10868–10878 *Nucleic Acids Research, 2021, Vol. 49, No. 19* *Published online 4 October 2021*
https://doi.org/10.1093/nar/gkab883

Identification and classification of antiviral defence systems in bacteria and archaea with PADLOC reveals new system types

Leighton J. Payne [1], Thomas C. Todeschini[2], Yi Wu[2], Benjamin J. Perry[1], Clive W. Ronson [1,3], Peter C. Fineran [1,3,4,5], Franklin L. Nobrega [2] and Simon A. Jackson [1,3,4,5,*]

[1]Department of Microbiology and Immunology, University of Otago, Dunedin, New Zealand, [2]School of Biological Sciences, Faculty of Environmental and Life Sciences, University of Southampton, Southampton, UK, [3]Genetics Otago, University of Otago, Dunedin, New Zealand, [4]Bioprotection Aotearoa, University of Otago, Dunedin, New Zealand and [5]Maurice Wilkins Centre for Molecular Biodiscovery, University of Otago, Dunedin, New Zealand

Received August 05, 2021; Revised September 13, 2021; Editorial Decision September 15, 2021; Accepted September 17, 2021

ABSTRACT

To provide protection against viral infection and limit the uptake of mobile genetic elements, bacteria and archaea have evolved many diverse defence systems. The discovery and application of CRISPR-Cas adaptive immune systems has spurred recent interest in the identification and classification of new types of defence systems. Many new defence systems have recently been reported but there is a lack of accessible tools available to identify homologs of these systems in different genomes. Here, we report the Prokaryotic Antiviral Defence LOCator (PADLOC), a flexible and scalable open-source tool for defence system identification. With PADLOC, defence system genes are identified using HMM-based homologue searches, followed by validation of system completeness using gene presence/absence and synteny criteria specified by customisable system classifications. We show that PADLOC identifies defence systems with high accuracy and sensitivity. Our modular approach to organising the HMMs and system classifications allows additional defence systems to be easily integrated into the PADLOC database. To demonstrate application of PADLOC to biological questions, we used PADLOC to identify six new subtypes of known defence systems and a putative novel defence system comprised of a helicase, methylase and ATPase. PADLOC is available as a standalone package (https://github.com/padlocbio/padloc) and as a webserver (https://padloc.otago.ac.nz).

INTRODUCTION

Bacteria and archaea possess a variety of defence systems to protect against diverse types of phages and mobile genetic elements (MGEs) (1,2) and to limit phages and MGEs from evading defence (3,4). The discovery and characterisation of novel defence systems has increased our understanding of the interactions between phages or MGEs and their hosts, and has led to the discovery of unique enzyme functionality that has been repurposed for new molecular tools, such as Cas9 for genome editing (5–7). Within genomes, defence systems are often concentrated in distinct genomic loci termed 'defence islands' (8,9). Many new types of defence systems have recently been discovered by studying the genomic 'dark matter' of defence islands using a guilt-by-association approach—uncharacterised genes that commonly reside next to genes of known phage defence systems often encode novel defence systems (10–14). As more genomic data are deposited into sequence databases, there are also renewed efforts to comprehensively identify and characterise known defence systems (15–19).

Several software tools have been developed to identify defence systems in prokaryotic genomes (20–26). However, these tools are often tailored to identify specific types of defence systems, such as CRISPR-Cas. Several precomputed databases are available for other defence systems including toxin-antitoxin and restriction-modification systems (27–31). However, these databases are limited to publicly available data and most lack the capability of searching user-supplied genomes on demand. As new types of defence systems are discovered and existing defence system classifications are revised, software tools will need to adapt to use this new information. To address the lack in capability of current tools to identify many types of phage defence systems, we have developed a

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

scalable, open-source Prokaryotic Antiviral Defence LOCator (PADLOC). PADLOC is available as a standalone package (https://github.com/padlocbio/padloc) and as a webserver (https://padloc.otago.ac.nz). Both resources allow analysis of user-supplied genomes, and the webserver includes precomputed PADLOC results from the RefSeq Bacteria and Archaea genome database (32).

Since most phage defence systems function through the coordinated action of multiple proteins that are encoded together in a single genomic locus, PADLOC uses a modified implementation of an approach previously developed to identify multi-gene macromolecular systems (20). Briefly, genes encoding defence system homologues are identified using profile Hidden Markov Models (HMMs), followed by validation of defence system completeness using gene presence/absence requirements specified in defence system classification files. For the initial release of PADLOC, we focused on two large groups of recently discovered phage defence systems, those identified in Doron *et al.* (13) (Druantia, Gabija, Hachiman, Kiwa, Lamassu, Septu, Shedu, Thoeris, Wadjet and Zorya, hereafter the 'Doron systems') and the cyclic-oligonucleotide-based anti-phage signalling system (CBASS) classifications described in Millman *et al.* (18). The Doron systems include the Wadjet systems that provide plasmid defence and are equivalent to the efficient plasmid transformation (*ept*) systems discovered in *Mycobacterium smegmatis* (33). We demonstrate that PADLOC can be used to identify phage and plasmid defence systems in prokaryotic genomes with high accuracy and specificity. In addition, we have used PADLOC to discover several new variant types of Doron systems, providing a foundation for further functional research. For future scalability, we used a modular approach for the organisation of HMMs and system classifications, which allows new defence systems to easily be added to the PADLOC defence systems database. As such, PADLOC provides a framework for continued community development into an all-in-one tool to identify the rapidly expanding set of known defence systems.

MATERIALS AND METHODS

PADLOC implementation

PADLOC uses a protein FASTA and corresponding Generic Feature Format (GFF3) file as input, which are commonly generated by genome annotation pipelines including the NCBI Prokaryotic Genome Annotation Pipeline (34), IMG Annotation Pipeline (35) and Prokka (36). Alternatively, a nucleotide FASTA file can be supplied as input, in which case Prodigal (37) is used to predict open reading frames and produce a protein FASTA and GFF3 file. Defence system proteins are identified using profile Hidden Markov Models with HMMER (38). Defence system classifications are described in YAML (a simple data serialisation language) formatted files (see Figure 1A, Supplementary Figure S1A and the PADLOC database GitHub repository https://github.com/padlocbio/padloc-db, for example system classification structure). All HMMs and classification files are available from the PADLOC database GitHub repository. In the YAML system classifications, proteins are designated as *core, optional or prohibited*. *Core* proteins are those expected to be present for a functional system. Proteins classed as *optional* are not strictly required for system identification. Specifying proteins as *prohibited* is useful when distinguishing between similar types of systems that may share *core* components but differ by a few key proteins. For each classification, a minimum number of *core* genes (*minimum_core*) and total *core*/*optional* genes (*minimum_total*) must be satisfied. When the requirements for a system are met, the location and details of the relevant corresponding genes are recorded as output. A simplified GFF file is also generated, allowing for annotation of the defence systems in genome viewing software. The typical run time of PADLOC for a genome encoding ∼4,500 proteins is less than one minute.

Building HMMs, system classifications and benchmarking

To build profile HMMs for the Doron and CBASS system proteins, we first retrieved the relevant protein sequences from defence system loci listed by Doron *et al.* (13) and Millman *et al.* (18). Redundant identical sequences were removed using SeqKit v0.13.2 (39). The sequences of each protein were then clustered at 30% minimum sequence identity and 80% alignment coverage with MMseqs2 v12.113e3 (40). If a cluster contained >100 sequences, redundancy was reduced at a threshold of 90% sequence identity and 90% pairwise alignment coverage with CDHIT v4.8.1 (41) using the accurate/slow clustering mode. Clusters with less than 200 sequences were aligned with MUSCLE v3.8.1551 (42) with anchor optimisation disabled and clusters with >200 sequences were aligned with MAFFT v7.471 (43) using one guide tree. An HMM was built for each cluster with at least five sequences using HMMER v3.3 with default parameters. PADLOC system classifications were written (in YAML format) to represent reported Doron and CBASS types/subtypes (13,18). In most cases the HMM scoring cut-offs for *E*-value and alignment coverage were set at 1×10^{-5} and 30%, respectively. For the single-gene Shedu system, *E*-value and alignment cut-offs were set at 1×10^{-25} and 50%, respectively. To benchmark the performance of PADLOC, we searched for Doron systems in the genomes listed in the original study and compared the accuracy of defence system recall to what was previously reported (13). To comprehensively identify Doron and CBASS defence systems in publicly available genomes, we used PADLOC to search all RefSeq v201 Archaea and Bacteria genomes ($n = 192{,}371$, July 2020) (32).

Identification of new defence system variants

To discover variants of Doron defence systems, we used a subset of RefSeq genomes ($n = 41{,}470$) with reduced redundancy, comprised of up to five genomes for each bacterial and archaeal species defined by either the NCBI or GTDB taxonomy (44). We first attempted to identify variants by searching for orphan (single) Doron system genes. This approach proved unproductive because the Doron systems often comprise proteins with nuclease, helicase and/or ATPase domains, which are abundant in prokaryotic genomes and implicated in a variety of functions. As a result, our search revealed a large number of hits to proteins unlikely

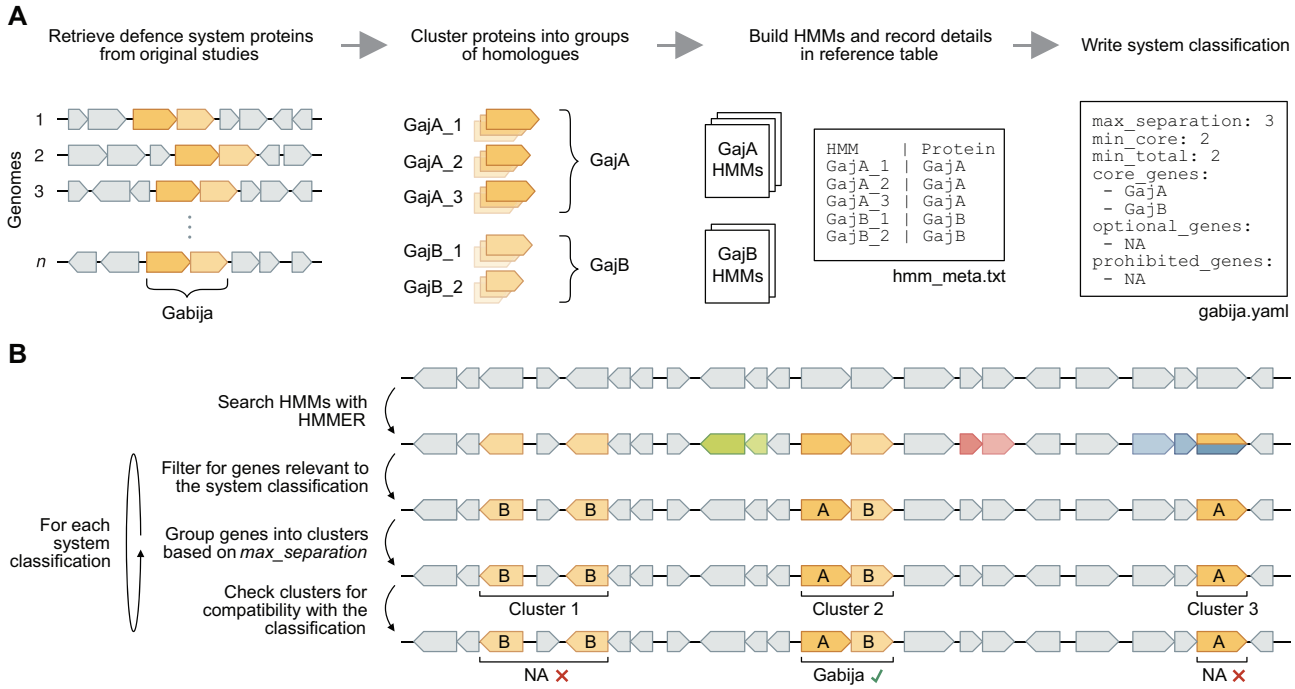


Figure 1. Workflow of data preparation and PADLOC functioning. (**A**) Preparation of data for PADLOC. For each type of defence system protein, sequences were retrieved and clustered into homologue groups. An HMM was built from each group of proteins, and the names of the HMMs (e.g. GajA_1) and their corresponding protein families (e.g. GajA) were recorded in a reference table (hmm_meta.txt), which allows a single family of defence system proteins to be represented by multiple HMMs. A simple classification file ([system].yaml) was written to represent each defence system, describing the typical genetic architecture of the system. (**B**) Automated functional workflow of PADLOC. HMMER is used to identify genes encoding defence protein homologues in the input genome. Each system classification is then analysed individually, filtering the HMM hits for genes relevant to the current type of system being searched. HMM hits are grouped into gene clusters based on the synteny requirements specified in the system classification. Each cluster is then checked against the system classification to determine whether the system requirements are fulfilled. Yellow genes represent Gabija; green, red or blue genes represent genes from other defence systems; genes with two colours (i.e. yellow/blue) represent genes matched by HMMs from two different defence systems.

to be related to defence. To reduce the number of false positives, we instead used PADLOC to search for gene clusters that encoded at least two canonical proteins from the Doron systems. Proteins that were encoded within three open reading frames (ORFs) either side of the identified gene clusters were then pooled to give a set of putative defence-associated proteins. Proteins that were already part of a Doron system were removed. We then clustered the putative defence-associated proteins into groups of homologues using MMseqs2 and built HMMs with HMMER (as above). The resulting HMMs were then grouped into larger protein families based on an all-against-all HMM-HMM comparison using HH-suite v3.1.0 (45) with cut-offs of 95% probability and 75% pairwise alignment coverage. We then calculated the frequency of association between each defence-associated protein family and canonical Doron system protein. For two proteins '*A*' and '*B*', the frequency of association was calculated as frequency = (loci encoding A and B)/(loci encoding A). To reduce bias from overrepresented loci (due to many similar genomes from closely related strains), only one representative locus was counted per distinct gene cluster (refer to Supplementary Figure S2A for more details). Defence-associated protein families were filtered for those above a threshold of >50 associations (loci) with at least two different canonical Doron system proteins with a frequency of association >0.5, and at least one with a frequency greater than 0.7; these cut-offs were determined empirically by inspection of the resulting network graphs. The remaining associations were then manually inspected for loci displaying features characteristic of defence systems, including conserved operon-like architecture, presence in diverse genetic contexts and indications of horizontal gene transfer (presence in multiple species). Candidate defence system variants that had indiscernible locus architectures or lacked context diversity were excluded from further analysis (Supplementary Figure S2B). To assign putative function to each Doron system–associated protein, pairwise comparison of their HMMs against PFAM v33 (46) and COG v1 (47) was carried out using HHpred (48).

Prevalence and phylogeny of defence systems

We analysed the prevalence of all CBASS, canonical Doron systems and candidate new system types by searching all RefSeq v201 Archaea and Bacteria genomes with PADLOC. To investigate the phylogeny of the new Doron system subtypes, we built trees of the shared core components of these systems, i.e. the proteins present in both the canonical and new subtypes. For each type of core protein, the sequences of all the proteins identified with PADLOC were clustered with MMseqs2 and filtered for the top *n* clusters containing ∼90% of the total sequences. Five random sequences were sampled from each of these clusters as representatives of each core protein. These sequences were

aligned using MUSCLE and phylogenetic trees were inferred with IQ-TREE v2.0.3 (49) using the best-fit model selected by ModelFinder (50) and ultrafast bootstrap with 1000 replicates (51).

Phenotypic analysis of defence systems

We assessed the activity of three new Doron system subtypes (Zorya type III, Hachiman type II, and Lamassu type II) *in vivo*. The systems were amplified from the genomic DNA of *Stenotrophomonas nitritireducens* DSM 12575 (NZ_LDJG01000021.1, Zorya type III), *Sphingopyxis witflariensis* DSM 14551 (NZ_NISJ01000011.1, Hachiman type II) and *Janthinobacterium agaricidamnosum* DSM 9628 (NZ_HG322949.1, Lamassu type II) using primers (Integrated DNA Technologies) listed in Supplementary Table S1 with Q5 DNA polymerase (New England Biolabs). The systems were cloned by restriction digestion with enzymes KpnI and BamHI (Zorya type III) or SbfI-HF and NotI-HF (Hachiman type II and Lamassu type II) (New England Biolabs) into a derivative of pACYCDuet-1 (pUOS001 or pUOS0014, Supplementary Table S2) amplified with primers FN0031, FN0032, FN0126, FN0127 (Supplementary Table S1) to introduce the restriction sites. After confirmation by Sanger sequencing (Eurofins Genomics), plasmids pZorya (pUOS004), pHachiman (pUOS002) and pLamassu (pUOS003) (Supplementary Table S2) were transformed into *Escherichia coli* BL21-AI (New England Biolabs), which naturally lacks the three defence systems, for subsequent phage challenge assays.

Phage propagation and plaque assays

Escherichia coli phages T1, T3, T4, T7 and Lambda-vir were obtained from the Fagenbank (Delft, Netherlands). *Salmonella* phage PVP-SE1, previously shown to infect *E. coli* strain BL21-AI (52), was kindly provided by the Azeredo Lab (University of Minho, Braga, Portugal). Phages were propagated on *E. coli* BL21-AI using the plate lysate method (53). The lysate titre was determined using the small drop plaque assay (54). For plaque assays, overnight cultures of BL21-AI containing pZorya, pHachiman or pLamassu were diluted in Lysogeny Broth (LB) supplemented with 25 μg ml^{-1} of chloramphenicol, 1 mM of IPTG and 0.2% (w/v) of L-arabinose and grown to early log phase (OD_{600} of ∼0.3) at 37°C with shaking at 200 rpm. Bacteria were mixed with LB top agar (0.6% (w/v) agar) supplemented with 1 mM IPTG and 0.2% L-arabinose, and with 10-fold serial dilutions of the phages. The mixture was poured on top of LB agar plates (1.5% (w/v) agar) and incubated at 37°C overnight. The efficiency of plaquing (EOP) was determined by comparing plaque formation in bacteria containing the defence systems with that in control bacteria with the empty vector.

Infection dynamics in liquid medium

Overnight cultures of bacteria containing a new Doron system subtype or control empty vector were diluted in LB supplemented with 25 μg ml^{-1} chloramphenicol, 1 mM IPTG, and 0.2% (w/v) L-arabinose. Cells were grown to early log phase (OD_{600} of ∼0.3), then diluted to a final OD of ∼0.1 and distributed into the wells of a 96-well plate. Phage PVP-SE1 was added to the wells at a multiplicity of infection (MOI) of 0.1 and 0.01. Infections were performed in biological triplicates, and a control without phage (MOI = 0) was used to determine normal bacterial growth. The OD_{600} was monitored every 20 min for 15 h at 37°C using a CLARIOstar Plus plate reader.

RESULTS

Defence system identification using profile HMMs and system classifications

To identify genes encoding defence system proteins, we use profile HMM-based homologue detection. HMMs are linked with their corresponding defence system proteins using a reference table (hmm_meta.txt), which includes the minimum requirements for E-value, and target/HMM alignment thresholds for each HMM (Supplementary Figure S1). Using this approach, each protein family can be represented by multiple HMMs. Each type of defence system is defined using a classification file ([system].yaml) that describes the typical genetic architecture of the system, including which genes are required and the maximum allowance for intervening non-system genes. The typical workflow for adding a new type of defence system to PADLOC involves retrieving the appropriate protein sequences, clustering based on sequence similarity, aligning and building profile HMMs, assigning the HMMs to their respective proteins, and writing a [system].yaml classification file to represent the system (Figure 1A). These data can then be used with PADLOC to search for the system. After a genome is searched for defence system genes, each system classification is analysed individually. First, the putative defence genes are filtered for those relevant to the current system classification being analysed (i.e. they have been labelled as core, optional or prohibited) (Figure 1B). Since different defence systems can comprise similar components, this pre-filtering prevents similar HMMs that belong to different system types from affecting defence system detection. The relevant genes are then grouped into gene clusters based on a maximum allowable number of unrelated genes separating them, defined by the *maximum_separation* parameter in the classification file (Figure 1B). Lastly, each gene cluster is checked against the system classification, to determine whether the system requirements are fulfilled. Gene clusters meeting the specifications are reported as encoding the corresponding defence system.

We first validated our PADLOC approach using the Doron defence systems (13). Of the systems reported in the GenBank 2016 dataset by Doron *et al.* (13), we reserved half the data as a testing set and used the remaining half to construct HMMs, then wrote PADLOC-formatted system classifications to define each system subtype. Running PADLOC over both the test and training sets demonstrated a high recall sensitivity (typically >97% recall) (Supplementary Figure S3A). We then expanded the PADLOC dataset to include all data from Doron *et al.* (13) and analysed the detection of systems by PADLOC compared to the

reported systems, which revealed a high recall and an increased sensitivity to identify several additional examples of most defence systems (Supplementary Figure S3B). For the single-gene Shedu system, we had to trade-off detection sensitivity for specificity by enforcing higher HMM scoring cut-offs, which resulted in detection of approximately 89% of the Shedu systems listed by Doron *et al.* (13) The sensitivity/specificity trade-off is a limitation of using PADLOC (or indeed any approach relying on profile HMMs to detect single proteins) to identify single gene defence systems and users should note that the accuracy of PADLOC for single-gene systems will be less than for multi-gene systems. Overall, these results demonstrate the quality of our protein models and system classifications and validate the ability of PADLOC to identify defence systems in prokaryotic genomes.

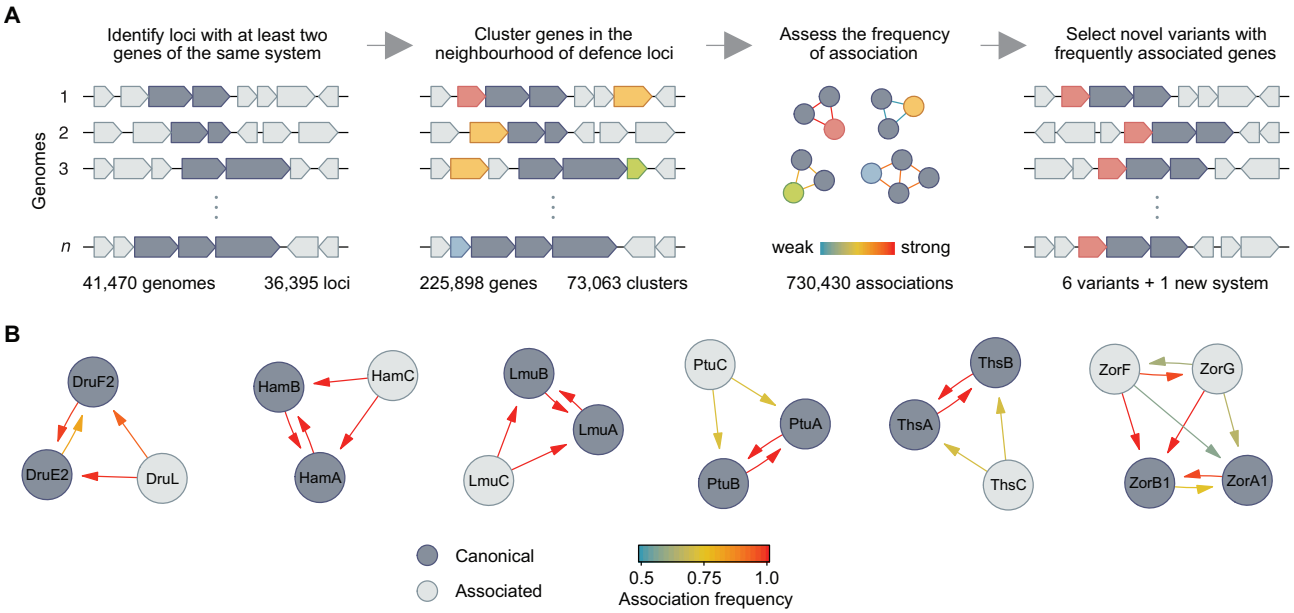
Identification of new defence system variants

Several types of defence systems have accessory proteins that can regulate, diversify or enhance the antiviral response (18,55,56). For instance, a recent analysis of CBASS systems revealed several distinct types/subtypes that share conserved components, with each type encoding several different ancillary proteins (18). Likewise, we hypothesised that additional subtypes of the Doron systems existed which had not been previously classified. To test this, we systematically identified genes that were frequently associated with each of the Doron systems (Figure 2A). First, we identified 36,395 loci comprised of co-localised genes encoding two or more proteins of the same Doron system, in a set of 41,470 representative genomes. We then clustered the 225,898 proteins encoded by the genes surrounding these defence gene clusters into 73,063 groups of homologues. We then calculated the frequency of association of each group of homologues to each Doron system gene, thereby revealing any protein families with frequent association to each type of Doron system (Figure 2B and Supplementary Figure S4). Loci containing these frequent associations were examined for features characteristic of defence systems (i.e. conserved operon-like genetic architecture, diverse genetic contexts, distribution across distantly related organisms). To further demonstrate that the additional genes of the new subtypes were associated with their respective systems, we searched for cases where the additional genes were present without their respective canonical genes (i.e. orphan occurrences of the subtype-specific genes). In general, the Doron-associated genes were identified more often as belonging to the new system subtypes than they were identified as orphan genes, and in all cases the observed associations were significant ($P < 0.001$, determined using one-sample proportion tests), suggesting functional association (Supplementary Figure S5). Additionally, genes identified as orphans generally were matched by their respective HMMs with a higher *E*-value (i.e. were weaker hits), indicating that the orphan genes were more divergent (Supplementary Figure S5). Altogether, the associations observed between the genes of these systems were robust and, using this method, we identified six putative new Doron system subtypes and an additional putative novel defence system, which we named Hma (Figure 2C).

To comprehensively identify the new Doron system subtypes and the Hma system, we wrote system classifications for PADLOC and searched for them in all RefSeq v201 Archaea and Bacteria genomes (Figure 2C). Altogether, we identified 168 instances of a Druantia-like system that, similar to Druantia type II, encodes DruE and DruF. However, the Druantia-like system lacks the type II requisite DruM and DruG proteins, instead encoding a hypothetical protein with no domain annotations, hereafter named DruL. As such, we have classified this system as a new type of Druantia, type IV. About 3.8% of Hachiman systems (379 systems) were associated with a gene encoding a DUF3223 protein (HamC) either upstream or downstream of *hamAB*. We refer to the new Hachiman systems as type II, and the original Hachiman systems as type I. Similarly, an additional hypothetical protein (LmuC) was identified in about 10.2% of the Lamassu systems detected (371 systems) (Lamassu type II). About 7.6% of Septu systems identified (1,449 systems) were preceded by a gene encoding a reverse-transcriptase (PtuC). This same gene cluster of *ptuC* (reverse-transcriptase), *ptuA* (ATPase) and *ptuB* (HNH endonuclease) was identified previously using several different approaches (14,57,58), and characterised as a retron phage defence system. Our mutual discovery of this association adds support to our method of system variant/subtype detection. Due to the similarity of this system with canonical Septu, we classified it as Septu type II. In 2.4% of Thoeris systems (167 systems), there was an additional gene encoding a histidine triad (HIT) domain protein (ThsC) (Thoeris type II). In addition to the 7,742 canonical Zorya systems found (types I and II), we identified 6,401 pairs of *zorBC* genes flanked by genes encoding a DUF3348 domain protein (ZorF) and a DUF2894 domain protein (ZorG), which we have named Zorya type III. While analysing genes associated with Doron systems, we also identified 1,638 instances of a three-gene operon that occurred frequently with Septu type I systems but was also often found elsewhere (not near Septu) in the genomes analysed. We designated this as a new candidate defence system named Hma, as it encodes three proteins with predicted helicase (HmaA), m5c methyltransferase (HmaB) and ATPase (HmaC) domains.

Doron system variants provide protection against phage infection

To determine whether the new Doron system subtypes defend against phages, we cloned representative Zorya type III, Hachiman type II and Lamassu type II systems onto plasmids for inducible expression in *E. coli*. We then challenged the bacteria with several types of phages (Siphoviridae: T1, LambdaVir; Myoviridae: T4, PVP-SE1; Podoviridae: T7) (Figure 3). In all cases, we observed significant reductions in the efficiency of plaquing for at least one phage tested for each system (Figure 3A). We also tested a phage (PVP-SE1) with liquid culture experiments at a range of different multiplicities of infection (MOI), and found that each of the systems provided protection at MOI <0.1, demonstrated by an absence of culture collapse upon infection (Figure 3B). These results confirm that Zorya type III, Hachiman type II and Lamassu type II encode functional defence systems.



C

Name	Description	Example locus	Loci (unique)
Druantia type II	DruM (methylase) + DruF (hyp) + DruG (hyp) + DruE (helicase-DUF1998)		636 (260)
Druantia type IV	DruF (hyp) + **DruL (hyp)** + DruE (helicase-DUF1998)		168 (146)
Hachiman type I	HamA (DUF1837) + HamB (helicase)		9,685 (4,279)
Hachiman type II	**HamC (DUF3223)** + HamA (DUF1837) + HamB (helicase)		379 (233)
Lamassu type I	LmuA (DUF4297) + LmuB (helicase)		3,250 (1,370)
Lamassu type II	LmuA (DUF4297) + **LmuC (hyp)** + LmuB (helicase)		371 (246)
Septu type I	PtuA (ATPase) + PtuB (HNH)		17,624 (9,416)
Septu type II	**PtuC (RT)** + PtuA (ATPase) + PtuB (HNH)		1,449 (513)
Thoeris type I	ThsA (TIR) + ThsB (Macro)		6,751 (2,199)
Thoeris type II	ThsA (TIR) + ThsB (Macro) + **ThsC (HIT)**		167 (102)
Zorya type I	ZorA (MotA), ZorB (MotB), ZorC (EH signature), ZorD (helicase-nuclease)		4,578 (1,675)
Zorya type II	ZorA (MotA) + ZorB (MotB) + ZorE (nuclease)		3,164 (888)
Zorya type III	**ZorF (DUF3348)** + ZorA (MotA) + ZorB (MotB) + **ZorG (DUF2894)**		6,401 (1,649)
Hma	HmaA (Helicase) + HmaB (Methylase) + HmaC (ATPase)		1,638 (1,152)

ATPase RT TIR Methylase DUF TM
Helicase EH signature Nuclease Macro HIT
1000 aa

Figure 2. Analysis of proteins associated with Doron systems reveals new system types. (**A**) Workflow of defence system variant identification. PADLOC was used to identify loci encoding Doron system proteins. The proteins encoded by up to three genes either side of each Doron system locus were clustered into families. The frequency of association of each protein family to each Doron system protein was analysed. Loci with frequent associations, conserved locus architecture and found in diverse genetic contexts were considered as new subtypes. (**B**) Network of defence gene associations after filtering for abundance greater than 50 distinct loci, association frequency greater than 0.5, conservation of genetic architecture and context variability. Arrow direction represents association frequency of protein '*A*' (start of arrow) with protein '*B*' (end of arrow). (**C**) Descriptions and schematic diagrams of the new Doron system types and their most similar canonical Doron system types. Domains: RT, reverse transcriptase; DUF, domain of unknown function; TIR, Toll-interleukin receptor; TM, Transmembrane; Hyp, hypothetical protein. Proposed system names: Hma; helicase, methylase and ATPase. Descriptions of proteins that differ between canonical Doron systems and the new types are shown in bold. Refer to Supplementary Table S3 for details of loci examples.

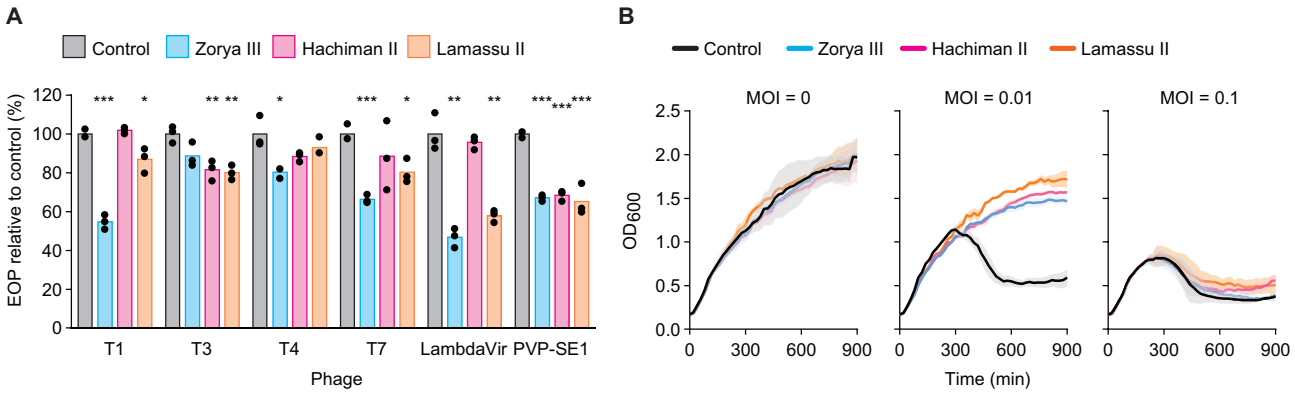


Figure 3. The new Doron system subtypes provide protection against phage infection. (**A**) The efficiency of plaquing (EOP) for *E. coli* BL21-AI possessing representative Zorya type III, Hachiman type II or Lamassu type II systems from *Stenotrophomonas nitritireducens* DSM 12575, *Sphingopyxis witflariensis* DSM 14551 and *Janthinobacterium agaricidamnosum* DSM, respectively, relative to the empty vector control. Graphs show the mean of three biological replicates with individual data points overlaid. Two-sided *t*-test; * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$. (**B**) Liquid culture infection time courses for BL21-AI strains possessing the Doron defence system variants, infected with phage PVP-SE1. Growth curves represent the mean of three biological replicates, the shaded area corresponds to the standard error of the mean.

The Doron system variants are found in multiple lineages

Defence systems are frequently transferred between prokaryotes via horizontal gene transfer (2,59,60). As a result, some defence systems are phylogenetically widespread, while others may be confined to specific taxa but show patchy distribution between closely related species (59). To further investigate the new Doron defence system subtypes, we analysed their prevalence and phylogenetic distribution in all RefSeq v201 Archaea and Bacteria. For comparison, we included the canonical Doron systems and also built HMMs and wrote PADLOC system definitions for the recently discovered CBASS systems described in Millman *et al.* (18). We identified each of our new system types in multiple phyla, as is expected for phage defence systems and observed for the canonical Doron and CBASS systems (Figure 4). Druantia type IV was the most widespread of the new Doron system subtypes, present in 11 phyla compared to the related canonical Druantia type II system, which was identified only in Proteobacteria. The putative Hma system was very widespread, present in 26 phyla, surpassed only by CBASS type I, Gabija and Septu type I. At the genus level, the defence systems also exhibited patchy distribution (Supplementary Figure S6), indicative of horizontal transfer, congruent with the function of these systems as phage defences. To determine whether the new Doron system subtypes were divergent from those of the archetypal systems, we analysed the sequence similarity of their core components (i.e. the proteins present in both the new and canonical types) (Supplementary Figure S7, Table S4). In most cases, the core components of the new subtypes were divergent from those of the canonical systems, being present in the same or closely related clans. This sequence divergence could correlate with the acquisition of the additional gene(s) followed by subsequent functional specialisation, although this remains to be determined. Overall, each new defence system subtype exhibited typical defence system characteristics including conserved operon-like genetic architecture, presence in diverse genetic contexts and distribution across distantly related organisms.

DISCUSSION

Many diverse defence systems have evolved in bacteria and archaea to defend against phages and other MGEs (1). Recently, there has been a surge in the discovery of new types of phage defence systems. However, the systematic identification and annotation of defence systems remains a challenge for biologists interested in searching the genome of their organism of interest. To address the lack in capability of current tools to identify newly discovered types of phage defence systems, we developed PADLOC. When benchmarked against the genomes searched by Doron *et al.* (13), PADLOC detected on average 97% of the multi-gene systems listed in the original study, with some additional systems detected. This demonstrates that PADLOC can identify multi-gene defence systems with high accuracy and specificity. One limitation of PADLOC is that, due to the constraint of genetic synteny, defence systems that are split by breaks in contigs will not be detected. However, this is an important trade-off in reducing false positives, firstly because HMMs detect proteins with greater sensitivity than traditional BLAST methods (38) and secondly because defence system proteins often comprise domains that are ubiquitous in other molecular systems. To aid in the identification of multi-gene systems split between contigs, we developed several relaxed system classifications (specified in [system]_other.yaml files) that require only two defence genes to be present and co-localised. The raw HMMER outputs can also be inspected, allowing users to identify potential orphan defence genes or highly divergent homologues.

Using PADLOC, we identified several clusters of Doron system genes that had strong associations with additional proteins. Based on these associations, we propose new types of Druantia, Hachiman, Lamassu, Septu, Thoeris, and Zorya systems. Septu type II was recently discovered independently and classified as a Type I-A bacterial retron (14,57,58). Members of the Type I-A retrons include Ec73 from *E. coli* and Vc95 from *Vibrio cholerae*, which provide defence against phages (14,58). Our detection of Septu type II demonstrates the capability of our approach for identification of variant defence systems. Recently, a type I Thoeris

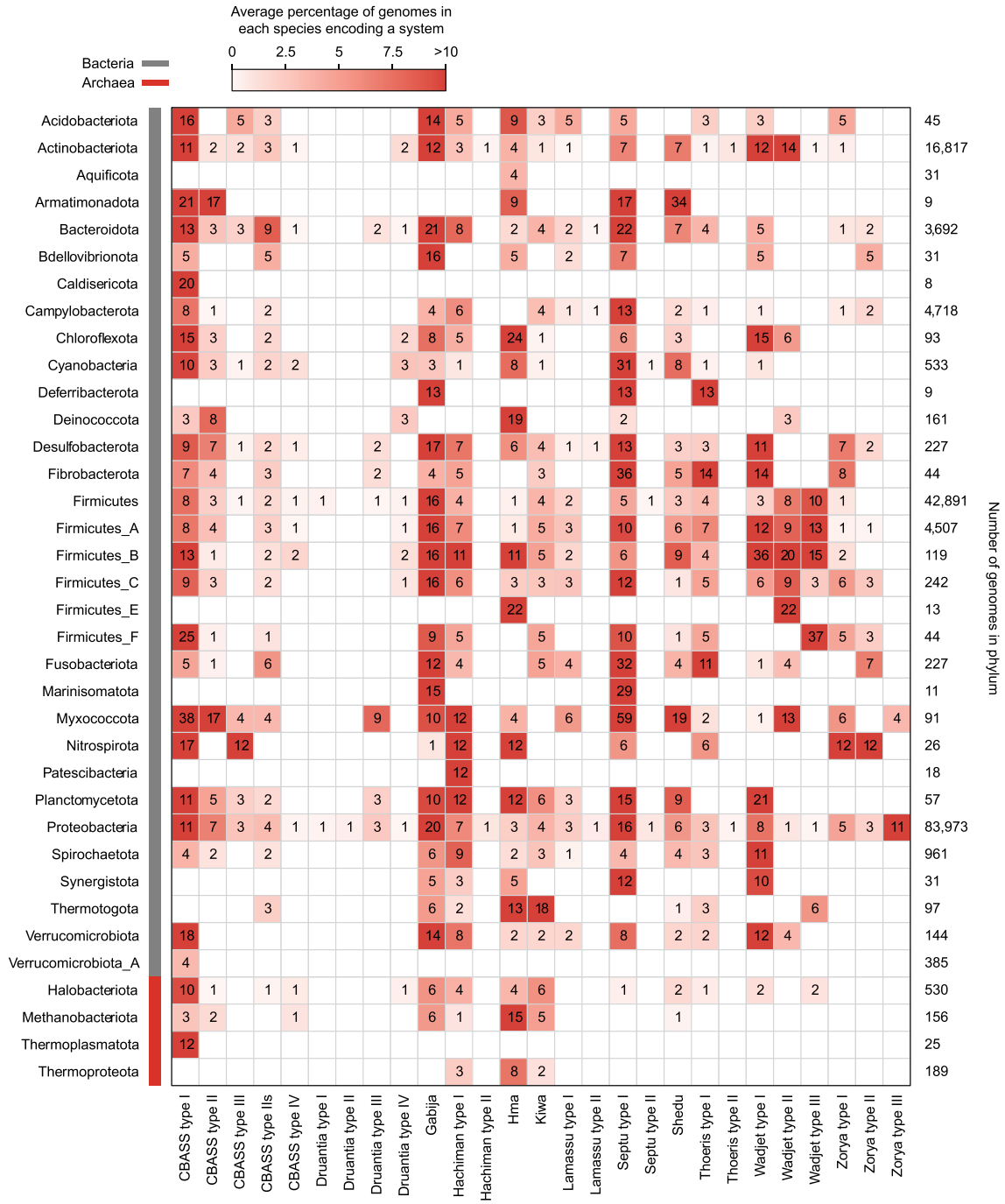


Figure 4. Abundance of defence systems identified with PADLOC in bacteria and archaea. All genomes from RefSeq v201 Archaea and Bacteria were searched with PADLOC. The values in the boxes represent, for each phylum, the average percentage of genomes in each species encoding a system, grouped using GTDB taxonomy (44); system prevalence is weighted in this way to limit biases in phyla that contain many closely related genomes of the same species. The colouring in each box provides a visual representation of these values. Shown are phyla with more than five genomes and at least one type of system. A species-level comparison is provided in Supplementary Figure S6 and the full data are provided in Supplementary Table S5.

defence system, comprised of ThsA and ThsB, was shown to generate an isomer of cyclic adenosine diphosphate ribose (v-cADPR) from NAD^+ in response to phage infection (61,62). It is proposed that v-cADPR is a second messenger that triggers further degradation of NAD^+ by ThsB to induce cell death (62). As a putative HIT family nucleotide hydrolase/transferase, we hypothesise that ThsC of the newly identified Thoeris type II systems might play a role in the formation or degradation of the v-cADPR second messenger to regulate NAD^+ degradation, perhaps as an off-switch analogous to the RING nucleases associated with some type III CRISPR-Cas systems that use cyclic oligonucleotide signalling (63–65). ZorA and ZorB from Zorya systems share sequence similarity with the inner

membrane flagella motor proteins MotA and MotB, respectively (13). However, ZorAB are not sufficient for defence and it has been proposed that a ZorAB complex forms a proton channel that facilitates abortive infection, whereas ZorC, ZorD, and ZorE perform additional essential roles as phage sensors or activators of ZorAB (13). Since our data demonstrate activity of the Zorya type III system comprised of ZorA, ZorB, ZorF and ZorG, we propose that ZorF and ZorG function are regulators of ZorAB activity in place of ZorC, ZorD and ZorE. From the other new system types we identified, DruL, HamC, and LmuC comprise domains of unknown function. An NMR structure for the HamC protein of *Rhodospirillum rubrum* ATCC 11170 has been solved (PDB ID: 2K0M; DOI: 10.2210/pdb2K0M/pdb), with similar topology to Nuclear Transport Factor 2 (66). However, the function of HamC in phage defence remains unknown. Altogether, the data presented here extend the spectrum of potential defence systems and provide a foundation for further experimental study of their mechanisms.

The discovery of new defence systems is progressing rapidly, and importantly PADLOC can be updated to incorporate these systems as they are characterised. Using our modular approach to the organisation of HMMs and system classifications, defence systems can be easily added or updated as required. For greater accessibility, we have also developed a PADLOC webserver that allows users to analyse their genomes of choice or browse a pre-computed database of defence systems identified in RefSeq genomes. PADLOC is an open-source project, with code, HMMs, and system classifications available on GitHub. Additional curation of high quality HMMs for additional defence systems will be required to establish PADLOC as a comprehensive resource for defence system identification. We encourage the community to submit new defence system data for addition to the PADLOC database.

DATA AVAILABILITY

The defence systems identified in this study can be viewed on the PADLOC webserver (https://padloc.otago.ac.nz). Additional genomes can be searched for defence systems by submitting them on the webserver or by downloading PADLOC from GitHub and running the software locally (https://github.com/padlocbio/padloc).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Chris Palmer (Department of Mathematics and Statistics, University of Otago) and members of the Information Technology Services Division at the University of Otago for assistance in establishing the PADLOC webserver. We thank members of the Fineran laboratory for helpful discussions. We acknowledge the use of the New Zealand eScience Infastructure (NeSI) high-performance computing facilities in this research. NeSI's facilities are provided by and funded jointly by NeSI's collaborator institutions and through the Ministry of Buisness, Innovation and Employment's Research Infrastructure programme.

FUNDING

Royal Society of New Zealand Te Apārangi (RSNZ) Marsden Fund; School of Biomedical Sciences Bequest Fund from the University of Otago; L.J.P. was supported by a University of Otago Doctoral Scholarship. Funding for open access charge: Laboratory research funding.
Conflict of interest statement. None declared.

REFERENCES

1. 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.
2. 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.
3. Samson,J.E., Magadán,A.H., Sabri,M. and Moineau,S. (2013) Revenge of the phages: defeating bacterial defences. *Nat. Rev. Microbiol.*, **11**, 675–687.
4. Davidson,A.R., Lu,W.-T., Stanley,S.Y., Wang,J., Mejdani,M., Trost,C.N., Hicks,B.T., Lee,J. and Sontheimer,E.J. (2020) Anti-CRISPRs: protein inhibitors of crispr-cas systems. *Annu. Rev. Biochem.*, **89**, 309–332.
5. Anzalone,A.V., Koblan,L.W. and Liu,D.R. (2020) Genome editing with CRISPR–Cas nucleases, base editors, transposases and prime editors. *Nat. Biotechnol.*, **38**, 824–844.
6. Hegge,J.W., Swarts,D.C. and Oost,J. (2018) Prokaryotic Argonaute proteins: novel genome-editing tools? *Nat. Rev. Microbiol.*, **16**, 5–11.
7. Loenen,W.A.M., Dryden,D.T.F., Raleigh,E.A., Wilson,G.G. and Murray,N.E. (2014) Highlights of the DNA cutters: a short history of the restriction enzymes. *Nucleic Acids Res.*, **42**, 3–19.
8. 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.
9. 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.
10. Cohen,D., Melamed,S., Millman,A., Shulman,G., Oppenheimer-Shaanan,Y., Kacen,A., Doron,S., Amitai,G. and Sorek,R. (2019) Cyclic GMP–AMP signalling protects bacteria against viral infection. *Nature*, **574**, 691–695.
11. 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.
12. 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.
13. 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.
14. 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.
15. Makarova,K.S., Haft,D.H., Barrangou,R., Brouns,S.J.J., Charpentier,E., Horvath,P., Moineau,S., Mojica,F.J.M., Wolf,Y.I., Yakunin,A.F. *et al.* (2011) Evolution and classification of the CRISPR–Cas systems. *Nat. Rev. Microbiol.*, **9**, 467–477.
16. Makarova,K.S., Wolf,Y.I., Alkhnbashi,O.S., Costa,F., Shah,S.A., Saunders,S.J., Barrangou,R., Brouns,S.J.J., Charpentier,E., Haft,D.H. *et al.* (2015) An updated evolutionary classification of CRISPR–Cas systems. *Nat. Rev. Microbiol.*, **13**, 722–736.
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. 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.
19. Atack,J.M., Guo,C., Litfin,T., Yang,L., Blackall,P.J., Zhou,Y. and Jennings,M.P. (2020) Systematic analysis of REBASE identifies

numerous type I restriction-modification systems with duplicated, distinct hsds specificity genes that can switch system specificity by recombination. *mSystems*, **5**, e00497–20.
20. 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.
21. 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.
22. 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.
23. Crawley,A.B., Henriksen,J.R. and Barrangou,R. (2018) CRISPRdisco: an automated pipeline for the discovery and analysis of CRISPR-Cas Systems. *CRISPR J.*, **1**, 171–181.
24. Grissa,I., Vergnaud,G. and Pourcel,C. (2007) CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. *Nucleic. Acids. Res.*, **35**, W52–W57.
25. Padilha,V.A., Alkhnbashi,O.S., Shah,S.A., de Carvalho,A.C.P.L.F. and Backofen,R. (2020) CRISPRcasIdentifier: machine learning for accurate identification and classification of CRISPR-Cas systems. *GigaScience*, **9**, giaa062.
26. 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.
27. 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.
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. 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.
30. Zhang,Y., Zhang,Z., Zhang,H., Zhao,Y., Zhang,Z. and Xiao,J. (2019) PADS Arsenal: a database of prokaryotic defense systems related genes. *Nucleic Acids Res.*, **48**, D590–D598.
31. Akarsu,H., Bordes,P., Mansour,M., Bigot,D.-J., Genevaux,P. and Falquet,L. (2019) TASmania: a bacterial Toxin-Antitoxin systems database. *PLoS Comput. Biol.*, **15**, e1006946.
32. 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.
33. 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. U.S.A.*, **111**, 13264–13271.
34. 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.
35. Chen,I.-M.A., Chu,K., Palaniappan,K., Pillay,M., Ratner,A., Huang,J., Huntemann,M., Varghese,N., White,J.R., Seshadri,R. *et al.* (2019) IMG/M v.5.0: an integrated data management and comparative analysis system for microbial genomes and microbiomes. *Nucleic Acids Res.*, **47**, D666–D677.
36. Seemann,T. (2014) Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, **30**, 2068–2069.
37. 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.
38. Eddy,S.R. (2011) Accelerated profile HMM searches. *PLoS Comput. Biol.*, **7**, e1002195.
39. Shen,W., Le,S., Li,Y. and Hu,F. (2016) SeqKit: a cross-platform and ultrafast toolkit for FASTA/Q file manipulation. *PLoS One*, **11**, e0163962.
40. Steinegger,M. and Söding,J. (2017) MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.*, **35**, 1026–1028.
41. Li,W., Jaroszewski,L. and Godzik,A. (2001) Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics*, **17**, 282–283.
42. Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.
43. Katoh,K. and Standley,D.M. (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.*, **30**, 772–780.
44. 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.
45. Steinegger,M., Meier,M., Mirdita,M., Vöhringer,H., Haunsberger,S.J. and Söding,J. (2019) HH-suite3 for fast remote homology detection and deep protein annotation. *BMC Bioinformatics*, **20**, 473.
46. 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.
47. 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.
48. 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.
49. Minh,B.Q., Schmidt,H.A., Chernomor,O., Schrempf,D., Woodhams,M.D., von Haeseler,A. and Lanfear,R. (2020) IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.*, **37**, 1530–1534.
50. Kalyaanamoorthy,S., Minh,B.Q., Wong,T.K.F., Haeseler,A. and Jermiin,L.S. (2017) ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods*, **14**, 587–589.
51. Hoang,D.T., Chernomor,O., von Haeseler,A., Minh,B.Q. and Vinh,L.S. (2018) UFBoot2: Improving the ultrafast bootstrap approximation. *Mol. Biol. Evol.*, **35**, 518–522.
52. Santos,S.B., Kropinski,A.M., Ceyssens,P.-J., Ackermann,H.-W., Villegas,A., Lavigne,R., Krylov,V.N., Carvalho,C.M., Ferreira,E.C. and Azeredo,J. (2011) Genomic and proteomic characterization of the broad-host-range Salmonella phage PVP-SE1: creation of a new phage genus. *J. Virol.*, **85**, 11265–11273.
53. Fortier,L.-C. and Moineau,S. (2009) Phage production and maintenance of stocks, including expected stock lifetimes. In: Clokie,M.R.J. and Kropinski,A.M. (eds). *Bacteriophages: Methods and Protocols*. Isolation, Characterization, and Interactions, Methods in Molecular Biology™. Humana Press, Totowa, NJ, Vol. **1**, pp. 203–219.
54. Mazzocco,A., Waddell,T.E., Lingohr,E. and Johnson,R.P. (2009) Enumeration of bacteriophages using the small drop plaque assay system. In: Clokie,M.R.J. and Kropinski,A.M. (eds). *Bacteriophages: Methods and Protocols*. Isolation, Characterization, and Interactions, Methods in Molecular Biology™. Humana Press, Totowa, NJ, Vol. **1**, pp. 81–85.
55. 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. U.S.A.*, **115**, E5307–E5316.
56. Makarova,K.S., Anantharaman,V., Grishin,N.V., Koonin,E.V. and Aravind,L. (2014) CARF and WYL domains: ligand-binding regulators of prokaryotic defense systems. *Front. Genet.*, **5**, 102.
57. Mestre,M.R., González-Delgado,A., Gutíerrez-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.
58. 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.
59. Bernheim,A. and Sorek,R. (2019) The pan-immune system of bacteria: antiviral defence as a community resource. *Nat. Rev. Microbiol.*, **18**, 113–119.

60. Houte,S., Buckling,A. and Westra,E.R. (2016) Evolutionary ecology of prokaryotic immune mechanisms. *Microbiol. Mol. Biol. Rev.*, **80**, 745–763.
61. Ka,D., Oh,H., Park,E., Kim,J.-H. and Bae,E. (2020) Structural and functional evidence of bacterial antiphage protection by Thoeris defense system via NAD + degradation. *Nat. Commun.*, **11**, 2816.
62. Ofir,G., Herbst,E., Baroz,M., Cohen,D., Millman,A., Doron,S., Tal,N., Malheiro,D.B.A., Malitsky,S., Amitai,G. *et al.* (2021) Antiviral activity of bacterial TIR domains via signaling molecules that trigger cell death. bioRxiv doi: https://doi.org/10.1101/2021.01.06.425286, 06 January 2021, preprint: not peer reviewed.
63. Athukoralage,J.S., Rouillon,C., Graham,S., Grüschow,S. and White,M.F. (2018) Ring nucleases deactivate type III CRISPR ribonucleases by degrading cyclic oligoadenylate. *Nature*, **562**, 277–280.
64. Kazlauskiene,M., Kostiuk,G., Venclovas,Č., Tamulaitis,G. and Siksnys,V. (2017) A cyclic oligonucleotide signaling pathway in type III CRISPR-Cas systems. *Science*, **357**, 605–609.
65. Niewoehner,O., Garcia-Doval,C., Rostøl,J.T., Berk,C., Schwede,F., Bigler,L., Hall,J., Marraffini,L.A. and Jinek,M. (2017) Type III CRISPR–Cas systems produce cyclic oligoadenylate second messengers. *Nature*, **548**, 543–548.
66. Sillitoe,I., Bordin,N., Dawson,N., Waman,V.P., Ashford,P., Scholes,H.M., Pang,C.S.M., Woodridge,L., Rauer,C., Sen,N. *et al.* (2021) CATH: increased structural coverage of functional space. *Nucleic Acids Res.*, **49**, D266–D273.