NAR Genomics and Bioinformatics, 2024, **6**, lqae105
https://doi.org/10.1093/nargab/lqae105
Advance access publication date: 20 August 2024
Standard Article

Distribution of specific prokaryotic immune systems correlates with host optimal growth temperature

Lisa H. Olijslager, Dolf Weijers and Daan C. Swarts*

Laboratory of Biochemistry, Wageningen University, Wageningen, Stippeneng 4, 6708WE, the Netherlands
*To whom correspondence should be addressed. Email: daan.swarts@wur.nl

Abstract

Prokaryotes encode an arsenal of highly diverse immune systems to protect themselves against invading nucleic acids such as viruses, plasmids and transposons. This includes invader-interfering systems that neutralize invaders to protect their host, and abortive-infection systems, which trigger dormancy or cell death in their host to offer population-level immunity. Most prokaryotic immune systems are found across different environments and prokaryotic phyla, but their distribution appears biased and the factors that influence their distribution are largely unknown. Here, we compared and combined the prokaryotic immune system identification tools DefenseFinder and PADLOC to obtain an expanded view of the immune system arsenal. Our results show that the number of immune systems encoded is positively correlated with genome size and that the distribution of specific immune systems is linked to phylogeny. Furthermore, we reveal that certain invader-interfering systems are more frequently encoded by hosts with a relatively high optimum growth temperature, while abortive-infection systems are generally more frequently encoded by hosts with a relatively low optimum growth temperature. Combined, our study reveals several factors that correlate with differences in the distribution of prokaryotic immune systems and extends our understanding of how prokaryotes protect themselves from invaders in different environments.

Introduction

Prokaryotes are under constant threat from mobile genetic elements (MGEs), including transposons, plasmids and viruses (1). To protect prokaryotes against MGEs, an arsenal of immune systems has evolved (2) and, in turn, various mechanisms to escape immunity have evolved in MGEs (3,4). The consequential evolutionary arms race between prokaryotes and their invaders has resulted in an extreme diversification of immune systems, yielding > 100 distinct immune system families, with more still being discovered (3). Although these immune system families rely on highly divergent mechanisms to achieve immunity, they generally adhere to one of two strategies: (i) invader interference (hereafter referred to as Invi) or (ii) abortive infection (hereafter referred to as Abi). Invi systems protect the host by neutralizing the invader, for example by degrading the invader DNA. Notable examples of Invi systems include restriction–modification (RM) systems (5) and most CRISPR-Cas [regularly interspaced palindromic repeats (CRISPR)/CRISPR-associated protein] systems (6). In contrast, Abi systems cause metabolic arrest or trigger cell death in the host cells to prevent propagation and spread of MGEs, thereby providing population-level immunity (7). Examples include various AbiX systems (where X indicates the specific Abi system), CBASS systems and short prokaryotic Argonaute (pAgo) systems (8–15). Prokaryotes typically encode a combination of Invi and Abi systems to provide immunity against different MGEs (16).

Fuelled by both fundamental curiosity and the successful repurposing of several prokaryotic immune systems as molecular tools in practical applications (17,18), there has been an increasing interest in the identification and characterization of prokaryotic immune systems. Together with the expanding availability of (meta)genomic data and newly developed methods to identify putative immune systems, this has led to the discovery and characterization of numerous novel prokaryotic immune systems (19–22). To facilitate rapid identification of known immune systems in (meta)genomes, bioinformatics tools have been developed, including CRISPRCasFinder (23), CRISPRCasTyper (24), Prokaryotic Antiviral Defense Locator (PADLOC) (25) and DefenseFinder (16). These tools typically rely on similar search strategies: they first identify genes that are involved in prokaryotic immunity through HMM-profile-based searches and subsequently identify immune systems by evaluating whether a genomic locus satisfies the genetic architecture of the immune system by scoring essential, optional and prohibited genes. However, specific search parameters vary per tool, which could result in the different tools identifying distinct immune systems in the same dataset.

Although prokaryotic immune systems are widely distributed in nature (16,26,27) and are frequently subject to horizontal gene transfer (HGT) between different prokaryotic species (28–31), they are not evenly distributed (16,26,27). We hypothesized that in specific prokaryotes, due to physiological and/or environmental factors, certain immune systems provide a larger selective advantage than other immune systems. Consequentially, an uneven distribution of prokaryotic immune systems should exist in which the immune system abundance correlates with specific physiological

Received: May 23, 2024. **Revised:** July 15, 2024. **Editorial Decision:** August 1, 2024. **Accepted:** August 2, 2024

and/or environmental factors. Indeed, previous studies have revealed differences in the total number of immune systems per genome based on phylogeny, genome size, co-localization of prophages, lifestyle and habitat (16,27). Other studies have shown that CRISPR-Cas subtype distribution is distinct depending on their environment (24,26), and that the overall abundance of CRISPR-Cas systems is higher in prokaryotic hosts with a high optimal growth temperature (from hereon: T_{opt}) (28–32). In contrast, for RM systems, both a weak positive and a weak negative correlation between abundance and T_{opt} have been reported, with both observations being only marginally statistically significant (31,32). To our knowledge, no systematic studies on the correlation between host T_{opt} and abundance of other prokaryotic immune systems have hitherto been reported.

Here, we combined DefenseFinder and PADLOC to identify prokaryotic immune systems in the species representative of the Genome Taxonomy database (SR-GTDB) (33). Their combined output reveals that immune system distribution varies between different prokaryotic phyla and that the total number of immune systems encoded per genome is positively correlated with genome size. Using a previously generated dataset (32) in which SR-GTDB genomes are linked to host T_{opt}, correlations between host T_{opt} and immune system abundance were analysed. This reveals that certain Invi systems and most Abi systems are, respectively, more and less abundant in genomes of hosts that thrive at higher temperatures. Analyses of metagenomic datasets confirms this and reveals similar correlations for other environmental parameters. The data presented in this study show that the general strategy of specific immune systems affects their distribution, and thereby suggest that Abi and Invi systems provide different fitness gains in distinct environments.

Materials and methods

Identification of prokaryotic immune systems

The species representatives in the Genome Taxonomy database (SR-GTDB) were used to obtain a broad phylogenetic and non-redundant sampling of prokaryotic species (4906 bacteria and 291 archaea) (33). Only accessions described as 'complete genome' or 'chromosome' (33) as described by Lan *et al.* (32) were used. The genomic sequences were retrieved as fasta files from ftp://ftp.ncbi.nlm.nih.gov/genomes/ (accessed November 2022). An automated script was used to predict the prokaryotic immune systems in these genomes (dx.doi.org/10.6084/m9.figshare.24632142 and https://github.com/LOlijslager/find_prokaryotic_immune_systems). In brief, the script predicts the encoded protein sequences using Prodigal (version 2.6.3 used) (34). Subsequentially, the script mines the resulting proteomes for prokaryotic immune systems using DefenseFinder (version 1.0.7; immune system models downloaded on June 9 2022) (16) and PADLOC (version 1.1.0; PADLOC database 1.4.0) (25), supported by HMMER version 3.3.1 (35) and MacSyFinder version e2.0rc6 (36). From the output, the script compares and combines the identified immune system families (Cas, RM, DISARM, etc.) by each of the tools. Because of the uncertain family identification of PADLOC systems denoted as 'GAO' in the PADLOC version used, these systems were omitted.

The combined output contains (i) immune systems comprising the same genes, classified identically by both tools; (ii) immune systems classified identically by both tools, but with one tool identifying more genes than the other; (iii) different immune systems identified by each tool, with one or more genes overlapping; (iv) immune systems comprising the same genes, but classified as different immune systems; and (v) immune systems uniquely identified by one of the two tools. In cases (i) and (v) there is no conflict, while in cases (ii), (iii) and (iv) there is a conflict that needs to be resolved (Supplementary Figure S1).

In case (ii), the script automatically keeps the system identified comprising the highest number of genes. In case (iii), it cannot automatically be determined if the (partially) overlapping systems share genes, are a hybrid system or are a single system to which PADLOC and DefenseFinder attribute a different gene set. To reduce redundancy in the identified systems, the script keeps one system while the other system is disregarded. In case (iv), the script will provide a warning for the user to see if the reference file (in which immune system classifications for PADLOC and DefenseFinder are listed) might need to be updated. After that, the script keeps one system, while the other system is disregarded. If this conflict (iv) occurs within one identification tool (i.e. DefenseFinder annotates the same genes as multiple immune systems), instead only the system is kept which is identified by the other tool. If the other tool did not identify this particular system, the system annotation is merged. For this manuscript, in conflict case (iii) and (iv), DefenseFinder output was always chosen as conflict winner. However, conflict winner can be adjusted in the script based on the user needs.

The script can be used in broad-identification mode or in high-confidence mode. In the broad-identification mode, all immune systems identified are kept, while in the high-confidence mode, the program keeps only the systems identified and annotated identically by both DefenseFinder and PADLOC. In order to obtain a broad identification of as many immune systems as possible, the broad-identification mode was used for this study. Data output can be found in Supplementary data S1 and S2, and raw data can be found on dx.doi.org/10.6084/m9.figshare.24632142.

As Defensefinder version 1.0.7 does not search for pAgos, and PADLOC version 1.1.0 uses non-canonical classification of pAgos (37,38), pAgos identified by PADLOC were reclassified for analysis. To this end, PADLOC pAgo types I, II, III and solo were reclassified as, respectively, effector-enzyme-associated long pAgos, short pAgos, PIWI-RE systems and stand-alone long pAgos.

Metagenomes (Supplementary data S3) used for metagenomics analyses were downloaded from ftp://ftp.ncbi.nlm.nih.gov/genomes/genbank/metagenomes/ or from specific metagenomic databases (39–41) as DNA fasta files (accessed May 2023). The metagenomic databases were analysed using the automated code as described above. Raw data can be found in Supplementary data S4.

Determination of host phyla and T_{opt}

For the SR-GTDB genomes, phyla were determined using the NCBI Taxonomy Database (42). Genome sizes were obtained by counting the number of nucleotides in the fasta file using SeqIO in Biopython v1.8 (43), including plasmids and additional chromosomes when these were present. For each

genome, the associated host T_{opt} was obtained from (32) for the SR-GTDB. Genome accessions in this database contain up to three listed T_{opt} values: (i) strain-specific T_{opt} values from the literature; (ii) T_{opt} values of related strains from the literature; and (iii) T_{opt} values predicted using the machine-learning method Tome v. 1.0.0 (44). We used the highest confidence T_{opt} available (i > ii > iii) for each genomic accession.

Quantification and statistical analysis

All statistical analyses were performed in Excel. Categorical comparisons were done using a χ^2 test of independence and verified using the Bonferroni test with an $\alpha = 0.05$/number of comparisons. Population-level distributions were compared using a two-tailed *t*-test assuming unequal variance. Correlations were determined with a Pearson correlation analysis. Bar graphs, scatter plots, pie charts and heat maps were made using Microsoft Excel and Adobe Illustrator, with the exception of the Sankey diagram [made using SankeyMATIC (https://sankeymatic.com/)], raincloud and bee-swarm plots (made using python package Plotly version 5.14.0) and the Venn diagram (made using the function venn2 of the python package matplotlib-venn version 0.11.7).

Results

Combining PADLOC and DefenseFinder increases immune system discovery rate

To investigate factors that influence the distribution of immune system families, we created an automated script (https://github.com/LOlijslager/find_prokaryotic_immune_systems and dx.doi.org/10.6084/m9.figshare.24632142) that identifies prokaryotic immune systems from (meta)genomic sequences by combining PADLOC (25) and DefenseFinder (16) output and generating a unified database. To obtain a broad phylogenetic and non-redundant genome dataset, we used the genomes of species representatives in the Genome Taxonomy Database (SR-GTDB) (33). While DefenseFinder and PADLOC identified a similar number of immune systems in total (DefenseFinder: 32 381; PADLOC: 29 405), only 36% (16 279 out of 45 507) of the immune systems identified in total were identified by both tools (Figure 1A; Supplementary Figure S1).

The limited overlap between PADLOC and DefenseFinder is partially explained by the fact that certain immune system families are only being searched for by one of the tools (4 823 of the total; e.g. DefenseFinder v1.0.7 does not search for pAgo systems, and PADLOC v1.1.0 does not search for Mokosh). Furthermore, conflicts in identification exist due to various reasons, for example when immune systems comprised of the same genes are identified by both tools, but are classified as distinct immune system families, or when immune systems (partially) overlap with genes of other immune systems identified by the other tool. By using information from both tools (see the Materials and methods and Supplementary Figure S1) redundancy between systems differentially identified and/or classified by PADLOC and DefenseFinder was removed. Combined, our search strategy resulted in the identification of 40 598 immune systems, significantly more than identified by DefenseFinder (32 381) or PADLOC (29 405) alone. In the remaining database, only 466 systems with partial overlap remain. While manual curation might further reduce redundancy, as these 466 systems only comprise a small fraction of the immune system database (1.15%) and might represent hybrid immune systems, we have chosen to include them in our final immune system database.

Both DefenseFinder (12 166) and PADLOC (8 244) uniquely identify immune systems not identified by the other tool (in total 20 410 of 40 598). Further investigation of these uniquely identified systems reveals that these consist of nearly every immune system family for both tools (Supplementary Figure S2). In addition, DefenseFinder identified substantially more AbiH, DISARM, Lamassu, Nhi and PARIS systems than PADLOC, while PADLOC identified more AbiE, DRT, dXTPases, Iet, PT, Vipirin and Zorya systems than DefenseFinder (Figure 1B; Supplementary Figure S2). The tools did not show substantial differences in the identification of immune systems in specific phyla (Supplementary Figure S3). In conclusion, combining PADLOC and DefenseFinder significantly increases the discovery rate of putative immune systems in (meta)genomic datasets and thereby provides an expanded view of the immune system arsenal of prokaryotes.

Distribution of immune systems in different phyla

To further investigate the distribution of prokaryotic immune systems, we used the database of non-redundant immune systems identified in the SR-GTDB. First, we investigated the distribution of immune systems over the distinct bacterial and archaeal phyla (Figure 2A). While immune systems are regularly horizontally transferred within and between bacterial and archaeal phyla (28–31), their distribution over these domains of life is unequal (Figure 2A; Supplementary Figure S4). Also within each of the domains, certain systems show a patchy distribution over the different phyla (Figure 2A). Examples of the latter include Wadjet systems, which are mainly found in Actinobacteria, and RM systems, which are less abundant in Crenarchaeota (Figure 2A; Supplementary Figure S4).

Next, we investigated if there are differences in the total number of immune systems encoded by prokaryotes belonging to each of the phyla. Bacteria from most phyla on average encode between 7.1 and 8.8 immune systems (Figure 2B). However, Cyanobacteria and Tenericutes encode on average 16.8 and 2.9 immune systems, respectively (Figure 2B). Archaea generally encode fewer immune systems than bacteria, on average between 4.2 and 6.8 immune systems (Figure 2B). It was previously shown that the total number of immune systems encoded and genome size are positively correlated (16). Indeed, Tenericutes and archaea, which encode relatively few immune systems (2.9 and 5.9, respectively, on average), also have relatively small genomes (1.3–2.8 Mbp on average), while Cyanobacteria, which encode a relatively high number of immune systems (16.8 on average), usually have relatively large genomes (5 Mbp on average) (Figure 2B, C). Corroborating that the number of immune systems and genome size are positively correlated, a highly significant positive linear correlation between genome size and the average number of immune systems encoded exists in both bacteria and archaea (Figure 2D; Pearson correlation bacteria: moderate positive coefficient 0.43, $P < 10^{-99}$; archaea: weak positive coefficient 0.38, $P < 10^{-10}$). This indicates that while there is a correlation between genome size and average number of immune systems encoded, other factors also affect how many immune systems are encoded.

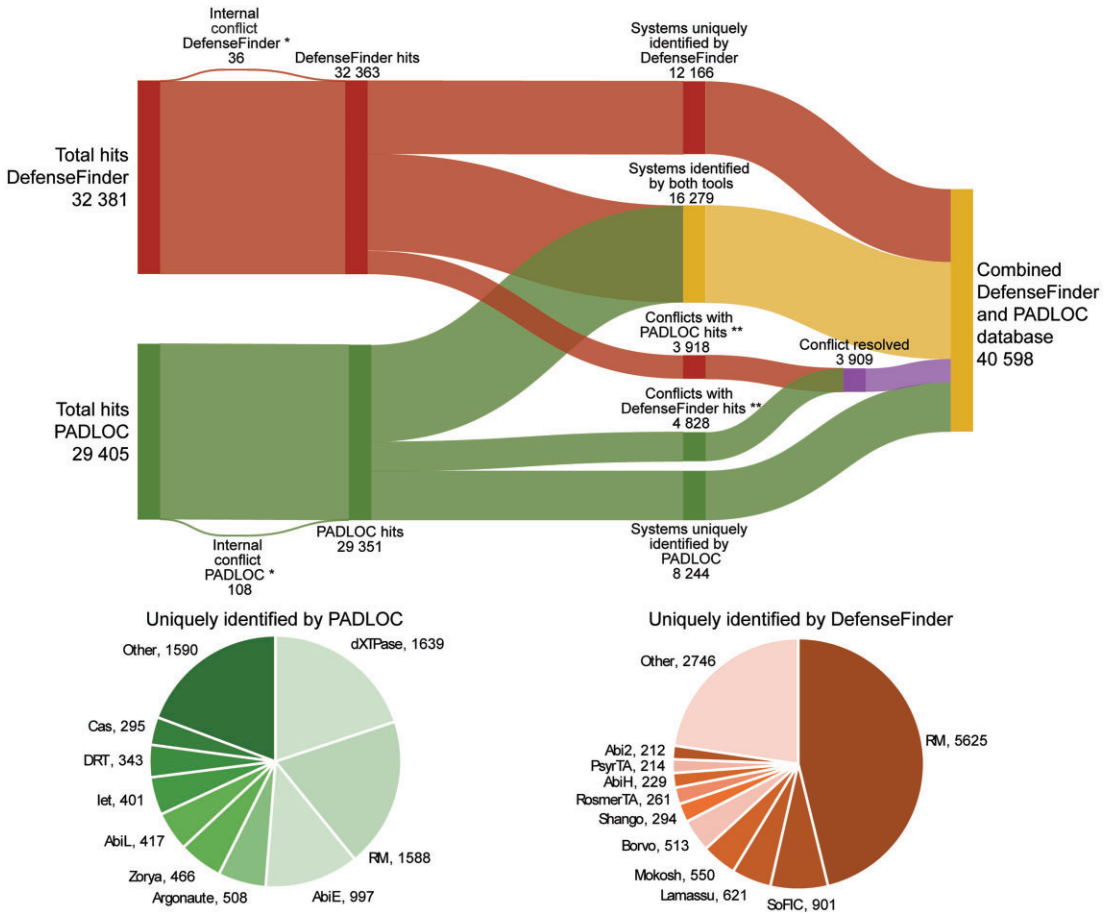


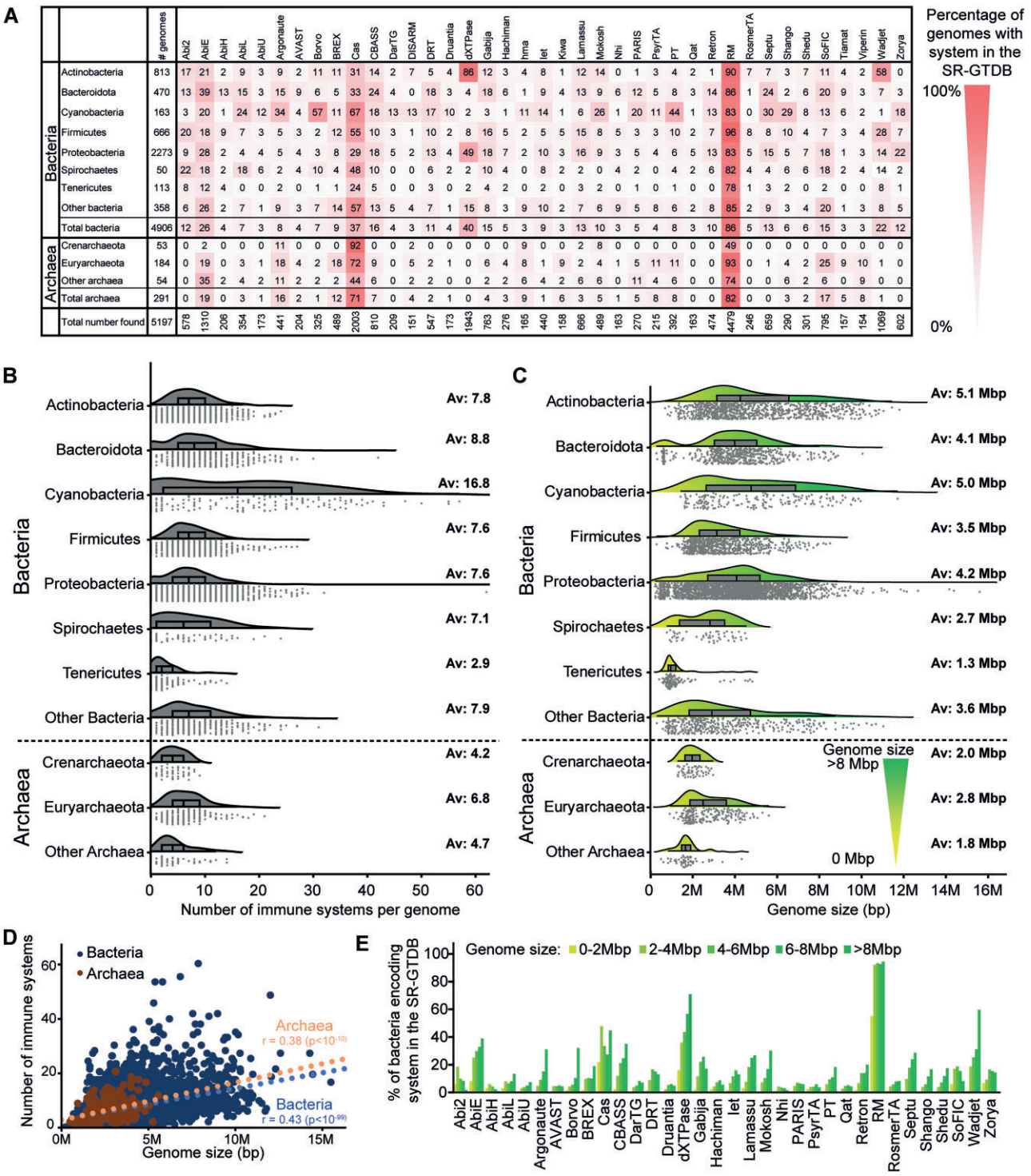
Figure 1. (**A**) Sankey diagram visualizing DefenseFinder and PADLOC prokaryotic immune system hits in the species representatives of the Genome Taxonomy Database (SR-GTDB). The diagram visualizes how all identified systems were combined into a single database (further details about how redundancy was removed can be found in Supplementary Figure S1 and the Materials and methods). *Prokaryotic immune systems identified by DefenseFinder or PADLOC that overlap completely with another system identified by the same tool. **Prokaryotic immune systems identified by PADLOC or DefenseFinder that share (partial or complete) overlap with a system identified by the other tool. In the combined database, 466 immune systems with partial overlap (possibly representing hybrid systems) remain. (**B**) Pie chart visualizing prokaryotic immune systems uniquely identified by PADLOC (left) or DefenseFinder (right). Prokaryotic immune systems uniquely identified < 200 times were grouped under 'Other'.

To investigate if there is a correlation between abundance and genome size for individual immune system families, we analysed this for each immune system family which was identified in > 150 genomes in our dataset. The positive correlation between abundance and genome size appears to exist for most immune system families (Figure 2E; Supplementary Figure S5A). However, certain immune system families show a stronger increase in abundance with increasing genome size than others (e.g. dXTPases; Figure 2E). In contrast, for other immune system families, the increase in abundance with increasing genomes is marginal (e.g. AVAST, CRISPR-Cas and RM) or even appears to decrease (e.g. Abi2, AbiH and Nhi). Combined, this shows that genome size is an important determinant for the total number of immune systems encoded, but not for each immune system family individually.

Distribution of immune systems in different T_{opt} categories

Based on previous reports that CRISPR-Cas systems are more abundant in hosts with a higher T_{opt} (28–32), we investigated if the abundance of immune systems correlates with host T_{opt} in general. Based on previously determined host T_{opt} values (32), all genomes in the SR-GTDB were grouped in five T_{opt} categories: psychrophiles (4–20°C), psychrotolerants (20–30°C), mesophiles (30–45°C), thermophiles (45–85°C) and hyperthermophiles (85–110°C) (Figure 3A, B). Bacterial hyperthermophiles and archaeal psychrophiles were excluded from analyses due to their poor representation in the SR-GTDB. Given the uneven distribution of specific prokaryotic immune systems over bacteria and archaea (Figure 2A), and the observation that bacteria and archaea are distributed unevenly over distinct T_{opt} classes (Figure 3A, B), the correlation between immune system abundance and host T_{opt} was analysed independently for bacteria and archaea.

It has previously been determined that CRISPR-Cas system abundance has a strong positive correlation with T_{opt} (28–32). To determine if this is the same for other prokaryotic immune systems, we determined the total number of immune systems encoded in bacterial genomes for each of the different T_{opt} categories. However, only relatively small differences between the average number of immune systems were observed for different T_{opt} categories in bacteria (Figure 3C; Pearson correlation coefficient: –0.11, $P < 10^{-14}$). As bacteria with a higher T_{opt} generally have smaller genomes (Figure 3D; Pearson correlation coefficient: –0.31, $P < 10^{-99}$), we corrected for genome size. Although a positive correlation between T_{opt} and the total number of immune systems encoded could be observed,



		# genomes	Abi2	AbiE	AbiH	AbiL	AbiU	Argonaute	AVAST	Borvo	BREX	Cas	CBASS	DarTG	DISARM	DRT	Druantia	dXTPase	Gabija	Hachiman	hma	Ier	Kiwa	Lamassu	Mokosh	Nhi	PARIS	PsyrTA	PT	Qat	Retron	RM	RosmerTA	Septu	Shango	Shedu	SoFIC	Tiamat	Viperin	Wadjet	Zorya
Bacteria	Actinobacteria	813	17	21	2	9	3	9	2	11	11	31	14	2	7	5	4	86	12	3	4	8	1	12	14	0	1	3	4	2	1	90	7	7	3	7	11	4	2	58	0
	Bacteroidota	470	13	39	13	15	3	15	9	6	5	33	24	4	0	18	3	4	18	6	1	9	4	13	9	6	12	5	8	3	14	86	1	24	2	6	20	9	3	7	3
	Cyanobacteria	163	3	20	1	24	12	34	4	57	11	67	18	13	13	17	10	2	3	1	11	14	1	6	26	1	20	11	44	1	13	83	0	30	29	8	13	6	2	1	18
	Firmicutes	666	20	18	9	7	3	5	3	2	12	55	10	3	1	10	2	8	16	5	2	5	5	15	8	5	3	3	10	2	7	96	8	8	10	4	7	3	4	28	7
	Proteobacteria	2273	9	28	2	4	4	5	4	3	8	29	18	5	2	13	4	49	18	7	2	10	3	16	9	3	5	4	6	5	13	83	5	15	5	7	18	1	3	14	22
	Spirochaetes	50	22	18	2	18	6	2	4	10	4	48	10	0	0	6	2	2	10	4	0	8	2	2	0	4	2	0	4	2	6	82	4	4	6	6	18	2	4	14	2
	Tenericutes	113	8	12	4	0	0	2	0	1	1	24	5	0	0	3	0	2	4	2	0	2	0	3	2	0	1	0	1	0	0	78	1	3	2	0	2	0	0	8	1
	Other bacteria	358	6	26	2	7	1	9	3	7	14	57	13	5	4	7	1	15	8	3	9	10	2	7	6	9	5	8	6	2	8	85	2	9	3	4	20	1	3	8	5
	Total bacteria	4906	12	26	4	7	3	8	4	7	9	37	16	4	3	11	4	40	15	5	3	9	3	13	10	3	5	4	8	3	10	86	5	13	6	6	15	3	3	22	12
Archaea	Crenarchaeota	53	0	2	0	0	0	11	0	0	0	92	0	0	2	0	0	0	0	0	9	0	0	2	8	0	0	0	0	0	0	49	0	0	0	0	0	0	0	0	0
	Euryarchaeota	184	0	19	0	3	1	18	4	2	18	72	9	0	5	2	1	0	6	4	8	1	9	2	4	1	5	11	11	0	0	93	0	1	4	2	25	9	10	1	0
	Other archaea	54	0	35	2	4	2	11	2	2	2	44	6	0	2	2	0	0	2	2	6	4	2	6	6	0	11	4	6	0	0	74	0	0	6	2	6	0	9	0	0
	Total archaea	291	0	19	0	3	1	16	2	1	12	71	7	0	4	2	1	0	4	3	9	1	6	3	5	1	5	8	8	0	0	82	0	0	3	2	17	5	8	1	0
	Total number found	5197	578	1310	206	354	173	441	204	325	489	2003	810	209	151	547	173	1943	763	276	165	440	158	666	489	163	270	215	392	163	474	4479	246	659	290	301	795	157	154	1069	602

Figure 2. Distribution of immune systems in different phyla and in correlation to genome size. (**A**) Percentage of SR-GTDB genomes that encodes a specific immune system family, subdivided for the prokaryotic phyla. Only immune system families identified in ≥ 150 genomes are shown; genomes from phyla with < 30 accessions in the SR-GTDB are grouped as 'other bacteria' or 'other archaea'. (**B**) Distribution of the total number of immune systems encoded in prokaryotes in the SR-GTDB for different phyla, represented by raincloud plots (visualizing individual data points as well as the probability density of the data) and a boxplot. The median is indicated in the boxplot, while the average (Av) is shown on the right. (**C**) Distribution of genome sizes for prokaryotes in the SR-GTDB for different phyla, represented by a raincloud plot and a boxplot, with averages indicated. (**D**) Correlation between (i) number of immune systems encoded and (ii) genome size for both bacteria (blue) and archaea (orange) represented in the SR-GTDB. Pearson correlation coefficients (*r*) and *P*-values are indicated. (**E**) Percentage of genomes encoding specific immune system families for different genome size categories. Only immune system families identified in ≥ 150 genomes are shown.

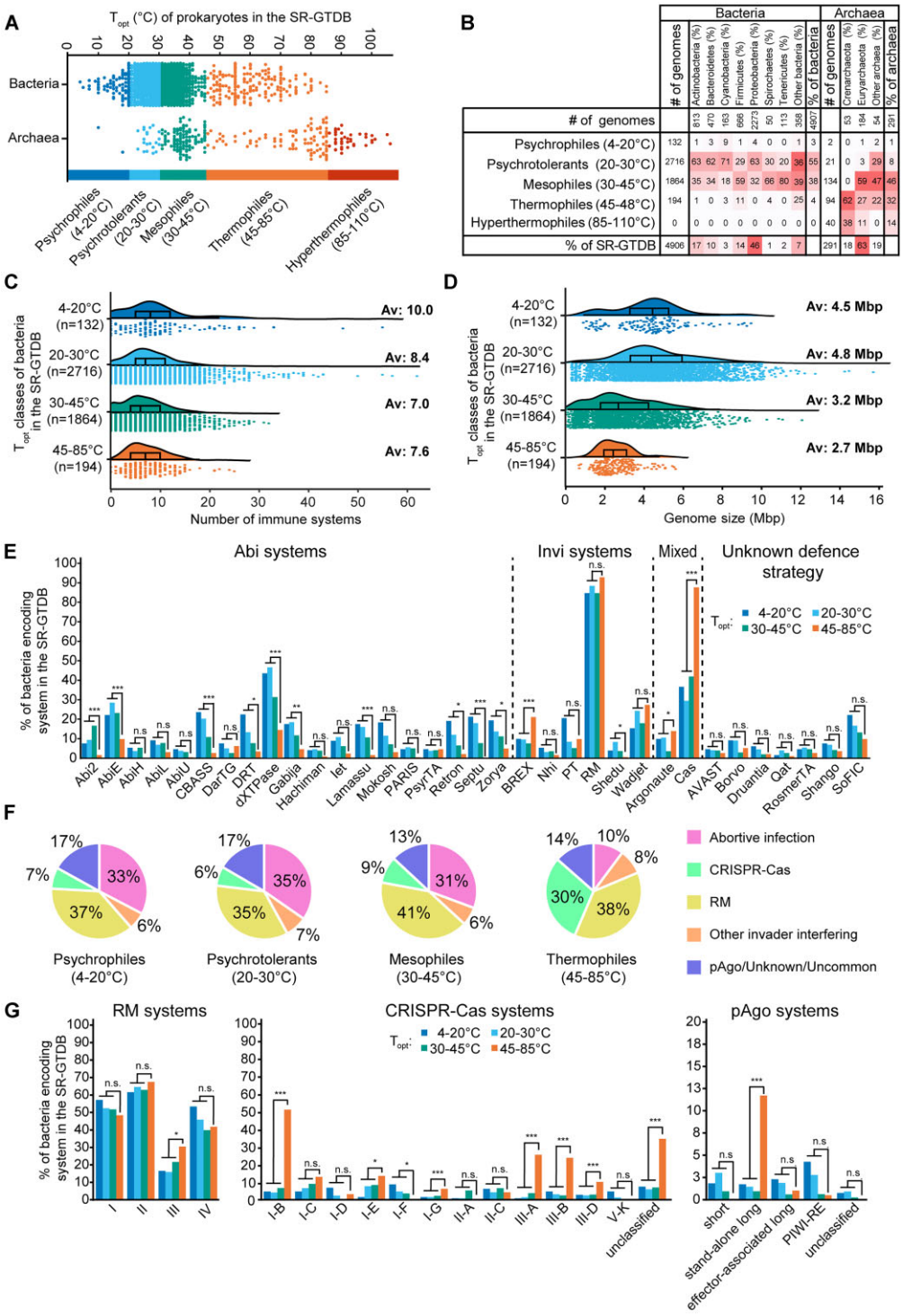


Figure 3. Distribution of immune systems in different T_{opt} categories. (**A**) Definition of T_{opt} categories and their bacterial and archaeal coverage in the SR-GTDB. (**B**) Distribution of genomes in the SR-GTDB for different T_{opt} categories and phyla, with averages indicated. Genomes from phyla with < 30 accessions are grouped as 'other bacteria' and 'other archaea'. (**C**) Distribution of the total number of immune systems in bacteria in the SR-GTDB per T_{opt} category, represented by a raincloud plot (visualizing both individual data points and the probability density of the data) and a boxplot. The median is indicated in the boxplot, while the average (Av) is shown on the right. (**D**) Distribution of bacterial genome sizes for different T_{opt} categories, represented by raincloud plots (visualizing both individual data points and the probability density of the data) and a boxplot. The median is indicated in the boxplot, while the average (Av) is shown on the right. (**E**) Percentage of bacterial genomes encoding specific immune system families (when found in ≥ 150 genomes) for different T_{opt} categories. *, ** and *** indicate *P*-values < 1.43×10^{-3}, 1.43×10^{-4} and 1.43×10^{-5}. *P*-values are determined by a χ^2 test of independence; cut-off values are determined by a Bonferroni test of $\alpha = 0.05$ / number of comparisons. (**F**) Proportional distribution of different immune system categories for different T_{opt} categories in bacteria. (**G**) Percentage of bacterial genomes with specific RM, CRISPR-Cas and pAgo subtypes identified (when found in ≥ 150 genomes) for different T_{opt} categories. For RM, *, ** and *** indicate *P*-values 1.43×10^{-3}, 1.43×10^{-4} and 1.43×10^{-5}. For CRISPR-Cas, *, ** and *** indicate *P*-values < 3.85×10^{-3}, 3.85×10^{-4} and 3.85×10^{-5}. For pAgos, *, ** and *** indicate *P*-values < 1×10^{-2}, 1×10^{-3} and 1×10^{-4}. *P*-values are determined by a χ^2 test of independence; cut-off values are determined by a Bonferroni test of $\alpha = 0.05$ / number of comparisons.

the strength of this correlation is weak (Pearson correlation coefficient: 0.17, $P < 10^{-31}$), unlike what has been reported for CRISPR-Cas systems individually.

We hypothesized that abundance/T_{opt} correlations vary for distinct prokaryotic immune systems, and therefore investigated this correlation for all immune systems individually (Figure 3E; Supplementary Figure S6A). Corroborating earlier studies (28–32), CRISPR-Cas systems are more abundant in bacterial hosts with a $T_{opt} \geq 45°C$. Akin to CRISPR-Cas systems, pAgos systems are also more abundant in bacterial hosts with a $T_{opt} \geq 45°C$. Both CRISPR-Cas and pAgo immune system families are highly diverse and can rely on Abi and/or Invi as the functional mechanism (45–51). Remarkably, for the Abi systems Abi2, AbiE, CBASS, DRT, dXTPases, Gabija, Lamassu, Retron, Septu and Zorya, a negative correlation between T_{opt} and abundance is observed, and a similar trend in abundance is observed for AbiH, AbiL, AbiU, Hachiman, Iet, Mokosh and PARIS, although this trend is non-significant (8–13,20,52–63) (Figure 3E). PsyrTA and DarTG (57,64) are the only Abi systems for which the trend is not observed. In contrast, for the Invi system BREX (65), a positive correlation between T_{opt} and abundance is observed, while for the Invi system Shedu (66,67) a negative correlation is observed. None of the other Invi systems shows a statistically significant correlation between T_{opt} and abundance.

As most prokaryotes generally encode multiple immune systems (Figure 2B), we investigated the composition of the immune system arsenal for the different T_{opt} classes (Figure 3F; Supplementary Figure S7A). To this end, the identified immune systems were classified as (i) Abi system, (ii) RM systems, (iii) CRISPR-Cas systems, (iv) other Invi systems or (v) pAgo/Unknown/Uncommon systems. While no general trends are observed for Invi systems, RM systems and pAgo/Unknown/Uncommon systems, the fraction and total number of CRISPR-Cas systems are higher in bacteria that belong to higher T_{opt} classes, while the fraction and total number of Abi systems are lower (Figure 3F; Supplementary Figure S7A).

In general, most trends observed in bacteria can also be observed in archaea (Supplementary Figure S5). However, due to the generally lower number of genomes and consequentially lower number of immune systems identified, not all correlations observed in bacteria are statistically significant in archaea. Of note, the T_{opt} class of psychrophiles is excluded for archaea (due to poor representation in the SR-GTDB) but, in contrast to the analyses performed for bacteria, the T_{opt} class of hyperthermophiles is included. Just like in bacteria, in archaea host T_{opt} and CRISPR-Cas abundance are positively correlated (Supplementary Figure S5D). In contrast to observations made for bacteria, a negative correlation is observed for SoFIC and for BREX in archaea (Supplementary Figure S5D). Investigation of the general composition of the immune system arsenal for the different T_{opt} classes in archaea corroborates observations made in bacteria: CRISPR-Cas systems are more abundant in high T_{opt} hosts, while Abi systems are more abundant in low T_{opt} hosts (Supplementary Figures S5E and S7B). In conclusion, while CRISPR-Cas system abundance is positively correlated with host T_{opt}, the abundance of almost all Abi systems is negatively correlated with host T_{opt}. This suggests that CRISPR-Cas systems provide a stronger selective advantage in organisms that thrive at relatively high temperatures, while Abi systems provide a stronger selective advantage in organisms that thrive at relatively low temperatures.

Distribution of RM, CRISPR-Cas and pAgo system (sub)types

RM, CRISPR-Cas and pAgo systems are highly diversified immune system families for which many different (sub)types exist (38,68,69). Depending on the (sub)type, these systems can mediate Abi, Invi or rely on mixed Abi/Invi strategies. RM systems are subdivided in four types (Figure 3G; Supplementary Figure S5F), all of which mediate Invi (68). CRISPR-Cas systems are classified in six types and > 40 subtypes (69). While most CRISPR-Cas systems mediate Invi by targeting and degrading invader nucleic acids, Type VI (45), subtype V-A2 (46) and potentially subtype V-G CRISPR-Cas systems (47,48) mediate Abi. Furthermore, subtype I-F and type III systems mediate both Invi and Abi (49–51). Based on phylogeny, pAgo systems are classified as long-A pAgo, long-B pAgo, short pAgo, SiAgo-like and PIWI-RE systems (37,38). While long-A pAgos generally act as Invi systems (38,70–72), characterized long-B pAgo, short pAgo and siAgo-like systems act as Abi systems (14,15,73).

As the distribution of various immune systems appears to be influenced by their immune system strategy, and because distinct immune system (sub)types mediate different immune strategies, we investigated the distribution of RM, CRISPR-Cas and pAgo (sub)types both in bacteria and in archaea (Figure 3F; Supplementary Figures S5F and S6B). In bacteria, the abundance of type III RM systems is positively correlated with host T_{opt} (Figure 3F), while in archaea the abundance of type I systems negatively correlates with host T_{opt} (Supplementary Figure S5F). CRISPR-Cas subtypes I-A (in archaea only), I-B (in bacteria only), III-A, III-B and III-D (in bacteria only), and unclassified CRISPR-Cas systems, show a > 5-fold increased abundance in prokaryotic hosts with a $T_{opt} \geq 45°C$ compared with prokaryotic hosts with a $T_{opt} < 45°C$ (Figure 3F; Supplementary Figures S5F and S6B). Furthermore, the abundances of CRISPR-Cas subtypes I-E and I-G are also positively correlated with the host T_{opt} in bacteria. In contrast, the abundance of mixed-strategy subtype I-F systems is negatively correlated with T_{opt} in bacteria. Standalone long pAgos, which generally act as Invi systems (38,70–72), are more abundant in bacteria with a high T_{opt} (Figure 3G). In contrast, Abi-conferring short pAgo systems and long pAgos that associate with effector enzymes appear more abundant in prokaryotes with a low T_{opt}, although this correlation is significant only for short pAgos in archaea (Figure 3G; Supplementary Figure S5F). Combined, this shows that certain Abi-conferring subtypes of CRISPR-Cas and pAgo systems are less abundant in hosts with a high T_{opt}, while Invi-conferring subtypes of CRISPR-Cas and pAgo systems are more abundant in hosts with a high T_{opt}.

Distribution of prokaryotic immune systems in distinct environments

While the SR-GTDB provides a broad taxonomic sampling of prokaryotes, the sampling might be skewed towards cultivatable microbes. In addition, T_{opt} does not necessarily correlate directly with environmental temperature, and other environmental parameters could also influence the distribution of immune systems. To investigate if environmental temperature and/or other environmental parameters affect prokaryotic immune system distribution, our custom script that combines the output of PADLOC and DefenseFinder was used to identify prokaryotic immune systems in metagenomes isolated from various environments (Figure 4). It should be noted

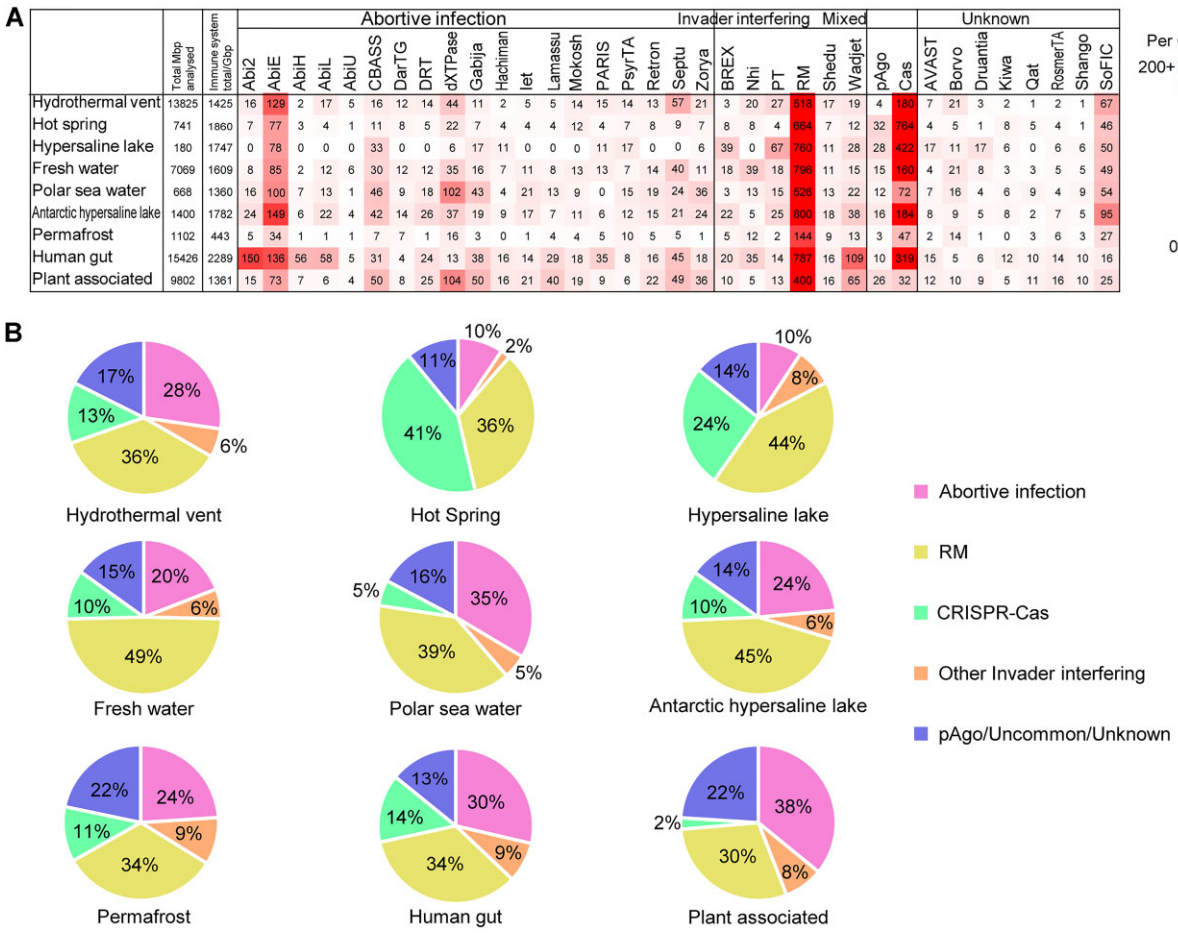


Figure 4. Distribution of prokaryotic immune systems in metagenomes isolated from different environments. (**A**) Number of immune systems identified per gigabase pair (Gbp) of each metagenomic dataset. (**B**) Proportional distribution of different immune system categories for different metagenomic datasets.

that, in contrast to analyses on SR-GTDB genomes (Figures 1–3), no phylogenetic curation was performed on these metagenomic datasets.

First, we investigated the general abundance of immune systems in distinct environments (Figure 4A). We observe considerable differences in the number of immune systems encoded per Gbp of metagenomics data, ranging from 461 systems per Gbp in permafrost metagenomes to 2 337 per Gbp in human gut metagenomes. Further investigation of the abundance of specific immune systems reveals large differences in abundance for certain immune systems in specific environments (Figure 4A). For example, in human gut metagenomes, Abi2 is far more abundant than in other environments and, in hypersaline lake metagenomes, most Abi systems are absent, despite these metagenomes encoding a generally high number of immune systems (1 791 systems per Gbp).

Next, we analysed the proportional distribution of (i) Abi, (ii) RM, (iii) CRISPR-Cas, (iv) other Invi systems and (v) pAgo/Uncommon/Unknown systems in these metagenomes. As observed for the different T_{opt} classes (Figure 3), the proportional distribution of Abi systems and CRISPR-Cas systems differs considerably in metagenomes from distinct environments, while differences in the proportional abundance of RM and other Invi systems are smaller (Figure 4B). Corroborating the results for host T_{opt}/abundance correlations described above, the metagenomes from a cold environment (polar sea water) have a relative high proportional abundance of Abi systems, while the metagenomes from a hot environments (hot spring) have a high abundance of CRISPR-Cas systems (Figure 4B). A more even distribution of CRISPR-Cas and Abi systems is observed in metagenomes isolated from hydrothermal vents (where temperatures can range from 4°C to > 100°C), fresh water, permafrost and Antarctic hypersaline lakes. In a high salt environment (hypersaline lake), we observe high CRISPR-Cas and low Abi abundance, similar to the proportional distribution of thermophilic organisms (Figure 3F). In host-associated microbiomes (human gut and plant), where there is a strong selection for specific taxa (74), we find relatively many Abi systems, and notably varying amounts of CRISPR-Cas systems. These results corroborate the correlation between temperature and abundance of specific immune systems, and make it evident that additional environmental parameters play a role in the distribution of prokaryotic immune systems.

Discussion

Prokaryotic immune systems are often subjected to horizontal gene transfer (28–31), resulting in most immune system families being present across bacterial and archaeal phyla and environments (16). Yet, the distribution of prokaryotic immune systems in nature is uneven (16). We hypothesize that under

distinct conditions, certain immune systems provide a greater selective advantage than other systems. Consequentially, this results in such immune systems becoming more abundant in the corresponding environment, causing the uneven distribution observed. In this study, we provide a systematic analysis of the abundance of different immune system families thereby showing that phylogeny, genome size and host T_{opt}, as well as other environmental parameters, are factors that correlate with the abundance of specific prokaryotic immune systems.

The evolutionary arms race between prokaryotes and their invaders has resulted in highly diversified immune systems (3), making their identification challenging. PADLOC and DefenseFinder rely on tool-specific HMM profiles, classification algorithms and cut-off scores to identify prokaryotic immune systems. Consequently, both tools identify immune systems with different success rates, and the immune systems identified by PADLOC and DefenseFinder only partially overlap: each tool additionally identifies a significant number of immune systems not identified by the other tool (Figure 1). When a user requires a broad identification of putative immune systems, the output of DefenseFinder and PADLOC can be merged, as done in this study. Based on identification by DefenseFinder alone, it was previously estimated that prokaryotes encode an average of 5.2 immune systems per genome (16). By combining the output of DefenseFinder and PADLOC, we identify an average of 7.9 and 5.9 immune systems in bacterial and archaeal genomes, respectively. It should be noted that novel immune systems are regularly identified and characterized. This results in various immune systems not yet being included in the search strategies of the tools. As such, it is likely that the used tools, even when combined, do not provide a complete image of the entire immune system arsenal encoded in prokaryotic genomes. However, DefenseFinder and PADLOC are regularly expanded and updated; while this manuscript was in preparation, new immune system databases have already been released. Therefore, future versions of the tools are likely to identify a larger fraction of the complete immune system arsenal. Furthermore, certain users might require predictions with a low chance of false positives. In that case, the user should select only immune systems identified by both tools. We have made an automated script that combines DefenseFinder and PADLOC for the identification of prokaryotic immune systems in either the broad or the selective mode https://github.com/LOlijslager/find_prokaryotic_immune_systems and dx.doi.org/10.6084/m9.figshare.24632142.

Although prokaryotic immune systems are regularly horizontally transferred between bacteria and archaea (28–31), in agreement with previous studies (16), our analysis shows that certain systems are completely absent in archaea or are encoded in specific bacterial phyla only. While we cannot rule out that more distant homologous systems exist and are not identified by PADLOC or DefenseFinder, we speculate that this patchy distribution of immune systems might, for example, be a result of these systems emerging later in evolution, the reliance on clade-specific accessory proteins or that these systems provide a larger selective advantage in specific hosts for other reasons. Our data corroborate earlier findings that genome size and the total number of immune systems encoded are positively correlated (16). While we additionally observe differences in the total number of immune systems encoded by prokaryotes from different phyla, this difference is largely explained by the prokaryotes from these phyla having different genome sizes. This implies that genome size serves as an important physical limitation for the total number of immune systems encoded. Yet, it is likely that other factors that contribute to the number of immune systems encoded exist. Of note, we show that the positive correlation between genome size and abundance does not hold for all immune system families: AVAST and CRISPR-Cas systems do not have a linear correlation between their abundance and genome size, and Abi2 and Nhi systems appear to decrease in abundance in larger genomes. Possible explanations include (but are not limited to) that larger genomes are at a higher risk for autoimmunity by these systems, or that these systems become redundant and/or interfere with increasing numbers of other systems.

Corroborating earlier studies (28–31), we show that CRISPR-Cas systems are more abundant in hosts with a higher T_{opt} (Figure 3): a 2-fold increase of CRISPR-Cas systems is observed in the genomes of hosts with a $T_{opt} > 45°C$. To determine whether this is a general trend for immune systems, we expanded this analysis to 35 other commonly encoded immune system families. A similar sharp increase in abundance in the genomes of bacterial hosts with a $T_{opt} > 45°C$ is observed for stand-alone long pAgos (6-fold increase). Long-A pAgos are programmed with small nucleic acid guides to recognize and neutralize invaders in a sequence-specific manner (38,70–72) akin to most CRISPR-Cas systems.

As many CRISPR-Cas systems that mediate Invi also show this correlation, it is tempting to speculate that a positive correlation between T_{opt} and abundance can generally be observed for Invi systems. However, while a positive correlation between host T_{opt} and abundance is also observed for certain RM system subtypes and BREX, no positive correlation is observed for other Invi systems. Possibly, the immune mechanism and downstream responses of these immune systems are not yet fully understood; we speculate that these exceptions might be explained by the hypothesis that certain Invi systems, especially RM systems, instigate downstream Abi responses (75). Additionally, an Abi response has also been observed for Invi systems Shedu and Wadjet in specific cases, indicating alternative responses or downstream responses for these systems too (60,76). For Abi systems, almost without exception, a negative correlation between T_{opt} and system abundance is observed, and none of the Abi systems investigated has a positive correlation between T_{opt} and system abundance. Corroborating these findings, analyses of the proportional distribution of different immune system strategies per genome and in environmental samples revealed that CRISPR-Cas systems and Abi systems increase and decrease, respectively, with increasing T_{opt}. Certain systems for which the underlying immune mechanism is unknown also show either a positive [e.g. hma in both bacteria (significant) and archaea (non-significant), Supplementary Figures S5D and S6A] or negative correlation (Azaca, in bacteria only, Supplementary Figure S6A) between abundance and host T_{opt}. While it is tempting to speculate that these are Invi and Abi systems, respectively, we stress that host T_{opt} is not the only determinant affecting immune system distribution.

These findings at least partially contradict various hypotheses previously put forward to explain the increased abundance of CRISPR-Cas systems in high T_{opt} hosts. For example, it has previously been hypothesized that the lack of cell-grazing predator organisms at temperatures above 45°C allows thermophilic prokaryotes to invest more of their energy in their immune system arsenal (32). However, this theory does not explain why most immune systems have a negative correlation with host T_{opt}. It has also been hypothesized that

at high temperatures viruses have a lower chance of escaping sequence-specific CRISPR-Cas immunity (30,77): in thermophilic environments, mutations are more frequently detrimental, resulting in lower mutation rates and lower MGE diversity (78,79). Indeed, an MGE mutation rate below a theoretical threshold value is necessary for CRISPR-Cas systems to mediate a fitness gain (30). The same hypothesis could also explain why stand-alone long pAgos, which are also sequence-specific Invi systems, follow a similar distribution. However, it does not explain why the abundance of other programmable sequence-specific immune systems, including short and effector-associated long pAgo systems, as well as CRISPR-Cas subtypes I-D, I-F, II-A, II-C and V-K, have no or a negative correlation with host T_{opt}.

We propose various hypotheses, none of which is mutually exclusive, to describe why Abi systems might be more abundant in hosts with a low T_{opt}. We hypothesize that Abi systems provide the highest fitness gain when viruses are likely to encounter kin of the original host. When the chance that this occurs is lower, for example due to low cell density and/or low virus stability, sacrificing the host cell is a relatively costly strategy. Possibly, prokaryotes with $T_{opt} < 45°C$ grow in clumps or biofilms, or form filaments relatively often, which makes Abi more beneficial. In addition, the extreme environments in which thermophiles thrive could reduce the stability of viruses, which leaves a shorter time window to infect new hosts. However, thermophiles frequently grow in biofilms (80) and viruses from thermophilic environments have adapted to withstand extreme conditions (81). Another explanation could be that because thermophilic hosts live in extreme environments and therefore close to what is physically still feasible, the fitness cost of Abi might be too high. Finally, we speculate that invader replication and spread occur faster at higher temperatures. Rapid virus propagation might make it difficult for Abi systems to shut down the host cell before lytic viruses can escape. Future experimental studies are required to establish if any of these hypotheses are valid.

Compared with eukaryotes, prokaryotes live in a much wider range of environments, and have highly diversified cellular mechanisms. Many of the factors that influence the complex ecology of prokaryotes and their invaders are unknown. Here we reveal that the abundance of specific Invi systems is positively correlated with host T_{opt}, while the abundance of Abi systems is negatively correlated with host T_{opt}. This suggests that the general strategy of prokaryotic immune systems (Invi or Abi) provides distinct fitness gains in different environments. Further studies are required to expand our understanding of what other environmental and/or additional factors affect the distribution of prokaryotic immune systems.

Data availability

Unprocessed data used to prepare the figures in this manuscript are available in Supplementary data S1, S2 and S4. Raw data are available on dx.doi.org/10.6084/m9.figshare.24632142. Scripts are available at: https://github.com/LOlijslager/find_prokaryotic_immune_systems and dx.doi.org/10.6084/m9.figshare.24632142.

Supplementary data

Supplementary Data are available at NARGAB Online.

Acknowledgements

We are grateful to Sumanth K. Mutte for valuable discussion and critical reading of the manuscript. We thank members of the Weijers and Swarts labs and other colleagues at the Laboratory of Biochemistry at Wageningen University for discussion.

Author contributions: L.H.O., D.W. and D.C.S. conceived the project. L.H.O. wrote scripts, generated and analysed all data, performed statistical analyses and made figures under supervision of D.W. and D.C.S.. L.H.O. and D.C.S. wrote the manuscript with input from D.W.

Funding

The Graduate School of Experimental Plant Sciences [EPS; EPS-1 036 to L.H.O];). the Netherlands Organization for Scientific Research (NWO) VENI grant [016.Veni.192.072 to D.C.S.]; and the European Research Council (ERC) [ERC-2020-STG 948783 to D.C.S.].

Conflict of interest statement

None declared.

References

1. Koonin,E.V., Makarova,K.S., Wolf,Y.I. and Krupovic,M. (2020) Evolutionary entanglement of mobile genetic elements and host defence systems: guns for hire. *Nat. Rev. Genet.*, **21**, 119–131.
2. Stern,A. and Sorek,R. (2011) The phage–host arms race: shaping the evolution of microbes. *Bioessays*, **33**, 43–51.
3. Safari,F., Sharifi,M., Farajnia,S., Akbari,B., Karimi Baba Ahmadi,M., Negahdaripour,M. and Ghasemi,Y. (2020) The interaction of phages and bacteria: the co-evolutionary arms race. *Crit. Rev. Biotechnol.*, **40**, 119–137.
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. Oliveira,P.H., Touchon,M. and Rocha,E.P. (2014) The interplay of restriction–modification systems with mobile genetic elements and their prokaryotic hosts. *Nucleic Acids Res.*, **42**, 10618–10631.
6. Makarova,K.S., Wolf,Y.I., Iranzo,J., Shmakov,S.A., Alkhnbashi,O.S., Brouns,S.J., Charpentier,E., Cheng,D., Haft,D.H. and Horvath,P. (2020) Evolutionary classification of CRISPR–Cas systems: a burst of class 2 and derived variants. *Nat. Rev. Microbiol.*, **18**, 67–83.
7. Lopatina,A., Tal,N. and Sorek,R. (2020) Abortive infection: bacterial suicide as an antiviral immune strategy. *Annu. Rev. Virol.*, **7**, 371–384.
8. McLandsborough,L., Kolaetis,K., Requena,T. and McKay,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.
9. Garvey,P., Fitzgerald,G. 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.
10. Prévots,F., Daloyau,M., Bonin,O., Dumont,X. and Tolou,S. (1996) Cloning and sequencing of the novel abortive infection gene abiH of *Lactococcus lactis* ssp. *lactis* biovar. *diacetylactis* S94. *FEMS Microbiol. Lett.*, **142**, 295–299.
11. Deng,Y.-M., Liu,C.-Q. and Dunn,N.W. (1999) Genetic organization and functional analysis of a novel phage abortive infection system, AbiL, from *Lactococcus lactis*. *J. Biotechnol.*, **67**, 135–149.

12. 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.
13. 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.
14. Zeng,Z., Chen,Y., Pinilla-Redondo,R., Shah,S.A., Zhao,F., Wang,C., Hu,Z., Wu,C., Zhang,C. and Whitaker,R.J. (2022) A short prokaryotic Argonaute activates membrane effector to confer antiviral defense. *Cell Host Microbe*, **30**, 930–943.
15. Koopal,B., Mutte,S.K. and Swarts,D.C. (2022) A long look at short prokaryotic Argonautes. *Trends Cell Biol.*, **33**, 605–618.
16. Tesson,F., Hervé,A., Mordret,E., Touchon,M., d'Humières,C., Cury,J. and Bernheim,A. (2022) Systematic and quantitative view of the antiviral arsenal of prokaryotes. *Nat. Commun.*, **13**, 2561.
17. Kropocheva,E., Lisitskaya,L., Agapov,A., Musabirov,A., Kulbachinskiy,A. and Esyunina,D. (2022) Prokaryotic Argonaute proteins as a tool for biotechnology. *Mol. Biol.*, **56**, 854–873.
18. 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.
19. 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.
20. 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.
21. Gao,L., Altae-Tran,H., Bohning,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.
22. Doug,H., 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 Bioinform.*, **11**, 119.
23. 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.
24. 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.
25. 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.
26. Meaden,S., Biswas,A., Arkhipova,K., Morales,S.E., Dutilh,B.E., Westra,E.R. and Fineran,P.C. (2022) High viral abundance and low diversity are associated with increased CRISPR-Cas prevalence across microbial ecosystems. *Curr. Biol.*, **32**, 220–227.
27. Beavogui,A., Lacroix,A., Wiart,N., Poulain,J., Delmont,T.O., Paoli,L., Wincker,P. and Oliveira,P.H. (2024) The defensome of complex bacterial communities. *Nat. Commun.*, **15**, 2146.
28. Jansen,R., van Embden,J.D., Gaastra,W. and Schouls,L.M. (2002) Identification of a novel family of sequence repeats among prokaryotes. *OMICS*, **6**, 23–33.
29. Anderson,R.E., Brazelton,W.J. and Baross,J.A. (2011) Using CRISPRs as a metagenomic tool to identify microbial hosts of a diffuse flow hydrothermal vent viral assemblage. *FEMS Microbiol. Ecol.*, **77**, 120–133.
30. Weinberger,A.D., Wolf,Y.I., Lobkovsky,A.E., Gilmore,M.S. and Koonin,E.V. (2012) Viral diversity threshold for adaptive immunity in prokaryotes. *mBio*, **3**, e00456-12.
31. Weissman,J.L., Laljani,R.M., Fagan,W.F. and Johnson,P.L. (2019) Visualization and prediction of CRISPR incidence in microbial trait-space to identify drivers of antiviral immune strategy. *ISME J.*, **13**, 2589–2602.
32. Lan,X.-R., Liu,Z.-L. and Niu,D.-K. (2022) Precipitous increase of bacterial CRISPR-Cas abundance at around 45°C. *Front. Microbiol.*, **13**, 773114.
33. Parks,D.H., Chuvochina,M., Rinke,C., Mussig,A.J., Chaumeil,P.-A. and Hugenholtz,P. (2022) GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Res.*, **50**, D785–D794.
34. White,J.R. (2011) Prodigal. University of New Orleans Theses and Dissertations. p. 1379.
35. Zhang,Z. and Wood,W.I. (2003) A profile hidden Markov model for signal peptides generated by HMMER. *Bioinformatics*, **19**, 307–308.
36. Abby,S.S., Néron,B., Ménager,H., Touchon,M. and Rocha,E.P. (2014) MacSyFinder: a program to mine genomes for molecular systems with an application to CRISPR-Cas systems. *PLoS One*, **9**, e110726.
37. 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.
38. Ugarte,P.B., Barendse,P. and Swarts,D.C. (2023) Argonaute proteins confer immunity in all domains of life. *Curr. Opin. Microbiol.*, **74**, 102313.
39. Levy,A., Salas Gonzalez,I., Mittelviefhaus,M., Clingenpeel,S., Herrera Paredes,S., Miao,J., Wang,K., Devescovi,G., Stillman,K. and Monteiro,F. (2018) Genomic features of bacterial adaptation to plants. *Nat. Genet.*, **50**, 138–150.
40. Cao,S., Zhang,W., Ding,W., Wang,M., Fan,S., Yang,B., Mcminn,A., Wang,M., Xie,B.-b. and Qin,Q.-L. (2020) Structure and function of the Arctic and Antarctic marine microbiota as revealed by metagenomics. *Microbiome*, **8**, 47.
41. Gomariz,M., Martinez-Garcia,M., Santos,F., Rodriguez,F., Capella-Gutiérrez,S., Gabaldon,T., Rossello-Mora,R., Meseguer,I. and Anton,J. (2015) From community approaches to single-cell genomics: the discovery of ubiquitous hyperhalophilic Bacteroidetes generalists. *ISME J.*, **9**, 16–31.
42. Schoch,C.L., Ciufo,S., Domrachev,M., Hotton,C.L., Kannan,S., Khovanskaya,R., Leipe,D., Mcveigh,R., O'Neill,K. and Robbertse,B. (2020) NCBI Taxonomy: a comprehensive update on curation, resources and tools. *Database*, **2020**, baaa062.
43. Cock,P.J., Antao,T., Chang,J.T., Chapman,B.A., Cox,C.J., Dalke,A., Friedberg,I., Hamelryck,T., Kauff,F. and Wilczynski,B. (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, **25**, 1422–1423.
44. Li,G., Rabe,K.S., Nielsen,J. and Engqvist,M.K. (2019) Machine learning applied to predicting microorganism growth temperatures and enzyme catalytic optima. *ACS Synth. Biol.*, **8**, 1411–1420.
45. Meeske,A.J., Nakandakari-Higa,S. and Marraffini,L.A. (2019) Cas13-induced cellular dormancy prevents the rise of CRISPR-resistant bacteriophage. *Nature*, **570**, 241–245.
46. Dmytrenko,O., Neumann,G.C., Hallmark,T., Keiser,D.J., Crowley,V.M., Vialetto,E., Mougiakos,I., Wandera,K.G., Domgaard,H. and Weber,J. (2023) Cas12a2 elicits abortive infection through RNA-triggered destruction of dsDNA. *Nature*, **613**, 588–594.
47. Chen,Y., Zeng,Z., She,Q. and Han,W. (2022) The abortive infection functions of CRISPR-Cas and Argonaute. *Trends Microbiol.*, **31**, 405–418.
48. Yan,W.X., Hunnewell,P., Alfonse,L.E., Carte,J.M., Keston-Smith,E., Sothiselvam,S., Garrity,A.J., Chong,S., Makarova,K.S. and Koonin,E.V. (2019) Functionally diverse type V CRISPR-Cas systems. *Science*, **363**, 88–91.
49. Watson,B.N., Vercoe,R.B., Salmond,G.P., Westra,E.R., Staals,R.H. and Fineran,P.C. (2019) Type IF CRISPR-Cas resistance against

virulent phages results in abortive infection and provides population-level immunity. *Nat. Commun.*, **10**, 5526.
50. 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.
51. 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.
52. González-Delgado,A., Mestre,M.R., Martínez-Abarca,F. and Toro,N. (2021) Prokaryotic reverse transcriptases: from retroelements to specialized defense systems. *FEMS Microbiol. Rev.*, **45**, fuab025.
53. Tal,N., Millman,A., Stokar-Avihail,A., Fedorenko,T., Leavitt,A., Melamed,S., Yirmiya,E., Avraham,C., Brandis,A. and Mehlman,T. (2022) Bacteria deplete deoxynucleotides to defend against bacteriophage infection. *Nat. Microbiol.*, **7**, 1200–1209.
54. Cheng,R., Huang,F., Lu,X., Yan,Y., Yu,B., Wang,X. and Zhu,B. (2022) The prokaryotic Gabija complex senses both viral transcription and DNA metabolism for antiviral defense. https://doi.org/10.21203/rs.3.rs-1703025/v1.
55. Millman,A., Melamed,S., Leavitt,A., Doron,S., Bernheim,A., Hör,J., Garb,J., Bechon,N., Brandis,A. and Lopatina,A. (2022) An expanded arsenal of immune systems that protect bacteria from phages. *Cell Host Microbe*, **30**, 1556–1569.
56. Rousset,F., Depardieu,F., Miele,S., Dowding,J., Laval,A.-L., Lieberman,E., Garry,D., Rocha,E.P., Bernheim,A. and Bikard,D. (2022) Phages and their satellites encode hotspots of antiviral systems. *Cell Host Microbe*, **30**, 740–753.
57. 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.
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. Wang,S., Sun,E., Liu,Y., Yin,B., Zhang,X., Li,M., Huang,Q., Tan,C., Qian,P. and Rao,V.B. (2023) Landscape of new nuclease-containing antiphage systems in *Escherichia coli* and the counterdefense roles of bacteriophage T4 genome modifications. *J. Virol.*, **97**, e0059923.
60. Costa,A.R., van den Berg,D.F., Esser,J.Q., Muralidharan,A., van den Bossche,H., Estrada Bonilla,B., van der Steen,B.A., Haagsma,A.C., Nobrega,F.L. and Haas,P.-J. (2022) Accumulation of defense systems drives panphage resistance in *Pseudomonas aeruginosa*. bioRxiv doi: https://doi.org/10.1101/2022.08.12.503731, 12 August 2022, preprint: not peer reviewed.
61. Cheng,R., Huang,F., Lu,X., Yan,Y., Yu,B., Wang,X. and Zhu,B. (2023) Prokaryotic Gabija complex senses and executes nucleotide depletion and DNA cleavage for antiviral defense. bioRxiv doi: https://doi.org/10.1101/2023.05.02.539174, 03 May 2023, preprint: not peer reviewed.
62. Tuck,O.T., Adler,B.A., Armbruster,E.G., Lahiri,A., Hu,J.J., Zhou,J., Pogliano,J. and Doudna,J.A. (2024) Hachiman is a genome integrity sensor. bioRxiv doi: https://doi.org/10.1101/2024.02.29.582594, 29 February 2024, preprint: not peer reviewed.
63. Stokar-Avihail,A., Fedorenko,T., Hör,J., Garb,J., Leavitt,A., Millman,A., Shulman,G., Wojtania,N., Melamed,S. and Amitai,G. (2023) Discovery of phage determinants that confer sensitivity to bacterial immune systems. *Cell*, **186**, 1863–1876.
64. LeRoux,M., Srikant,S., Teodoro,G.I., Zhang,T., Littlehale,M.L., Doron,S., Badiee,M., Leung,A.K., Sorek,R. and Laub,M.T. (2022) The DarTG toxin–antitoxin system provides phage defence by ADP-ribosylating viral DNA. *Nat. Microbiol.*, **7**, 1028–1040.
65. Gordeeva,J., Morozova,N., Sierro,N., Isaev,A., Sinkunas,T., Tsvetkova,K., Matlashov,M., Truncaitė,L., Morgan,R.D. and Ivanov,N.V. (2019) BREX system of *Escherichia coli* distinguishes self from non-self by methylation of a specific DNA site. *Nucleic Acids Res.*, **47**, 253–265.
66. Gu,Y., Li,H., Deep,A., Enustun,E., Zhang,D. and Corbett,K.D. (2023) Bacterial Shedu immune nucleases share a common enzymatic core regulated by diverse sensor domains. bioRxiv doi: https://doi.org/10.1101/2023.08.10.552793, 10 August 2023, preprint: not peer reviewed.
67. Loeff,L., Walter,A., Rosalen,G.T. and Jinek,M. (2023) DNA end sensing and cleavage by the Shedu anti-phage defense system. bioRxiv doi: https://doi.org/10.1101/2023.08.10.552762, 11 August 2023, preprint: not peer reviewed.
68. Roberts,R.J., Belfort,M., Bestor,T., Bhagwat,A.S., Bickle,T.A., Bitinaite,J., Blumenthal,R.M., Degtyarev,S.K., Dryden,D.T. and Dybvig,K. (2003) A nomenclature for restriction enzymes, DNA methyltransferases, homing endonucleases and their genes. *Nucleic Acids Res.*, **31**, 1805–1812.
69. Makarova,K.S., Wolf,Y.I. and Koonin,E.V. (2022) Evolutionary classification of CRISPR-Cas systems. In: Barrangou,R., Sontheimer,E.J. and Marraffini,L.A. (eds.) *CRISPR: Biology and Applications*. pp.13–38.
70. Kuzmenko,A., Oguienko,A., Esyunina,D., Yudin,D., Petrova,M., Kudinova,A., Maslova,O., Ninova,M., Ryazansky,S. and Leach,D. (2020) DNA targeting and interference by a bacterial Argonaute nuclease. *Nature*, **587**, 632–637.
71. Swarts,D.C., Hegge,J.W., Hinojo,I., Shiimori,M., Ellis,M.A., Dumrongkulraksa,J., Terns,R.M., Terns,M.P. and Van Der Oost,J. (2015) Argonaute of the archaeon *Pyrococcus furiosus* is a DNA-guided nuclease that targets cognate DNA. *Nucleic Acids Res.*, **43**, 5120–5129.
72. Swarts,D.C., Jore,M.M., Westra,E.R., Zhu,Y., Janssen,J.H., Snijders,A.P., Wang,Y., Patel,D.J., Berenguer,J. and Brouns,S.J. (2014) DNA-guided DNA interference by a prokaryotic Argonaute. *Nature*, **507**, 258–261.
73. Song,X., Lei,S., Liu,S., Liu,Y., Fu,P., Zeng,Z., Yang,K., Chen,Y., Li,M. and She,Q. (2023) Long-B prokaryotic Argonaute systems employ various effectors to confer immunity via abortive infection. bioRxiv doi: https://doi.org/10.1101/2023.03.09.531850, 09 march 2023, preprint: not peer reviewed.
74. Adair,K.L. and Douglas,A.E. (2017) Making a microbiome: the many determinants of host-associated microbial community composition. *Curr. Opin. Microbiol.*, **35**, 23–29.
75. 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.
76. Vassallo,C.N., Doering,C.R., Littlehale,M.L., Teodoro,G.I. and Laub,M.T. (2022) A functional selection reveals previously undetected anti-phage defence systems in the *E. coli* pangenome. *Nat. Microbiol.*, **7**, 1568–1579.
77. Iranzo,J., Lobkovsky,A.E., Wolf,Y.I. and Koonin,E.V. (2013) Evolutionary dynamics of the prokaryotic adaptive immunity system CRISPR-Cas in an explicit ecological context. *J. Bacteriol.*, **195**, 3834–3844.
78. Drake,J.W. (2009) Avoiding dangerous missense: thermophiles display especially low mutation rates. *PLoS Genet.*, **5**, e1000520.
79. Zeldovich,K.B., Chen,P. and Shakhnovich,E.I. (2007) Protein stability imposes limits on organism complexity and speed of molecular evolution. *Proc. Natl Acad. Sci. USA*, **104**, 16152–16157.
80. Parkar,S., Flint,S. and Brooks,J. (2003) Physiology of biofilms of thermophilic bacilli—potential consequences for cleaning. *J. Ind. Microbiol. Biotechnol.*, **30**, 553–560.
81. Zablocki,O., van Zyl,L. and Trindade,M. (2018) Biogeography and taxonomic overview of terrestrial hot spring thermophilic phages. *Extremophiles*, **22**, 827–837.

Received: May 23, 2024. **Revised:** July 15, 2024. **Editorial Decision:** August 1, 2024. **Accepted:** August 2, 2024