are currently 45 HMMs from various sources representing different Cas10 clades). Potential matches are filtered to remove low-scoring hits (based on E-value and coverage thresholds). PADLOC then uses a set of system definition models to determine whether the genetic synteny requirements are met for each possible system classification (Figure 1). This approach to multi-gene system identification has been applied successfully in the past for the detection of CRISPR-Cas (7,8) and protein secretion systems (9). In addition to detecting protein-coding genes, the PADLOC web server includes new functionality to detect CRISPR arrays and ncRNAs, such as retron msr-msd elements. Here, we discuss this added PADLOC functionality, the addition of many new systems to the database, limitations and important considerations for interpretation of the results, and guidance for subsequent analyses.

MATERIALS AND METHODS

Expansion of the PADLOC defence system database

At its core, the PADLOC web server is based on the PADLOC command line tool (https://github.com/padlocbio/padloc), with a curated database of profile HMMs, HMM scoring thresholds, and system models (https://github.com/padlocbio/padloc-db). Our current understanding of defence system genetics is the result of a collective scientific effort, and the HMMs and system models in the PADLOC database were built and curated using data from many sources. Construction of HMMs for the CBASS and Doron systems, plus several variant systems we discovered, are as previously described (6). We have since expanded the PADLOC web server database to contain > 180 system definitions, including > 3,500 profile HMMs. Where profile HMMs were made available by the authors of papers describing new defence system types (10–12), or databases and tools to detect subsets of systems (13–15), these HMMs were assigned PADLOC HMM accessions (e.g. PLDC12345) and added to the PADLOC database (the original HMM names were retained for traceability). Where multiple sequence alignments were available (16,17), we realigned the sequences using MUSCLE (18) and built HMMs with HMMER3 (19). For cases without HMMs or sequence alignments but a list of relevant proteins was available (8,20–39), we either used our sequence clustering and HMM generation pipeline described previously (6), or aligned the sequences with MUSCLE, manually curated the alignments to remove outlier sequences, then built HMMs with HMMER3. In the absence of supplied lists of homologs, we used example sequences from experimentally verified defence systems as seeds for BLAST searches, then aligned, curated and built HMMs, as above (22,40–69).

Where possible, the data source and appropriate reference for each HMM is listed in the PADLOC database HMM metadata file (hmm_meta.txt, available from the PADLOC database repository). We encourage PADLOC users to recognize the importance and value of these data used to build the PADLOC web server by citing the original sources. Models will continue to be added and updated periodically as more defence systems are discovered. We welcome and encourage submissions of new defence systems, including HMMs, multiple sequence alignments, or lists for relevant protein sequences or database accessions. Similarly, feedback to improve the sensitivity and specificity of PADLOC, updates to citation links, and suggestions to improve defence system and protein nomenclature will help ensure PADLOC remains a useful community resource.

Detection of non-coding sequences

Since the command line PADLOC tool detects only protein-coding genes, yet many defence systems contain non-coding RNAs, we integrated detection of CRISPR arrays and retron-associated ncRNAs (msr-msd elements) into the web server. CRISPR arrays are detected with a customized version of CRISPRDetect (70) using the arguments: array_quality_score_cutoff 2.5; minimum_word_repeatation 3; word_length 11; minimum_no_of_repeats 3; repeat_length_cutoff 11; max_gap_between_crisprs 250. The resulting crispr.gff output file is supplied to PADLOC using the –crispr input option and the human-readable crispr.txt output file is made available for user download. Potential ncRNAs associated with retrons are identified by searching a database of msr-msd element covariance models (available from the PADLOC database repository), against each genome sequence using Infernal's cmsearch (71) with the arguments: Z 10; FZ 500. The Infernal output is filtered to only include hits passing the inclusion threshold (E-value = 0.01), then loaded into PADLOC using the –ncrna input option. In specific PADLOC system definition models (e.g. for retrons), ncRNAs are listed as 'ncRNA' in the core, accessory or prohibited gene lists, as required. As such, any identified ncRNAs contribute to the total required gene count for each relevant defence system.

Precomputed RefSeq data and pseudogenes

For the precomputed PADLOC dataset (currently based on RefSeq v209 (72)), we used the [assembly]_genomic.fna, [assembly]_genomic.gff and [assembly]_protein.faa files for each genome assembly from the RefSeq FTP server. First, CRISPR arrays and retron-associated ncRNAs were identified by running CRISPRDetect and Infernal, respectively (as above), with [assembly]_genomic.fna as input. The resulting GFF-formatted CRISPRDetect outputs were saved for input to PADLOC and the more detailed, human-readable output files were loaded to the PADLOC web

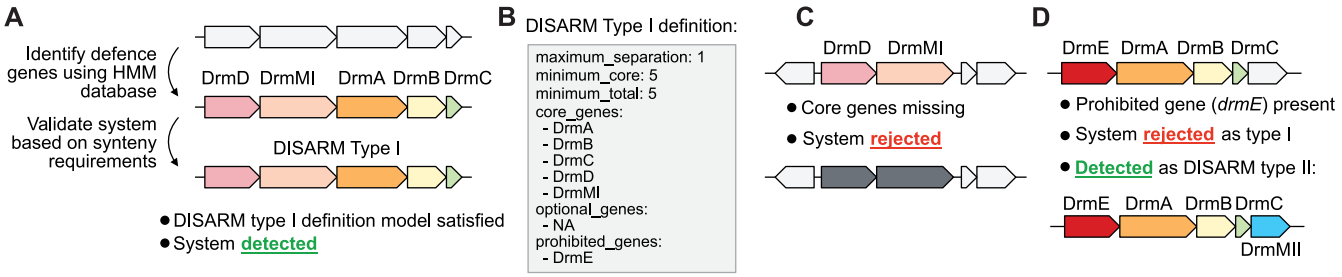


Figure 1. Defence system detection with PADLOC. (**A**) Genes encoding putative defence system proteins are identified using profile HMMs, then compared against the system definition models to determine whether a complete system is present. (**B**) For example, detection of a DISARM Type I system requires genes encoding all five core components (DrmA, DrmB, DrmC, DrmD, DrmMI) to be present. (**C**) If the minimum number of genes is not met, the system is not reported. (**D**) If any genes prohibited for a specific system definition are present (in the case of DISARM type I, *drmE* is prohibited), the system is rejected, but can instead be reported as a match to a different system definition—in this example as a type II DISARM (requiring genes encoding DrmA, DrmB, DrmC, DrmE and DrmMII).

server for user download. Next, we pre-processed the [assembly]_genomic.gff and [assembly]_protein.faa files to allow increased detection of defence systems containing pseudogenes. The NCBI prokaryotic genome annotation pipeline (PGAP) identifies genes that contain frameshifts, nonsense stop codons, or appear otherwise incomplete, as pseudogenes (73). The coordinates of each pseudogene are reported in the [assembly]_genomic.gff, but the corresponding protein sequences are not included in the [assembly]_protein.faa. Since PADLOC relies on any potential defence system protein sequences to be present in the input file, pseudogenes belonging to defence systems would not normally be identified. The PGAP annotates each pseudogene with the accession of the full protein sequence used to infer the product of the pseudogene (e.g. the *Bacillus cereus* VD146 assembly GCF_000399425.1 has a pseudogene with locus tag IK1_RS32735 that is labelled as similar to the CRISPR-associated protein Cas4 of *Oceanobacillus massiliensis* WP_010649895.1). Therefore, we substituted the pseudogenes in each RefSeq genome with the sequence of their inferential protein (where available). Lastly, PADLOC was run using the pseudogene-corrected .gff and .faa inputs, plus the CRISPR array and ncRNA inputs (as above).

Implementation

The core PADLOC tool is implemented in R, with some input handling using Bash and Python (primarily Biopython (74)). The PADLOC web server was built using the Django Framework (https://www.djangoproject.com). User jobs are identified by unique and anonymous job identifiers and are not accessible by other users. Users can access their results (tracked using cookies) until their browser cookies are cleared, the results are removed manually by the user, or until they expire (currently after 10 days). On the user side, anonymous job identifiers are replaced with the user-specific job name and output files downloaded are prefixed by the job name. When input files are uploaded to the server, a pre-processing script is used to detect the source format of the files and convert these to the default inputs for PADLOC (.gff and .faa with RefSeq formatting). For example, RAST formatted Genbank files are identified by the presence of the string 'rasttk', next the 'db_xref' field is changed to 'locus_tag', finally Biopython is used to output PADLOC-compatible input files. PADLOC also contains a '–fix-prodigal' option to natively parse Prodigal-formatted .gff and .faa file pairs, which the pre-processing script detects by searching for the string 'Prodigal' in the uploaded .gff file. Although the command line version of PADLOC includes a wrapper for gene-calling with Prodigal (allowing input of unannotated nucleotide sequences), the web server runs Prodigal during the pre-processing stage, allowing different gene-calling settings to be used for inputs > 100 kb, versus shorter sequences (see the Prodigal documentation of an explanation of the rationale behind this). For users wanting to analyse unannotated plasmid sequences, we recommend including the host genome sequence within a multi-fasta input file before uploading to the PADLOC web server (to improve the quality of gene predictions with Prodigal). Once PADLOC is run, the output is passed to a post-processing script that reformats and outputs the data in a custom machine-readable format that is used to generate an interactive genome annotation display (produced using d3.js (https://d3js.org/)) on the corresponding user-job result page.

RESULTS

Input and file handling

Users can analyse archaea, bacteria, metagenome, phage, archaeal virus and plasmid genome files from the 'Run PADLOC' page. The PADLOC web server accepts GenBank flat files, nucleotide FASTA, or paired amino acid FASTA and general feature format (GFF3) files as input (Figure 2A). If a GenBank file is provided, nucleotide, protein, and feature information (e.g. gene locations) is extracted using Biopython. If a nucleotide FASTA file is provided, Prodigal (75) is used to predict open reading frames and produce a protein FASTA and GFF3 file. For the best quality results, it is recommended that users supply a GenBank file or amino acid FASTA and GFF3 file where coding sequences have already been called and verified (e.g. with Prodigal or Prokka). Users may wish to download the example genome files to see the expected formatting for each file type (Figure 2B). When nucleotide information is provided, either through a GenBank or nucleotide FASTA file, Infernal (71) is used to detect the ncRNA components of retrons. Users also have the option to run CRISPRDetect

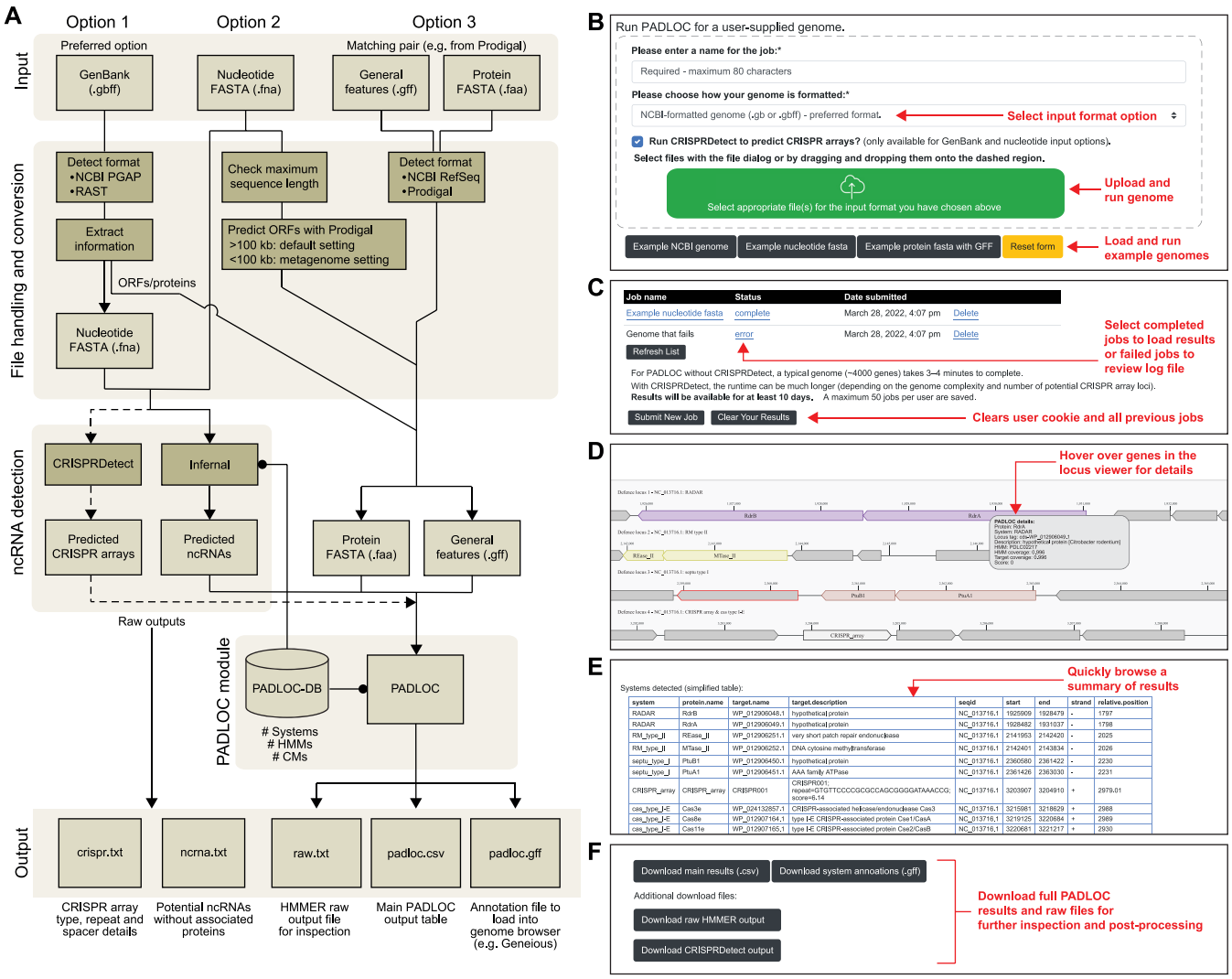


Figure 2. The PADLOC web server pipeline and results. (**A**) The PADLOC web server handles the pre-processing of user input to allow for additional input types and the identification of CRISPR arrays and ncRNAs. (**B**) Users can upload their own genomes or run an example genome through the 'Run PADLOC' page. (**C**) User jobs are listed on the 'My Results' page. Completed jobs link to their individual results pages, failed jobs link to a log file that provides information on why the genome could not be analysed. (**D**) The individual results pages display a locus viewer, which shows each identified defence system in the context of its surrounding genes. (**E**) A summarised version of the results is displayed under the locus viewer for an initial inspection of the systems identified. (**F**) Users can download the full PADLOC output and other raw files from the bottom of the page.

(70) to predict CRISPR arrays. This additional information helps to verify and enhance the quality of defence system detection. All input files are passed to the core PADLOC module, which then detects the defence systems specified in the PADLOC database (a current list of systems is provided on the web server). The expected processing time for a typical genome encoding ~5,000 proteins is less than five minutes. Including the CRISPRDetect option can substantially increase the run time in some cases, but usually only adds a couple of minutes.

Output and interpretation

Each job submitted by a user is listed on the 'My Results' page (Figure 2C), which includes details of the job status and links to completed jobs. If the job failed, a link is provided to download the corresponding log file. The most common reason for a failed job is incorrect input formatting. For successful jobs, individual result pages contain information about interpreting the output, an interactive view of the locus structure of each detected defence system (Figure 2D), a summary table of systems detected (Figure 2E), and options to download the output files (Figure 2F). The main PADLOC output file (.csv format) lists all systems identified, with one gene per row. The file contains information regarding the type of system detected, the proteins present, their location in the genome, and details about the confidence of detection. The most important values to consider when evaluating detection confidence are the full sequence and domain E-values (full.seq.E.value and domain.iE.value), and the target and HMM coverages (target.coverage and hmm.coverage). Usually, hits with large E-values (indicating low statistical significance) and low coverages should be treated with caution. In general, multi-gene

defence systems are detected with greater specificity than systems encoded by singe genes.

Considerations and limitations

As with any computational approach to inferring gene function, there are several potential limitations that users should consider when interpreting PADLOC web server outputs. Many defence proteins contain ubiquitous domains and their HMMs are more likely to detect spurious hits. For example, PtuA (Septu), several retron proteins and Old nucleases contain similar ATPase domains (11). As a result, PADLOC sometimes reports overlapping system classifications (typically less than 1% of results), which should be resolved via subjective evaluation of the reported scoring parameters and genetic context. For multi-gene systems, the synteny requirements resolve many ambiguities and increase the confidence of system classification (due to the reduced probability of two adjacent false-positive hits). By contrast, identification of single-gene systems is more challenging and requires trade-offs with the HMM scoring cut-offs (E-value and HMM/target coverage thresholds) to achieve an acceptable balance between sensitivity versus specificity. As such, false positive and negative results are inevitably more frequent for single gene systems. In general, the PADLOC scoring thresholds are set more toward sensitivity, with the intention that users interested in further study of identified potential defence system homologs will undertake additional analyses.

As a first step to curating PADLOC results, inspection of the HMM and target alignment coverage scores can reveal potential false-positive classifications (Figure 3A–C). In some cases, very similar proteins differ in function due to the presence or absence of enzymatic sites or functional motifs (Figure 3D). We encourage users to explore subsequent domain-based analyses of defence system proteins using tools such as HHpred (76) to identify protein domains with more granularity. We have also found structure prediction tools such as AlphaFold2 (77) and ColabFold (78), useful in identifying domain folds and boundaries. Once demarcated, predicted domain structures can be searched against protein structure databases such as the PDB (79) or AlphaFold-based databases, using tools like DALI (80) or Foldseek (81). In many cases, this structure-based approach reveals homologs with characterised active sites or functional motifs, which can aid in discrimination between defence system proteins and similar non-orthologous proteins. Users may also find it informative to compare their PADLOC results with DefenseFinder, another tool recently developed for defence system identification (82,83).

The similarity of several defence systems to other molecular systems inevitably leads to a background rate of false-positive system identifications. For example, Wadjet systems are similar to Muk structural maintenance of chromosomes (SMC) systems (21,84,85) (Figure 3E). The canonical Wadjet system comprises four proteins JetABCD, where JetA, JetB, and JetC share similarity with MukF, MukE, and MukB, respectively. The PADLOC Wadjet system definition requires the full JetABCD set to be present, whereas the MukFEB cases are reported as part of Wadjet 'other' systems. Several '[system]_other' models (which generally require only two components of a system to be co-localised) are run alongside the stricter canonical system definitions, to enable identification of systems that might otherwise be overlooked due to their being split by contig boundaries (particularly in metagenomes) (Figure 3F), fragmented by multiple intervening genes (e.g. due to MGE insertions), genes missing due to sequencing, assembly or gene-calling errors, mutations, or high sequence divergence of some defence proteins (Figure 3G). For the precomputed RefSeq data, we substituted pseudogene products with similar full-length protein sequences, which allows higher-confidence assignment of the example system as Wadjet type I (Figure 3H). The identification of defence system pseudogenes will also be helpful for studies of defence system evolution and turnover. For some systems, such as the Dnd and Pbe phosphorothioation systems, several genes are relatively short (including *dndE, pbeB/D*) and are often missed by gene prediction tools (33,35). The 'PT_other' model will detect the remaining genes and users should then manually check for short coding sequences in the vicinity of the expected location of any absent genes. Lastly, several system definition models specify 'optional' genes (e.g. 'cas_associated' proteins) that are not necessarily functionally associated with the system. In some cases, these proteins may be uncharacterised independent *bona fide* defence systems, or might have non-defence functions. To guide users in interpreting results for 'other' models and resolving ambiguities, each 'Results' page on the PADLOC web server contains a list of known potential ambiguities.

The current snapshot of antiviral defence systems

To provide a current and comprehensive quantitative view of the defence systems in bacteria and archaea, we used PADLOC to search all RefSeq v209 Bacteria and Archaea genomes. These results are available for browsing on the PADLOC web server under the 'RefSeq results' page and will be updated periodically with new RefSeq versions and as new defence systems are discovered. Overall, the distribution of different defence systems across different bacteria and archaea is highly varied (Supplementary Figure S1). It should be noted that this analysis includes assemblies of varied completeness, and systems may be underrepresented in taxa with incomplete genomes. Clear differences in the diversity and abundance of defence systems were apparent between phyla (Figure 4A), with several notable outliers including Chlamydiota (very few known defence systems) and Cyanobacteria (many types of defence systems in high abundance). Campylobacterota had a significant bimodal distribution of defence system abundance (Hartigan's dip test (86), $P < 0.001$), which relates to the *Helicobacteraceae* relying on a remarkably large number of restriction modification systems in each strain (Figure 4B,C) (87,88). Another interesting phyla was Spirochaetota, which had a significant bimodal distribution of defence system diversity due to *Borreliaceae* having very few types of defence systems (Hartigan's dip test (86), $P < 0.001$) (Figure 4D). Almost all *Borreliaceae* are tick-borne pathogens (89) with characteristically small genomes typical of obligate host-associated pathogenic bacteria (90). Similarly, *Treponema pallidum* (family *Treponemataceae*) are host-

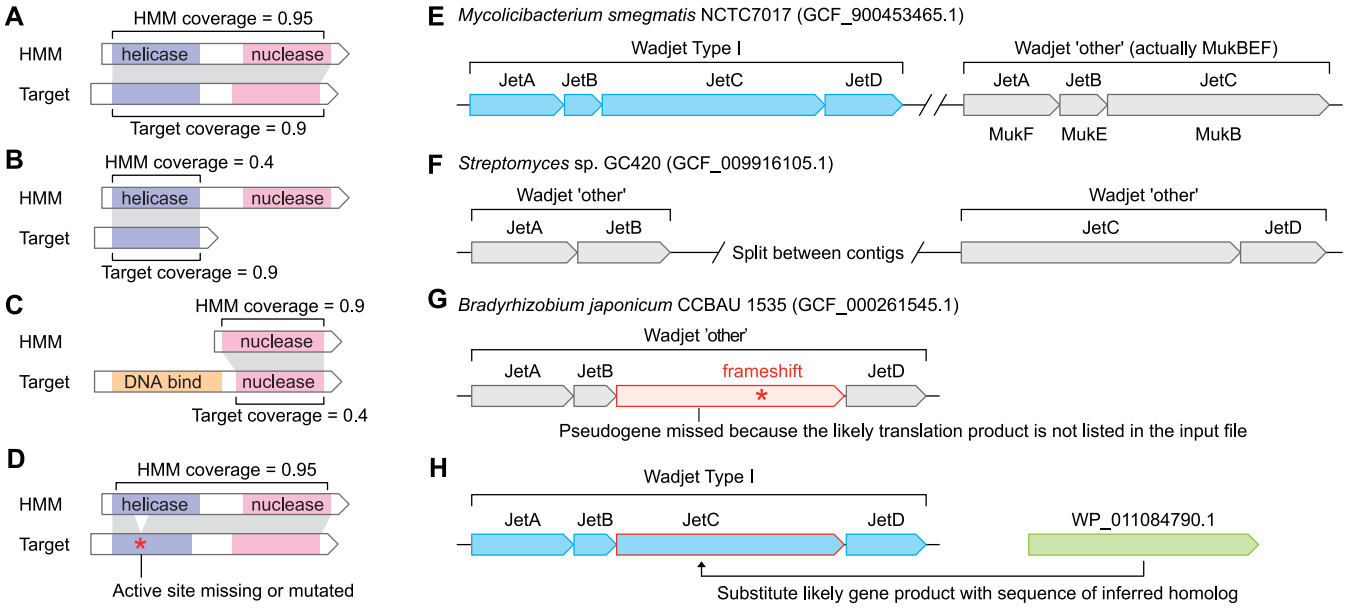


Figure 3. Considerations and limitations of defence system detection using profile HMMs and synteny criteria. (**A**) Likely hits to defence protein homologs have high alignment coverage between the HMM and target protein, including for multi-domain proteins, as the PADLOC HMMs were typically built from whole proteins rather than individual domains. (**B**) Users should be wary of cases where only part of the HMM aligns to the target protein, which may lack a domain important for function of defence system homologs. (**C**) Conversely, some defence protein domains have similarity to domains found within non-defence proteins, which can result in the PADLOC HMMs matching only part of the target protein. However, these cases might also represent defence system fusion proteins, or divergent homologs. (**D**) Where possible, users should follow up by also verifying the presence of expected active site residues and motifs important for domain fold and function. (**E**) Some defence systems are similar to non-defence molecular systems, such as the similarity between Wadjet and Muk systems. This is typically not an issue with multi-gene systems where all genes are present, but some [system]_other models may detect such cases where some genes are allowed to be absent. (**F**) Several [system]_other models allow the detection of fragmented multi-gene systems, which can be reconstructed manually after reviewing the results. (**G**) Pseudogenes within multi-gene systems are not detected by the default PADLOC workflow, so users should check [system]_other models for the potential presence of additional genes. In this example, a frameshift within *jetC* means the Wadjet system criteria are not fulfilled (because JetC was not detected) (**H**) In the precomputed PADLOC RefSeq dataset, pseudogenes were substituted with the protein homologs inferred by the PGAP pipeline, allowing the above example to be classified as a Wadjet Type I system. Pseudogenes are indicated by a red outline in the locus viewer and the prefix 'pseudo_sub' in the 'target.name' column of the output.

associated pathogens with small genomes and lack known defence systems (Figure 4D). It remains to be resolved whether the low defence diversity is driven by genome reduction and due to less exposure to phages and MGEs than free-living relatives. Overall, these examples illustrate that the PADLOC web server can be used to interrogate hypotheses such as these and to then identify candidate taxa and strains in which predictions can be experimentally tested.

DISCUSSION

To address the lack of capable and accessible tools for comprehensive defence system identification, we developed the PADLOC web server. Here, we have described the current state of PADLOC and important information regarding usage of the web server and interpretation of the output. PADLOC will continue to evolve as new defence systems are discovered and additional biological insight allows us to fine-tune the parameters of identification. Evaluating the accuracy of detection and adjusting these thresholds accordingly is difficult when only a few experimentally verified examples are available, as is currently the case with most defence systems. PADLOC provides a key initial step for further improvement, by facilitating the identification of many putative systems that can be followed up by functional investigation. Feedback and contribution to PADLOC and the web server is encouraged via the GitHub repository (https://github.com/padlocbio/padloc/issues), including but not limited to the addition of systems and HMMs, suggestions for adjusting thresholds or system classifications and nomenclature, and reporting bugs.

The comprehensive identification of many systems made possible with PADLOC also opens avenues for investigating many interesting biological questions. For example, many putative defence system genes are annotated as pseudogenes. Although pseudogenes can arise from sequencing or assembly errors, they might also include the remnants of defence systems that have become inactivated through mutation. Investigation of these remnants could provide insight into the ancestry and evolution of defence system arsenals that would otherwise remain undetected. In addition, our broad taxonomic analyses revealed notable differences in the diversity and abundance of defence systems in many taxa, raising the question of what factors drive the requirement for more defence against phages, and whether differences are driven by ecological factors such as phage diversity and encounter rates (reviewed in 92). Recent studies have also identified interplay between defence systems, with synergistic or antagonistic effects (42,93,94) and as more

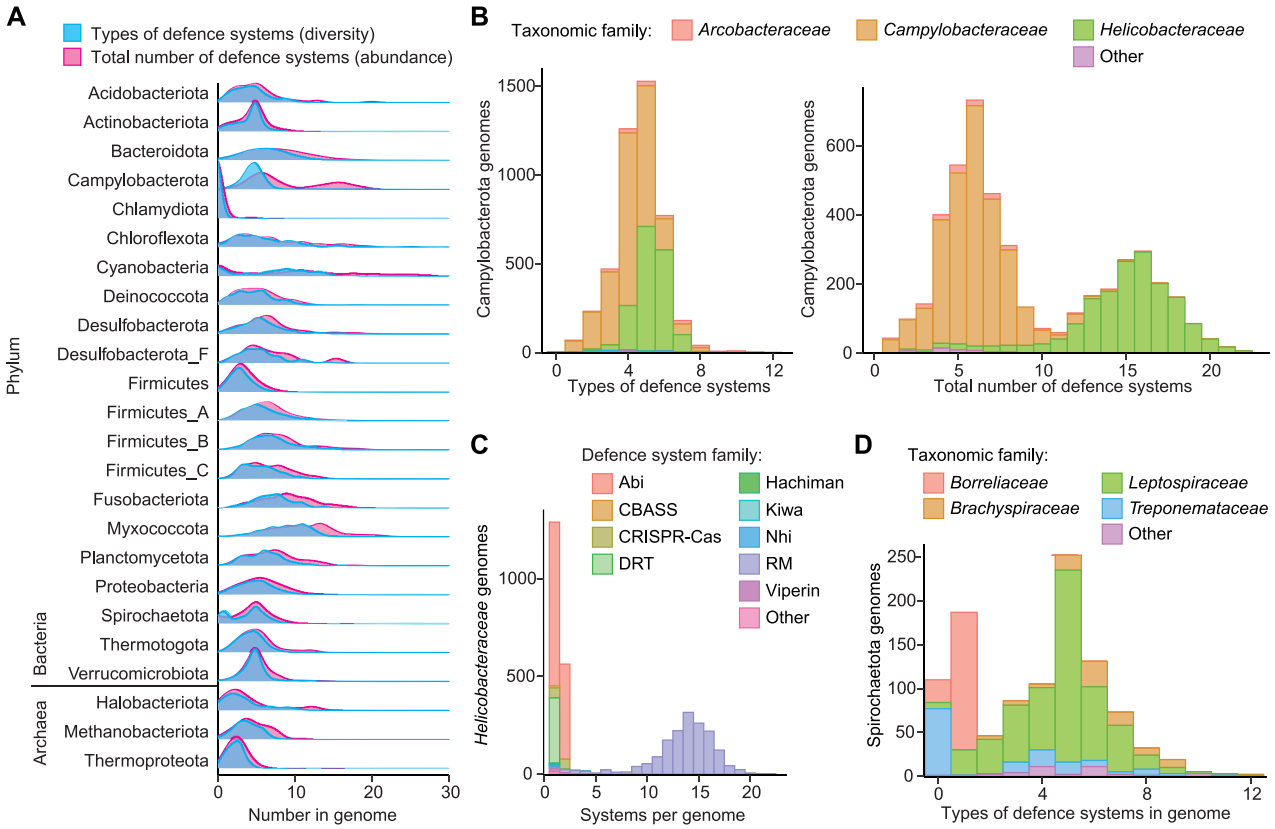


Figure 4: Example analyses of PADLOC data reveal lineage-specific differences in defence system diversity and abundance. (**A**) Overview of the defence system diversity (the number of unique types of defence systems within a host genome) and abundance (total number of defence systems within a host genome), separated by phyla (per the GTDB taxonomy (91)). Only phyla with more than 50 genomes are displayed. (**B**) A closer look at defence system diversity and abundance within the Campylobacterota phylum. (**C**) A breakdown of the abundance of different defence system types within the *Helicobacteraceae* family of Campylobacterota. Defence system types occurring in less than five genomes are grouped under 'Other types'. (**D**) Defence system diversity within Spriochaetota, revealing low defence diversity within *Borreliaceae*. The *Treponemataceae* lacking defence systems (types = 0) in this dataset are comprised entirely of *Treponema pallidum*.

types of defence systems are discovered, our understanding of compatibility between systems needs to be revisited. The PADLOC web server provides a convenient and accessible platform to detect suitable candidate strains for experimental work to resolve these outstanding questions.

DATA AVAILABILITY

The PADLOC web server is freely available at https://padloc.otago.ac.nz. This website is open to all users and there is no login requirement. Source code and documentation for installing and running PADLOC locally are freely available from the PADLOC GitHub repository (https://github.com/padlocbio/padloc). The HMMs and system models used by the PADLOC web server are available from the PADLOC database repository (https://github.com/padlocbio/padloc-db).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

Firstly, we thank the many researchers whose contributions that have contributed to development of PADLOC and the users that have provided constructive feedback to improve the PADLOC database and web server. We are also grateful to members of the Phage-host interactions laboratory at the University of Otago for helpful discussions and feedback. We acknowledge and appreciate the use of the New Zealand eScience Infastructure (NeSI) high-performance computing facilities in this research, which are funded jointly by NeSI collaborator institutions and the Ministry of Buisness, Innovation and Employment. N.T. is supported by grant PID2020-113207GB-I00 from the MCIN/ AEI /10.13039/501100011033.

FUNDING

This work was supported by the Royal Society of New Zealand Te Apārangi (RSNZ) Marsden Fund, the School of Biomedical Sciences Bequest Fund from the University of Otago and Bioprotection Aotearoa (Tertiary Education Commission, NZ). L.J.P. was supported by a University of Otago Doctoral Scholarship.

Conflict of interest statement. None declared.

REFERENCES

1. Bernheim,A. and Sorek,R. (2020) The pan-immune system of bacteria: antiviral defence as a community resource. *Nat. Rev. Microbiol.*, **18**, 113–119.
2. Hampton,H.G., Watson,B.N.J. and Fineran,P.C. (2020) The arms race between bacteria and their phage foes. *Nature*, **577**, 327–336.
3. Isaev,A.B., Musharova,O.S. and Severinov,K.V. (2021) Microbial arsenal of antiviral defenses – part I. *Biochem. Mosc.*, **86**, 319–337.
4. Isaev,A.B., Musharova,O.S. and Severinov,K.V. (2021) Microbial arsenal of antiviral defenses. Part II. *Biochem. Mosc.*, **86**, 449–470.
5. Tal,N. and Sorek,R. (2022) SnapShot: bacterial immunity. *Cell*, **185**, 578–578.
6. Payne,L.J., Todeschini,T.C., Wu,Y., Perry,B.J., Ronson,C.W., Fineran,P.C., Nobrega,F.L. and Jackson,S.A. (2021) Identification and classification of antiviral defence systems in bacteria and archaea with PADLOC reveals new system types. *Nucleic Acids Res.*, **49**, 10868–10878.
7. Abby,S.S., Néron,B., Ménager,H., Touchon,M. and Rocha,E.P.C. (2014) MacSyFinder: a program to mine genomes for molecular systems with an application to CRISPR-Cas systems. *PLoS ONE*, **9**, e110726.
8. Russel,J., Pinilla-Redondo,R., Mayo-Muñoz,D., Shah,S.A. and Sørensen,S.J. (2020) CRISPRCasTyper: automated identification, annotation, and classification of CRISPR-Cas loci. *CRISPR J.*, **3**, 462–469.
9. Abby,S.S., Cury,J., Guglielmini,J., Néron,B., Touchon,M. and Rocha,E.P.C. (2016) Identification of protein secretion systems in bacterial genomes. *Sci. Rep.*, **6**, 23080.
10. Burroughs,A.M., Iyer,L.M. and Aravind,L. (2013) Two novel PIWI families: roles in inter-genomic conflicts in bacteria and Mediator-dependent modulation of transcription in eukaryotes. *Biol. Direct*, **8**, 13.
11. Mestre,M.R., González-Delgado,A., Gutiérrez-Rus,L.I., Martínez-Abarca,F. and Toro,N. (2020) Systematic prediction of genes functionally associated with bacterial retrons and classification of the encoded tripartite systems. *Nucleic Acids Res.*, **48**, 12632–12647.
12. Rousset,F., Depardieu,F., Solange,M., Dowding,J., Laval,A.-L., Lieberman,E., Garry,D., Rocha,E.P.C., Bernheim,A. and Bikard,D. (2022) Phages and their satellites encode hotspots of antiviral systems. *Cell*, **30**, 740–753.
13. Couvin,D., Bernheim,A., Toffano-Nioche,C., Touchon,M., Michalik,J., Néron,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.
14. 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.
15. Shmakov,S.A., Makarova,K.S., Wolf,Y.I., Severinov,K.V. and Koonin,E.V. (2018) Systematic prediction of genes functionally linked to CRISPR-Cas systems by gene neighborhood analysis. *Proc. Natl. Acad. Sci.*, **115**, E5307–E5316.
16. Galperin,M.Y., Kristensen,D.M., Makarova,K.S., Wolf,Y.I. and Koonin,E.V. (2019) Microbial genome analysis: the COG approach. *Brief. Bioinform.*, **20**, 1063–1070.
17. Makarova,K.S., Wolf,Y.I., Iranzo,J., Shmakov,S.A., Alkhnbashi,O.S., Brouns,S.J.J., Charpentier,E., Cheng,D., Haft,D.H., Horvath,P. *et al.* (2020) Evolutionary classification of CRISPR–Cas systems: a burst of class 2 and derived variants. *Nat. Rev. Microbiol.*, **18**, 67–83.
18. Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.
19. Eddy,S.R. (2011) Accelerated profile HMM searches. *PLoS Comput. Biol.*, **7**, e1002195.
20. Bernheim,A., Millman,A., Ofir,G., Meitav,G., Avraham,C., Shomar,H., Rosenberg,M.M., Tal,N., Melamed,S., Amitai,G. *et al.* (2021) Prokaryotic viperins produce diverse antiviral molecules. *Nature*, **589**, 120–124.
21. 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.
22. Gao,L., Altae-Tran,H., Böhning,F., Makarova,K.S., Segel,M., Schmid-Burgk,J.L., Koob,J., Wolf,Y.I., Koonin,E.V. and Zhang,F. (2020) Diverse enzymatic activities mediate antiviral immunity in prokaryotes. *Science*, **369**, 1077–1084.
23. 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.
24. Johnson,A.G., Wein,T., Mayer,M.L., Duncan-Lowey,B., Yirmiya,E., Oppenheimer-Shaanan,Y., Amitai,G., Sorek,R. and Kranzusch,P.J. (2022) Bacterial gasdermins reveal an ancient mechanism of cell death. *Science*, **375**, 221–225.
25. Makarova,K.S., Wolf,Y.I., van der Oost,J. and Koonin,E.V. (2009) Prokaryotic homologs of Argonaute proteins are predicted to function as key components of a novel system of defense against mobile genetic elements. *Biol. Direct*, **4**, 29.
26. Millman,A., Melamed,S., Amitai,G. and Sorek,R. (2020) Diversity and classification of cyclic-oligonucleotide-based anti-phage signalling systems. *Nat. Microbiol.*, **5**, 1608–1615.
27. 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.
28. 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.
29. Shah,S.A., Alkhnbashi,O.S., Behler,J., Han,W., She,Q., Hess,W.R., Garrett,R.A. and Backofen,R. (2019) Comprehensive search for accessory proteins encoded with archaeal and bacterial type III CRISPR-cas gene cassettes reveals 39 new cas gene families. *RNA Biol.*, **16**, 530–542.
30. Tal,N., Millman,A., Stokar-Avihail,A., Fedorenko,T., Leavitt,A., Melamed,S., Yirmiya,E., Avraham,C., Amitai,G. and Sorek,R. (2021) Antiviral defense via nucleotide depletion in bacteria. bioRxiv doi: https://doi.org/10.1101/2021.04.26.441389, 26 April 2021, preprint: not peer reviewed.
31. Tal,N., Morehouse,B.R., Millman,A., Stokar-Avihail,A., Avraham,C., Fedorenko,T., Yirmiya,E., Herbst,E., Brandis,A., Mehlman,T. *et al.* (2021) Cyclic CMP and cyclic UMP mediate bacterial immunity against phages. *Cell*, **184**, 5728–5739.
32. Thiaville,J.J., Kellner,S.M., Yuan,Y., Hutinet,G., Thiaville,P.C., Jumpathong,W., Mohapatra,S., Brochier-Armanet,C., Letarov,A.V., Hillebrand,R. *et al.* (2016) Novel genomic island modifies DNA with 7-deazaguanine derivatives. *Proc. Natl. Acad. Sci.*, **113**, E1452–E1459.
33. Tong,T., Chen,S., Wang,L., Tang,Y., Ryu,J.Y., Jiang,S., Wu,X., Chen,C., Luo,J., Deng,Z. *et al.* (2018) Occurrence, evolution, and functions of DNA phosphorothioate epigenetics in bacteria. *Proc. Natl. Acad. Sci.*, **115**, E2988–E2996.
34. Wang,S., Wan,M., Huang,R., Zhang,Y., Xie,Y., Wei,Y., Ahmad,M., Wu,D., Hong,Y., Deng,Z. *et al.* (2021) SspABCD-SspFGH constitutes a new type of DNA phosphorothioate-based bacterial defense system. *Mbio*, **12**, e00613-21.
35. Xiong,L., Liu,S., Chen,S., Xiao,Y., Zhu,B., Gao,Y., Zhang,Y., Chen,B., Luo,J., Deng,Z. *et al.* (2019) A new type of DNA phosphorothioation-based antiviral system in archaea. *Nat. Commun.*, **10**, 1688.
36. Xiong,X., Wu,G., Wei,Y., Liu,L., Zhang,Y., Su,R., Jiang,X., Li,M., Gao,H., Tian,X. *et al.* (2020) SspABCD–SspE is a phosphorothioation-sensing bacterial defence system with broad anti-phage activities. *Nat. Microbiol.*, **5**, 917–928.
37. Xu,T., Yao,F., Zhou,X., Deng,Z. and You,D. (2010) A novel host-specific restriction system associated with DNA backbone S-modification in Salmonella. *Nucleic Acids Res.*, **38**, 7133–7141.
38. Yuan,Y., Hutinet,G., Valera,J.G., Hu,J., Hillebrand,R., Gustafson,A., Iwata-Reuyl,D., Dedon,P.C. and de Crécy-Lagard,V. (2018) Identification of the minimal bacterial 2′-deoxy-7-amido-7-deazaguanine synthesis machinery. *Mol. Microbiol.*, **110**, 469–483.
39. Zeng,Z., Chen,Y., Pinilla-Redondo,R., Shah,S.A., Zhao,F., Wang,C., Hu,Z., Zhang,C., Whitaker,R.J., She,Q. *et al.* (2021) A short prokaryotic argonaute cooperates with membrane effector to confer

antiviral defense. bioRxiv doi: https://doi.org/10.1101/2021.12.09.471704, 11 December 2021, preprint: not peer reviewed.
40. Anba,J., Bidnenko,E., Hillier,A., Ehrlich,D. and Chopin,M.C. (1995) Characterization of the lactococcal abiD1 gene coding for phage abortive infection. *J. Bacteriol.*, **177**, 3818–3823.
41. Bergsland,K.J., Kao,C., Yu,Y.-T.N., Gulati,R. and Snyder,L. (1990) A site in the T4 bacteriophage major head protein gene that can promote the inhibition of all translation in Escherichia coli. *J. Mol. Biol.*, **213**, 477–494.
42. Birkholz,N., Jackson,S.A., Fagerlund,R.D. and Fineran,P.C. (2022) A mobile restriction–modification system provides phage defence and resolves an epigenetic conflict with an antagonistic endonuclease. *Nucleic Acids Res.*, **50**,3348–3361.
43. Bouchard,J.D., Dion,E., Bissonnette,F. and Moineau,S. (2002) Characterization of the two-component abortive phage infection mechanism AbiT from Lactococcus lactis. *J. Bacteriol.*, **184**, 6325–6332.
44. Cluzel,P.J., Chopin,A., Ehrlich,S.D. and Chopin,M.C. (1991) Phage abortive infection mechanism from Lactococcus lactis subsp. lactis, expression of which is mediated by an Iso-ISS1 element. *Appl. Environ. Microbiol.*, **57**, 3547–3551.
45. Cram,D., Ray,A. and Skurray,R. (1984) Molecular analysis of F plasmid pif region specifying abortive infection of T7 phage. *Mol. Gen. Genet. MGG*, **197**, 137–142.
46. Dai,G., Su,P., Allison,G.E., Geller,B.L., Zhu,P., Kim,W.S. and Dunn,N.W. (2001) Molecular characterization of a new abortive infection system (AbiU) from Lactococcus lactis LL51-1. *Appl. Environ. Microbiol.*, **67**, 5225–5232.
47. Deng,Y.-M., Liu,C.-Q. and Dunn,W., N. (1999) Genetic organization and functional analysis of a novel phage abortive infection system, AbiL, from Lactococcus lactis. *J. Biotechnol.*, **67**, 135–149.
48. Deng,Y.-M., Harvey,M.L., Liu,C.-Q. and Dunn,N.W. (2006) A novel plasmid-encoded phage abortive infection system from Lactococcus lactis biovar. diacetylactis. *FEMS Microbiol. Lett.*, **146**, 149–154.
49. Dinsmore,P.K. and Klaenhammer,T.R. (1994) Phenotypic consequences of altering the copy number of abiA, a gene responsible for aborting bacteriophage infections in Lactococcus lactis. *Appl. Environ. Microbiol.*, **60**, 1129–1136.
50. Domingues,S., Chopin,A., Ehrlich,S.D. and Chopin,M.-C. (2004) The lactococcal abortive phage infection system AbiP prevents both phage DNA replication and temporal transcription switch. *J. Bacteriol.*, **186**, 713–721.
51. Durmaz,E., Higgins,D.L. and Klaenhammer,T.R. (1992) Molecular characterization of a second abortive phage resistance gene present in Lactococcus lactis subsp. lactis ME2. *J. Bacteriol.*, **174**, 7463–7469.
52. Dy,R.L., Przybilski,R., Semeijn,K., Salmond,G.P.C. and Fineran,P.C. (2014) A widespread bacteriophage abortive infection system functions through a type IV toxin–antitoxin mechanism. *Nucleic Acids Res.*, **42**, 4590–4605.
53. Emond,E., Holler,B.J., Boucher,I., Vandenbergh,P.A., Vedamuthu,E.R., Kondo,J.K. and Moineau,S. (1997) Phenotypic and genetic characterization of the bacteriophage abortive infection mechanism AbiK from Lactococcus lactis. *Appl. Environ. Microbiol.*, **63**, 1274–1283.
54. Emond,E., Dion,E., Walker,S.A., Vedamuthu,E.R., Kondo,J.K. and Moineau,S. (1998) AbiQ, an abortive infection mechanism from Lactococcus lactis. *Appl. Environ. Microbiol.*, **64**, 4748–4756.
55. Garvey,P., Fitzgerald,G.F. and Hill,C. (1995) Cloning and DNA sequence analysis of two abortive infection phage resistance determinants from the lactococcal plasmid pNP40. *Appl. Environ. Microbiol.*, **61**, 4321–4328.
56. Jabbar,M.A. and Snyder,L. (1984) Genetic and physiological studies of an Escherichia coli locus that restricts polynucleotide kinase- and RNA ligase-deficient mutants of bacteriophage T4. *J. Virol.*, **51**, 522–529.
57. Lindahl,G., Sironi,G., Bialy,H. and Calendar,R. (1970) Bacteriophage lambda; abortive infection of bacteria lysogenic for phage P2. *Proc. Natl. Acad. Sci.*, **66**, 587–594.
58. McLandsborough,L.A., Kolaetis,K.M., Requena,T. and McKay,L.L. (1995) Cloning and characterization of the abortive infection genetic determinant abiD isolated from pBF61 of Lactococcus lactis subsp. lactis KR5. *Appl. Environ. Microbiol.*, **61**, 2023–2026.
59. Millman,A., Bernheim,A., Stokar-Avihail,A., Fedorenko,T., Voichek,M., Leavitt,A., Oppenheimer-Shaanan,Y. and Sorek,R. (2020) Bacterial retrons function in anti-phage defense. *Cell*, **183**, 1551–1561.
60. O'Connor,L., Coffey,A., Daly,C. and Fitzgerald,G.F. (1996) AbiG, a genotypically novel abortive infection mechanism encoded by plasmid pCI750 of Lactococcus lactis subsp. cremoris UC653. *Appl. Environ. Microbiol.*, **62**, 3075–3082.
61. Owen,S.V., Wenner,N., Dulberger,C.L., Rodwell,E.V., Bowers-Barnard,A., Quinones-Olvera,N., Rigden,D.J., Rubin,E.J., Garner,E.C., Baym,M. *et al.* (2021) Prophages encode phage-defense systems with cognate self-immunity. *Cell Host Microbe*, **29**, 1620–1633.
62. Parma,D.H., Snyder,M., Sobolevski,S., Nawroz,M., Brody,E. and Gold,L. (1992) The Rex system of bacteriophage lambda: tolerance and altruistic cell death. *Genes Dev.*, **6**, 497–510.
63. Parreira,R., Ehrlich,S.D. and Chopin,M.-C. (1996) Dramatic decay of phage transcripts in lactococcal cells carrying the abortive infection determinant AbiB. *Mol. Microbiol.*, **19**, 221–230.
64. Prévots,F. and Ritzenthaler,P. (1998) Complete sequence of the new lactococcal abortive phage resistance gene abiO. *J. Dairy Sci.*, **81**, 1483–1485.
65. Prévots,F., Daloyau,M., Bonin,O., Dumont,X. and Tolou,S. (1996) Cloning and sequencing of the novel abortive infection gene *abiH* of *Lactococcuslactis* ssp. *lactis* biovar. *diacetylactis* S94. *FEMS Microbiol. Lett.*, **142**, 295–299.
66. Prévots,F., Tolou,S., Delpech,B., Kaghad,M. and Daloyau,M. (1998) Nucleotide sequence and analysis of the new chromosomal abortive infection gene abiN of Lactococcus lactis subsp. cremoris S114. *FEMS Microbiol. Lett.*, **159**, 331–336.
67. Sberro,H., Leavitt,A., Kiro,R., Koh,E., Peleg,Y., Qimron,U. and Sorek,R. (2013) Discovery of functional toxin/antitoxin systems in bacteria by shotgun cloning. *Mol. Cell*, **50**, 136–148.
68. Smith,H.S., Pizer,L.I., Pylkas,L. and Lederberg,S. (1969) Abortive infection of shigella dysenteriae P2 by T2 bacteriophage. *J. Virol.*, **4**, 162–168.
69. Su,P., Harvey,M., Im,H.J. and Dunn,N.W. (1997) Isolation, cloning and characterisation of the abiI gene from Lactococcus lactis subsp. lactis M138 encoding abortive phage infection. *J. Biotechnol.*, **54**, 95–104.
70. Biswas,A., Staals,R.H.J., Morales,S.E., Fineran,P.C. and Brown,C.M. (2016) CRISPRDetect: a flexible algorithm to define CRISPR arrays. *BMC Genomics*, **17**, 356.
71. Nawrocki,E.P. and Eddy,S.R. (2013) Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*, **29**, 2933–2935.
72. Haft,D.H., DiCuccio,M., Badretdin,A., Brover,V., Chetvernin,V., O'Neill,K., Li,W., Chitsaz,F., Derbyshire,M.K., Gonzales,N.R. *et al.* (2018) RefSeq: an update on prokaryotic genome annotation and curation. *Nucleic Acids Res.*, **46**, D851–D860.
73. Tatusova,T., DiCuccio,M., Badretdin,A., Chetvernin,V., Nawrocki,E.P., Zaslavsky,L., Lomsadze,A., Pruitt,K.D., Borodovsky,M. and Ostell,J. (2016) NCBI prokaryotic genome annotation pipeline. *Nucleic Acids Res.*, **44**, 6614–6624.
74. Cock,P.J.A., Antao,T., Chang,J.T., Chapman,B.A., Cox,C.J., Dalke,A., Friedberg,I., Hamelryck,T., Kauff,F., Wilczynski,B. *et al.* (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, **25**, 1422–1423.
75. Hyatt,D., Chen,G.-L., LoCascio,P.F., Land,M.L., Larimer,F.W. and Hauser,L.J. (2010) Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, **11**, 119.
76. Zimmermann,L., Stephens,A., Nam,S.-Z., Rau,D., Kübler,J., Lozajic,M., Gabler,F., Söding,J., Lupas,A.N. and Alva,V. (2018) A completely reimplemented MPI bioinformatics toolkit with a new HHpred server at its core. *J. Mol. Biol.*, **430**, 2237–2243.
77. Jumper,J., Evans,R., Pritzel,A., Green,T., Figurnov,M., Ronneberger,O., Tunyasuvunakool,K., Bates,R., Žídek,A., Potapenko,A. *et al.* (2021) Highly accurate protein structure prediction with AlphaFold. *Nature*, **596**, 583–589.
78. Mirdita,M., Schütze,K., Moriwaki,Y., Heo,L., Ovchinnikov,S. and Steinegger,M. (2022) ColabFold - Making protein folding accessible to all. bioRxiv doi: https://doi.org/10.1101/2021.08.15.456425, 15 August 2021, preprint: not peer reviewed.
79. Berman,H.M. (2000) The protein data bank. *Nucleic Acids Res.*, **28**, 235–242.

80. Holm,L. (2020) DALI and the persistence of protein shape. *Protein Sci.*, **29**, 128–140.
81. Kempen,M., Kim,S.S., Tumescheit,C., Mirdita,M., Söding,J. and Steinegger,M. (2022) Foldseek: fast and accurate protein structure search. bioRxiv doi: https://doi.org/10.1101/2022.02.07.479398, 09 February 2022, preprint: not peer reviewed.
82. Cazares,A., Figueroa,W. and Cazares,D. (2022) Diversity of microbial defence systems. *Nat. Rev. Microbiol.*, **20**, 191.
83. Tesson,F., Hervé,A., Touchon,M., Humières,C., Cury,J. and Bernheim,A. (2021) Systematic and quantitative view of the antiviral arsenal of prokaryotes. bioRxiv doi: https://doi.org/10.1101/2021.09.02.458658, 03 September 2021, preprint: not peer reviewed.
84. Krishnan,A., Burroughs,A.M., Iyer,L.M. and Aravind,L. (2020) Comprehensive classification of ABC ATPases and their functional radiation in nucleoprotein dynamics and biological conflict systems. *Nucleic Acids Res.*, **48**, 10045–10075.
85. Panas,M.W., Jain,P., Yang,H., Mitra,S., Biswas,D., Wattam,A.R., Letvin,N.L. and Jacobs,W.R. (2014) Noncanonical SMC protein in Mycobacterium smegmatis restricts maintenance of Mycobacterium fortuitum plasmids. *Proc. Natl. Acad. Sci.*, **111**, 13264–13271.
86. Hartigan,J.A. and Hartigan,P.M. (1985) The dip test of unimodality. *Ann. Stat.*, **13**, 70–84.
87. Krebes,J., Morgan,R.D., Bunk,B., Spröer,C., Luong,K., Parusel,R., Anton,B.P., König,C., Josenhans,C., Overmann,J. *et al.* (2014) The complex methylome of the human gastric pathogen Helicobacter pylori. *Nucleic Acids Res.*, **42**, 2415–2432.
88. Lin,L.F., Posfai,J., Roberts,R.J. and Kong,H. (2001) Comparative genomics of the restriction-modification systems in Helicobacter pylori. *Proc. Natl. Acad. Sci. U. S. A.*, **98**, 2740–2745.
89. Barbour,A.G. and Gupta,R.S. (2021) The family borreliaceae (Spirochaetales), a diverse group in two genera of tick-borne spirochetes of mammals, birds, and reptiles. *J. Med. Entomol.*, **58**, 1513–1524.
90. Moran,N.A. (2002) Microbial minimalism: genome reduction in bacterial pathogens. *Cell*, **108**, 583–586.
91. Parks,D.H., Chuvochina,M., Chaumeil,P.-A., Rinke,C., Mussig,A.J. and Hugenholtz,P. (2020) A complete domain-to-species taxonomy for Bacteria and Archaea. *Nat. Biotechnol.*, **38**, 1079–1086.
92. van Houte,S., Buckling,A. and Westra,E.R. (2016) Evolutionary ecology of prokaryotic immune mechanisms. *Microbiol. Mol. Biol. Rev.*, **80**, 745–763.
93. Hynes,A.P., Villion,M. and Moineau,S. (2014) Adaptation in bacterial CRISPR-Cas immunity can be driven by defective phages. *Nat. Commun.*, **5**, 4399.
94. Maguin,P., Varble,A., Modell,J.W. and Marraffini,L.A. (2022) Cleavage of viral DNA by restriction endonucleases stimulates the type II CRISPR-Cas immune response. *Mol. Cell*, **82**, 907–919.