Published in final edited form as:
Science. 2018 March 02; 359(6379): . doi:10.1126/science.aar4120.

Systematic discovery of anti-phage defense systems in the microbial pan-genome

Shany Doron[1,*], **Sarah Melamed**[1,*], **Gal Ofir**[1], **Azita Leavitt**[1], **Anna Lopatina**[1], **Mai Keren**[1], **Gil Amitai**[1], and **Rotem Sorek**[1,#]

[1]Department of Molecular Genetics, Weizmann Institute of Science, Rehovot 76100, Israel

Abstract

The arms race between bacteria and phages led to the development of sophisticated anti-phage defense systems, including CRISPR-Cas and restriction-modification systems. Evidence suggests that unknown defense systems are located in "defense islands" in microbial genomes. We comprehensively characterized the bacterial defensive arsenal by examining gene families that are clustered next to known defense genes in prokaryotic genomes. Candidate defense systems were systematically engineered and validated in model bacteria for their anti-phage activities. We report nine previously unknown anti-phage and one anti-plasmid systems that are widespread in microbes and strongly protect against foreign invaders. These include systems that adopted components of the bacterial flagella and condensin complexes. Our data also suggest a common, ancient ancestry of innate immunity components shared between animals, plants and bacteria.

Introduction

Bacteria and archaea are frequently attacked by viruses (phages), and as a result have developed multiple, sophisticated lines of active defense (1–3) that can collectively be referred to as the prokaryotic "immune system". Anti-phage defense strategies include restriction-modification (R-M) systems that target specific sequences on the invading phage (4), CRISPR-Cas, which provides acquired immunity through memorization of past phage attacks (5), abortive infection systems (Abi) that lead to cell death or metabolic arrest upon infection (6), and additional systems whose mechanism of action is not yet clear such as BREX (7), prokaryotic Argonautes (pAgos) (8) and DISARM (9). Different bacteria encode different sets of defense systems: CRISPR-Cas systems are found in about 40% of all sequenced bacteria (10, 11), R-M systems are found in about 75% of prokaryote genomes (12) while pAgos and BREX appear in about 10% (7, 13). It has been suggested that many currently unknown defense systems reside in genomes and plasmids of non-model bacteria and archaea and await discovery (2, 14).

Anti-phage defense systems were found to be frequently physically clustered in bacterial and archaeal genomes such that, for example, genes encoding restriction enzymes commonly reside in the vicinity of genes encoding abortive infection systems and other

[#]Correspondence: rotem.sorek@weizmann.ac.il.
[*]These authors contributed equally

Europe PMC Funders Author Manuscripts

phage resistance systems (14, 15). The observation that defense systems are clustered in genomic "defense islands" has led to the suggestion that genes of unknown function residing within such defense islands may also participate in anti-phage defense (15, 16). Indeed, recent studies that focused on individual genes enriched next to known defense genes resulted in the discovery of new systems that protect bacteria against phages (7, 9, 17).

Results

Identification of putative defense gene families

We have set out to comprehensively identify new defense systems enriched within defense islands, in attempt to systematically map the arsenal of defense systems that are at the disposal of bacteria and archaea in their fight against phages. As a first step in this discovery effort we sought to identify gene families that are enriched near known defense systems in the microbial pan-genome. For this, we analyzed 14,083 protein families (pfams) in >45,000 available bacterial and archaeal genomes (overall encoding >120 million genes). Each pfam represents a set of genes sharing a common protein domain (18). We calculated, for each pfam, the tendency of its member genes to reside in the vicinity of one or more known defense genes (Figure 1A&B; Methods). We further selected pfams that at least 65% of their member genes were found next to defense genes, and that their member genes appeared in diverse defense contexts within different genomes (at least 10% variability; Figure 1C). These thresholds were selected as they capture the majority of pfams that comprise known defense systems, e.g., restriction enzymes and Abi genes (Figure 1B&C; Table S1; Methods). The resulting set of 277 candidate pfams was supplemented with 35 non-pfam gene families that were previously predicted to be associated with known defense systems (15), as well as 23 pfams that were predicted in the same study as putatively defensive but did not pass our thresholds, altogether yielding a list of 335 candidate gene families (Table S2).

From defense genes to defense systems

Anti-phage defense systems are usually composed of multiple genes that work in concert to achieve defense – for example, *cas1*, *cas2*, *cas3* and the cascade genes in type I CRISPR-Cas systems (19), and the R, M and S genes in type I restriction-modification systems (3). Genes functioning within the same defense system are frequently encoded on the same operon, and the gene order within the operon is highly conserved among distantly related organisms sharing the same system (3, 7, 9, 16, 19, 20). To check whether the defense-associated pfams belong to multi-gene systems, we used each such pfam as an anchor around which we searched for commonly associated genes (Figure 1A). For this, we collected all the neighboring genes (10 genes from each side) from all the genomes in which members of the anchor pfam occurred, and clustered these genes based on sequence homology (Methods). We then searched for cassettes of gene clusters that, together with the anchor gene, show conserved order across multiple different genomes, marking such cassettes as candidate multi-gene systems (Methods; Figure 1A).

The gene annotations in the resulting candidate systems were manually inspected in order to filter out likely false predictions. We found that 39% of the cases (129/335) represented non-

defense, mobile genetic elements, such as transposons and integrases, that are known to co-localize with defense islands (15) (Table S2). Additional 30% (102/335) represented known defense systems whose pfams were not included in our original set of known defense pfams, and 17% belonged to operons probably performing metabolic or other functions not associated with defense (Figure S1A). The remaining systems possibly represent putative new defense systems. To expand our predictions with new pfams that may be specifically enriched next to the putative new defense systems, a second prediction cycle was performed, this time adding the members of the predicted new systems to the positive defense pfam set (Figure 1A; Figure S1B; Methods). Altogether, 41 candidate single-gene or multi-gene systems were retrieved from the two prediction cycles of this analysis (Table S3). We further filtered from this set systems that were largely confined to a specific taxonomic clade (e.g., systems appearing only in cyanobacteria), resulting in a set of 28 candidate systems that showed broad phylogenetic distribution.

Experimental verification strategy

We selected two bacteria, *Escherichia coli* str. MG1655 and *Bacillus subtilis* str. BEST7003, as model organisms to experimentally examine whether the predicted systems confer defense against phages (Figure 2A). None of the candidate new systems are naturally present in the genomes of these two bacterial strains. For each candidate system we selected source organisms from which the system was taken and heterologously cloned into one of the model organisms. To increase the probability that the cloned system would be compatible and functionally expressed within the receiving bacterium, we selected systems from mesophilic organisms as close phylogenetically as possible to *E. coli* or to *B. subtilis*, and included the upstream and downstream intergenic regions so that promoters, terminators or other regulatory sequences would be preserved. Where possible we took at least two instances of each system (from two different source genomes), to account for the possibility that some systems may not be active in their source organism (21, 22). The DNA of each system, spanning the predicted genes and the intergenic spaces, was synthesized or amplified from the source genome and cloned into the phylogenetically closest model organism - either to *E. coli* (on a plasmid) or to *B. subtilis* (genomically integrated). As a control, we repeated the procedure with 5 known defense systems (instances of types I, II and III R-M systems, a type III toxin/antitoxin system and an abortive infection gene of the AbiH family) for which source organisms were similarly selected and cloning was performed into *B. subtilis*, as well as a 6th control comprised of the recently discovered DISARM defense system (9) (Table S4).

Altogether, we attempted to heterologously clone 61 representative instances of the 28 candidate new systems, and successful cloning was verified by whole genome sequencing (Table S4). For 27 of these 28 systems there was at least one candidate locus for which cloning was successful, and RNA-seq of the transformants showed that for 26 of the systems, at least one of the candidate loci was expressed in the receiving *E. coli* or *B. subtilis* strain.

The engineered bacteria were then challenged by an array of phages consisting of 10 *B. subtilis* and 6 *E. coli* phages, spanning the three major families of tailed dsDNA phages

(myo-, sipho- and podo-phages), as well as one ssDNA phage infecting *E. coli* (Figure 2B-C). Measuring phage efficiency of plating (EOP) on system-containing bacteria vs. control cells, we found that 9 of the 26 tested systems (35%) showed protection from infection by at least one phage (Figure 2B-C; Figures S2-S3). In comparison, three of the six positive control systems showed defense, with the remaining 3 showing no protection against the 10 *B. subtilis* phages tested (see Discussion).

We named the 9 verified new systems after protective deities from various world mythologies. These defense systems comprise between 1 and 5 genes and span between 2 and 12 kb of genomic DNA (Table 1; Table S5). Where possible, we verified system consistency by testing for phage resistance in systems where individual genes were deleted (Figures 3-5; Figures S4-S5). We found between several hundreds and several thousand representations of each of the defense systems in sequenced microbial genomes, usually with broad phylogenetic distribution (Figure S6; Tables S6-S15). Most systems were detected in >10 taxonomic phyla, and 7 of them appear in archaea (Figure S6). Some of the systems seem to target a specific family of phages (e.g., the Thoeris system appears to specifically protect from myophages), while others, such as the Hachiman system, provide broader defense (Figure 2B). The genes comprising the new systems encode many protein domains that are commonly present in antiviral systems such as CRISPR-Cas and RNAi, including helicases, nucleases, and nucleic acid binding domains, in addition to many domains of unknown function and also atypical domains as described below. Three of the systems contain membrane-associated proteins as predicted by the presence of multiple transmembrane helices. Below we focus on further functional analyses for a selected set of systems.

The Zorya defense system

The Zorya system (named after a deity from Slavic mythology) was identified based on the enrichment of the anchor pfam15611, representing a domain of unknown function, within defense islands. Pfam15611-containing gene clusters were previously reported as genomically associated with tellurium- and stress-resistance genes (23). The reconstructed system is comprised of the 4 genes *zorABCD*, overall encompassing ~9kb of DNA, with pfam15611 being the third gene in the system (*zorC*; Figure 3C; Table 1). A representative Zorya operon from *E. coli* E24377A was cloned into *E. coli* MG1655 and provided 10-10000 fold protection against infection by T7, SECphi27 and lambda-vir phages (Figure 3; Figure S3). Further searches based on homologies to the first two genes of the system, *zorA* and *zorB*, revealed a second type of Zorya, comprised of the 3 genes *zorABE*. A Type II Zorya was cloned from *E. coli* ATCC8739 into *E. coli* MG1655 and provided defense against T7 and the ssDNA phage SECphi17 (Figure 2C, 3B-3C).

The first two genes of the Zorya system, *zorA* and *zorB*, contain protein domains sharing distant, but clear homology with domains in *motA* and *motB*, respectively (Figure 3C). MotA and MotB are inner membrane proteins that are part of the flagellar motor of bacteria. They assemble into a MotAB complex, which forms the stator of the flagellar motor (the static part within which the flagellar rotor swivels) (24). The MotAB complex also forms the proton channel that provides the energy for flagellar rotation, coupling transport of protons

into the cell with the rotation (Figure 3D) (25, 26). While *zorB* shares the same size and domain organization with *motB* (including the pfam13677 and pfam00691 domains), *zorA* contains, in addition to the MotA domain (pfam01618), a long C-terminal helical domain that is sometime identified as a "methyl-accepting chemotaxis domain" (COG0840). In addition to these two genes Type I Zorya contains *zorC*, a gene of unknown function, and *zorD*, which encodes a large protein (1200aa) with a helicase domain that in some cases also encodes a C-terminal Mrr-like nuclease domain. Type II Zorya lacks *zorC* and *zorD* and instead contains *zorE*, a smaller gene encoding an HNH-endonuclease domain.

The gene composition of the Zorya system may point to several hypotheses as to its mechanism of action. It is possible that the system has adopted the MotAB proton channel to achieve depolarization of membrane potential upon phage infection. Possibly, ZorC, ZorD and ZorE may be involved in the sensing and inactivation of phage DNA, and if phage inactivation fails, the ZorAB proton channel opens up, leading to membrane depolarization and cell death. Under this hypothesis Zorya may be a conditional abortive infection system. Indeed, while Zorya-containing cells that were infected by phage T7 did not yield phage progeny in >80% of infection events, infection of Zorya-containing cells in liquid cultures has led to an eventual culture collapse, suggesting that Zorya-mediated defense involves death or metabolic arrest of the infected cells (Figure S7).

We further experimented with mutated forms of Type I Zorya. All four genes in the system appear to be essential for its functionality, as deletion of each of the genes resulted in loss of protection from phage infection (Figure 3E). Moreover, the activity of the ZorAB putative proton channel is necessary for system's functionality, as point mutations in residues predicted to be critical for proton translocation through the channel (either ZorA:T147A/S184A or ZorB:D26N) yielded a non-functional system (Figure 3E). Similarly, point mutations inactivating the Walker B motif of the ZorD helicase domain, predicted to prevent ATP hydrolysis, resulted in loss of protection from phage infection.

We identified 1829 instances of the Zorya system within 1663 sequenced bacterial genomes, belonging to 12 phyla, marking this system as prevalent in at least 3% of sequenced bacteria (Figure S6; Table S12). We did not find the system in archaea. The system is enriched in Proteobacteria and is markedly under-represented in Gram positive bacteria (Firmicutes and Actinobacteria; Figure S6), suggesting that its functionality may depend on the double membrane organization of Gram negative bacteria, or on differences between flagellar organization of Gram negative and Gram positive bacteria. As the Zorya system protects against phages that do not use flagella as their receptor (e.g. T7 (27)), Zorya protection is unlikely to stem from a receptor-masking effect.

The Thoeris defense system

Thoeris (Egyptian protective deity of childbirth and fertility) is a system that was detected based on the enrichment of pfam08937 (TIR domain, acronym for Toll-Interleukin Receptor) next to known defense genes (Figure 4A). This domain was previously reported as associated with prokaryotic argonaute genes (28). The first gene in the Thoeris system, denoted *thsA*, has an NAD-binding domain that is sometimes annotated as sirtuin (SIR2)-like domain or Macro domain. The second gene, *thsB*, contains the TIR domain, and can

appear in one or more copies (Figure 4A). In some Thoeris versions, ThsA has a multi-transmembrane N-terminal domain. Two instances of this system, one from *Bacillus amyloliquefaciens* Y2 (where ThsA is predicted to be membrane-associated) and the other from *Bacillus cereus* MSX-D12 (ThsA predicted as cytoplasmic), were engineered into *B. subtilis* BEST7003 and were found to confer defense against myophages (Figure 2; Figure 4B&C; Figure S2). As the 3 myophages we tested are very different from each other and share few homologous genes, it is possible that the Thoeris system senses or targets a general feature in the biology of myophages rather than a specific sequence or genome modification. Both Thoeris genes, *thsA* and *thsB*, are essential in the system as deletion of either of them rendered the system inactive (Figure 4C).

Interestingly, the TIR domain is an important component of the innate immune systems of mammals, plants and invertebrates, where it mainly serves as a connector domain that transfers the immune signal once a molecular pattern of an offensive pathogen is sensed (29). In animals this domain frequently forms the intracellular portion of membrane-bound Toll-like receptors, whereas in plants it is often connected to intracellular R genes (30) and can also be involved in direct recognition of pathogens (31, 32). Our finding marks a common involvement of TIR domains in innate immunity across the three domains of life, and implies that the ancestry of this important component of eukaryotic innate immune systems may have stemmed from prokaryotic defense against phages.

The Thoeris system is broadly distributed in bacteria and archaea, and can be detected in at least 4% of the sequenced genomes we analyzed (2070 genomes; Table S6; Figure S6). The TIR domain gene, *thsB*, has a strong tendency (52% of cases) to appear in multiple, diverse copies clustered around the *thsA* gene (Figure 4A; Table S6). Presence in multiple copies is typical to specificity-conferring genes in defense systems (such as the S subunit in type I R-M systems), where duplication followed by diversification serves for multiple specificities of the system (33–35). It is therefore possible that the TIR domain gene is responsible for identification of specific phage patterns, with multiple TIR domain genes serving for recognition of different phage components. Under this hypothesis, it is tempting to suggest that Thoeris is the prokaryotic ancestral form of pathogen-associated molecular pattern (PAMP) receptors.

A recent study showed that TIR domains can have enzymatic NAD+ hydrolase activities (36), which is in line with predictions that these domains process nucleotide derivatives (37). In *C. elegans*, this activity was shown to be involved in anti-fungal and anti-bacterial defense (38), while in animal neurons NAD+ hydrolysis by the SARM1 TIR domain-containing gene leads to NAD+ depletion and generation of linear and cyclic ADP-ribose signaling molecules that regulate axonal degeneration (39). An E99A point mutation in the *B. amyloliquefaciens* Y2 ThsB protein, which aligns with the catalytic residue in the SARM1 NAD-cleaving TIR domain (Figure S8) abolished the protective activity of Thoeris (Figure 4C). Moreover, point mutations in the ThsA NAD+ binding site, predicted to abolish NAD+ binding, also resulted in system inactivation (ThsA:N112A and ThsA:D100A/N115A for the *B. cereus* and *B. amyloliquefaciens* systems, respectively). These results suggest NAD+ binding and hydrolysis as essential for the anti-phage activity of the Thoeris system.

The Druantia system

Another system worth discussing briefly is the Druantia system (named after a deity from Gallic mythology). This system is characterized by a gene encoding a very large protein (~1800-2100aa) containing a domain of unknown function (DUF1998) as well as a helicase signature and a Walker A/B motif suggestive of ATP utilization. This large gene is typically preceded by a set of highly variable genes with no recognizable domains or function prediction – either 3 genes sized 350-600aa each (Type I), or two genes sized 700-900aa (Type II), or a single large gene of 1000-1200aa (Type III) (Figure S5A&B). In some cases Type I systems are preceded by a gene annotated as DUF4338, encoding yet another domain of unknown function; and Type II systems are also associated with a cytosine methylase (Figure S5A&B). A Type I system cloned from *E. coli* UMEA 4076-1 into *E. coli* MG1655 rendered the engineered strain resistant against 4 of the 6 phages tested, and by serially deleting four of the genes in this system we verified that all four are essential for its activity (Figure S5C). Notably, DUF1998-containing genes are among the components of the recently reported DISARM (9) and Dpd (40) defense systems, where their function is also unknown. The sheer size of the Druantia system (12kb of genomic DNA) suggests a complex function, and the near-complete absence of recognizable domains in its genes suggests a new mode of defense not shared by prokaryotic defense systems whose mechanism is currently understood.

Defense against plasmid transformation

Some of the putative defense systems we experimentally tested did not show any anti-phage activity despite being strongly associated with known defense genes. We reasoned that some of these systems may defend against other forms of foreign DNA. To test this hypothesis we selected one such system, which we hereby denote Wadjet (god protector of ancient Egypt) for further experimentation. Wadjet is a 4-gene system, *jetABCD*, which is common in microbial genomes and is very frequently found next to defense genes (Figure 5A). Three instances representing three different types of Wadjet (see below) were cloned from three separate *Bacillus* species into *B. subtilis* BEST7003. While none of these systems provided protection against any of the 10 *Bacillus* phages in our array, all three consistently and significantly reduced transformation efficiency of the episomal plasmid pHCMC05 (Figure 5C). These results suggest that Wadjet may be a defense system specifically targeting foreign plasmids.

We identified three different domain compositions, each encoding a different set of pfams, but all with common sequence signatures marking them as three types of Wadjet (Figure 5B). While the pfam domains of Wadjet genes are mostly defined as “domains of unknown function”, structural modeling using Phyre2 (41) showed structural homology between JetA, JetB and JetC and genes belonging to the housekeeping condensin system MukF, MukE and MukB, respectively. Bacterial condensins are chromosome-organizing complexes that are responsible for DNA condensation and accurate segregation during replication (42), and mutations in the housekeeping condensins lead to severe defects in chromosome segregation and viability (43). Several versions of housekeeping codensins appear in bacterial genomes: SMC, MukBEF and MksBEF (44); the Wadjet system was previously noted as a distant homolog of the MksBEF system described in *P. aeruginosa* (45).

While the domain organization of the jetABC genes may lead to the hypothesis that Wadjet is an alternative condensin system involved in bacterial chromosome maintenance, our data imply that its role is defensive. This system is highly enriched within defense islands, undergoes extensive horizontal gene transfer, and is only sporadically found within strains of the same species, all of which is inconsistent with a core, essential role in chromosome maintenance. We hypothesize that the Wadjet system has been adapted from a MukBEF condensin ancestor to become a defense system. Possibly, the system identifies foreign plasmids and uses its condensin properties to interfere with proper plasmid segregation into daughter cells. Notably, plasmid transformation in *B. subtilis* takes place via the natural competence of this organism, during which the plasmid DNA is transformed to the cell through dedicated transporters as single-stranded DNA (ssDNA) (46). It is possible that the Wadjet system protects against rampant natural transformation or, alternatively, may specifically target ssDNA phages. However, as no ssDNA phage was reported for *B. subtilis*, we were not able to test whether ssDNA phages are specifically blocked by the Wadjet systems cloned in *B. subtilis* BEST7003.

The Wadjet system is broadly spread in bacterial and archaeal genomes (found in ~6% of the genomes we studied), where it presents high sequence diversity (Table S15; Figure S6). Deletion of each of the four genes in Type I Wadjet from *B. cereus* Q1 abolished its activity and restored plasmid transformation, indicating that each of the genes is essential for anti-plasmid defense (Figure 5C). Moreover, point mutations E59K/K60E in JetB, predicted to disrupt the MukE-MukF-like protein-protein interactions, resulted in loss of protective activity against plasmids, and so has the E1025Q mutation in the Walker B motif of JetC that is predicted to abolish ATPase activity. The JetD gene, which has no homology to genes in the Muk system, has a putative topoisomerase VI domain based on structural predictions; a point mutation JetD:E226A, predicted to diminish binding of the topoisomerase VI domain to DNA, also abolished the protective activity of the system.

Discussion

Our study significantly expands the known arsenal of defense systems used by prokaryotes for protection against phages. However, our results do not yet expose the complete set of prokaryotic defense systems. Out of the 26 candidate systems we tested, nine were verified as anti-phage defense systems and an additional one showed protection against plasmids. The remaining 16, although not verified by our experiments, do not necessarily represent false predictions, as exemplified by the fact that only 50% of our positive control systems showed defense in our assays. Lack of activity of positive control systems or candidate systems could possibly stem from incompatibility of some tested systems with the recipient organism (*E. coli* or *B. subtilis*), or due to pseudogenization of some systems in their genome of origin. Some systems may be highly specific against a certain type of phages or foreign genetic element not represented in our phage set, while others may work in a specific condition not tested in our study. Clade-specific potential systems such as those found only in archaea or cyanobacteria (Table S3) were not tested in this study and can represent a more specialized defense arsenal unique only to a subset of organisms. Finally, we may have missed some true systems by falsely tagging them as belonging to the “mobilome” (Table

S2), as mobile genetic elements have an intimate evolutionary relationship with defense systems (47).

In the past, the discovery and mechanistic understanding of anti-viral defense systems led to the development of important biotechnological tools. For example, the discovery of restriction enzymes resulted in a revolution in genetic engineering, and CRISPR-Cas now revolutionizes the genome editing field. Eukaryotic immune systems, such as RNAi and antibodies, have also become widely utilized tools. The tendency of defense systems to turn into revolutionary molecular tools stems from their intrinsic high degree of flexible molecular specificity (to differentiate between self and non-self), as well as their inherent capability to target the identified molecule. One may envision that some of the new systems we discovered, once their mechanism is deciphered, may also be adapted into useful molecular tools in the future.

Materials and Methods

Computational prediction of defense systems

A set of gene families known to participate in defense—A set of pfams and COGs that are known to participate in anti-phage defense was compiled based on the gene families present in Table S10 from Makarova *et al.* 2011 (15) with the addition of pfams/COGs present in the BREX (7) and DISARM (9) anti-phage systems. This set is found in Table S1.

Identification of pfams enriched near defense genes—The genome sequences, gene annotations and taxonomy annotations of all publicly available sequenced bacterial and archaeal genomes were downloaded from the NCBI FTP site (ftp.ncbi.nih.gov/genomes/genbank/bacteria/ and ftp.ncbi.nih.gov/genomes/genbank/archaea/, respectively) on April 2016. Pfam annotations for bacterial and archaeal genes were obtained from the Integrated Microbial Genomes (IMG) database (48) on December 2015, and cross-referenced to the genes in the genomes downloaded from NCBI using the locus_tag Genbank field. All pfams annotated in at least 20 genes ("members") across the analyzed genomes (14,083 pfams) were scanned. For each pfam, the number of member genes for which a gene having an annotation of a known defense gene family (Table S1) was present in proximity (up to ten genes upstream and ten genes downstream) was recorded. The fraction of defense-associated members out of total members ("defense score") was calculated per pfam. A second score ("defense context variability score") was calculated for each pfam as follows: for each member gene occurring with at least one defense gene in proximity, a list of the proximal defense genes was recorded, and the fraction of unique lists out of total number of lists for that pfam represents the score (for example: if pfamX is found within 20 genes in our set, with 15 of them having Cas9 nearby and 5 having type I R-M nearby, the number of unique lists is 2, and the "defense context variability score" is 2/20 = 0.1). Pfams with defense score >= 65% and defense context variability score >= 0.1 were taken for further analysis. This list was supplemented with 35 non-pfam gene families that were predicted to be associated with defense by Makarova *et al.* 2011 (15), as well as 23 pfams that were predicted in the same study but did not pass the thresholds above (Table S2).

From genes to systems—Each of the putative defense-related gene families was used as an anchor to search for multi-gene systems, as follows. The protein coding sequences for neighboring genes (+/-10 genes) for all family members were clustered based on sequence homology (for example, if pfamY is found within 50 genomes in our set, the 20 neighboring genes in each genome, plus the pfamY gene in each genome, were taken – altogether 50*21 = 1050 genes to be clustered). Clustering was done with OrthoMCL software v2.0.9 (49) with blastp parameters [-F 'm S' -v 100000 -b 100000 -e 1e-5 -m 8] and with mcl v12.068 downloaded from http://micans.org/mcl/ (50, 51) with inflation value of 1.1. When the number of blastp hits for a given anchor pfam was too large and prohibitive for OrthoMCL to generate clusters (>75 million blastp hits), a subset of genomes, containing only bacterial and archaeal genomes annotated as "complete" (rather than "draft") was used for clustering.

To detect the most prevalent genes around the anchor pfam, only the 10% largest clusters ("frequent clusters") were considered. For the sake of cluster size calculation, genes originating from the same species (derived from the strain name in the NCBI annotation) were counted as one gene, to prevent organisms for which many strains have been sequenced from inflating the cluster size. An edge between cluster(i) and cluster(j) was defined if a gene from cluster(j) followed a gene from cluster(i) in a given genome with no other genes belonging to frequent clusters found in between, with edge weight ("thickness") defined as the number of such adjacency cases. Again, edge weights were adjusted such that multiple appearances of a cluster pair originating from the same species were recorded as a single appearance. Only the 10% thickest edges were retained for further analysis. In each genome, the maximal "path" that included the anchor pfam gene and was composed of the retained (largest) clusters and the retained (thickest) edges was recorded. Such a "path", representing a set of genes appearing in a conserved order in multiple genomes, was considered a candidate multi-gene system. Infrequent variations on the gene order and composition of common systems were merged into the common system if they shared at least 50% of their clusters and had less than 25% appearances than the common system. Only systems with five or more appearances from different species were further analyzed.

The domains within the gene members of each system were analyzed bioinformatically using the tools HHpred (52, 53), Phyre2 (41), PSI-BLAST (54) and NCBI's Conserved Domain Database (CDD) (55). The systems were then manually filtered, based on this analysis, to remove (i) known defense systems whose domains did not appear in our initial set of gene families known to participate in defense; (ii) systems likely representing mobile genetic elements ("mobilome") and; (iii) systems likely participating in non-defensive functions or house-keeping systems (Table S2).

A second cycle of prediction was then performed, expanding the set of "positive" gene families from Table S1 to include the gene families participating in the candidate new defense systems, as well as the gene families participating in known defense systems that were previously missing from our set and detected in the first round. All pfams were again scanned and the same thresholds were applied (defense score 65%, context variability score 0.1). New pfams retrieved from the second cycle were analyzed as above to generate and annotate multi-gene systems.

Candidate new systems were further prioritized to select instances for experimental validations. Systems tagged as “questionable”, due to uncertainty whether they represent defense genes or mobile genetic elements, were filtered out (Table S3). Systems existing in only a narrow range of organisms, as well as systems that were not found in organisms phylogenetically close either to *E. coli* or *B. subtilis*, were not tested experimentally (Table S3).

For system selection for experimental testing, we first attempted to select candidate systems from organisms close to *B. subtilis* as the receiving model organism, as in this organism genomic integration of large fragments of DNA is straightforward and results in a single-copy addition of the system. In case no source organisms sufficiently close to *B. subtilis* were found, we switched to *E. coli* as the model organism for experimentation.

Phylogenetic distribution analysis of new systems—For each validated defense system, several loci including the locus that was experimentally verified were taken as seeds for psi-Blast. psi-Blast version 2.5.0 of BLAST+ (54, 56) with parameters [-num_iterations 10 -max_hsps 1 -max_target_seqs 100000 -evalue 1e-10] was performed for each protein of each system, against all microbial genomes downloaded from NCBI on April 2016. When the hits of all proteins of a system were found closely localized on a genome, spanning no more than 150% of the length of the original system, this genome was recorded as containing the system. For the Druantia and Wadjet systems, -evalue 1e-5 was used to enable detection of distant homologs. For systems with 4 or 5 genes (Zorya type I, Druantia types I and II, Wadjet), systems were reported if at least 3 of their genes were identified. For the Druantia system, systems with hits to the DruE protein were retained if the DruE size was >1300aa. For the Thoeris system, multiple *thsB* genes near the *thsA* gene were recorded if they were within 10 kb of genomic DNA around the identified *thsA*. Phylum for each genome was obtained using the JGI taxonomy server (https://taxonomy.jgi-psf.org/).

Experimental validation of defense systems

Cloning of candidate systems into *E. coli* MG1655 and *B. subtilis* BEST7003—A cloning shuttle vector for large fragments was constructed as previously described (9). The vector contains a p15a origin of replication and ampicillin resistance for plasmid propagation in *E. coli*, and *amyE* integration cassette with spectinomycin resistance for genomic integration into *B. subtilis*. The backbone of this vector was amplified using primers OGO309+OGO310, adding to it a BamHI restriction site and a terminator site upstream to the insert cloning site. The multiple cloning site of plasmid pBS1C (57), received from BGSC (accession ECE257), was amplified using primers OGO311+OGO312. Both fragments were digested using AscI and BamHI, ligated using T4 ligase and transformed into *E. coli*, resulting in plasmid pSG1-rfp.

The loci of most systems were commercially synthesized and cloned, by Genscript corp., directly into pSG1-rfp between the AscI and NotI sites of the multiple cloning site (Table S4, "Cloning method" column). In one case (the Type I Wadjet system) the DNA was synthesized by Gen9 (Boston, MA) with synonymous modifications to optimize GC content. In case the donor strains were readily available the system was not synthesized but instead

was directly amplified from the genomic DNA of the donor strain using KAPA HiFi HotStart ReadyMix (Kapa Biosystems KK2601) with primers as detailed in Table S16. For long systems (>10000 bases) when the donor strain was not available, the system was commercially synthesized in overlapping fragments (Table S4, "Cloning method" column). Systems amplified from genomic DNA or ordered as overlapping fragments were cloned into pSG1-rfp between the AscI and NotI sites using NEBuilder HiFI DNA Assembly cloning kit (NEB E5520S). The full list of sources used for cloning the systems into our model organisms is found in Table S4, including the accessions of all strains ordered.

Transformation to *B. subtilis* was performed using MC medium as previously described (58). MC medium was composed of 80 mM K_2HPO_4, 30 mM KH_2PO_4, 2% Glucose, 30 mM Trisodium citrate, 22 μg/ml Ferric ammonium citrate, 0.1% Casein Hydrolysate (CAA), 0.2% potassium glutamate. From an overnight starter of bacteria, 10 μl were diluted in 1 ml of MC medium supplemented with 10 μl 1M $MgSO_4$. After 3 hours of incubation (37 °C, 200 rpm), 300 μl was transferred to a new 15 ml tube and ~200 ng of plasmid DNA was added. The tube was incubated for another 3 hours (37 °C, 200 rpm), and the entire reaction was plated on LB agar plates supplemented with 100 μg/ml spectinomycin and incubated overnight at 30 °C.

For systems tested in *E. coli*, the cloned vector was transformed into *E. coli* MG1655 cells (ATCC 47076), and the resulting transformants were verified by PCR. For systems to be tested in *B. subtilis*, the cloned vector was transformed into *B. subtilis* BEST7003 cells, kindly provided previously by M. Itaya. The system was integrated into the *amyE* locus, and resulting transformants were screened on starch plates for amylase-deficient phenotype. Whole-genome sequencing was then applied to all transformed *B. subtilis* and *E. coli* clones as described in (9) to verify system's integrity and lack of mutations.

As a negative control for transformation into *B. subtilis*, a transformant with an empty plasmid, containing only the spectinomycin resistance gene in the *amyE* locus, was used. As a negative control for transformation into *E. coli*, the wild-type *E. coli* MG1655 carrying an empty plasmid was used.

For strains with gene deletions and point mutations, plasmids containing systems with these deletions/mutations were commercially synthesized by Genscript. The mutated systems were transformed into *B. subtilis* and *E. coli* as described above, and clones used were fully sequenced to verify proper integration and sequence of the mutated systems.

Phage strains, cultivation and plaque assay—The following *B. subtilis* phages were obtained from the Bacillus Genetic Stock Center (BGSC): SPO1 (BGSCID 1P4), phi3T (BGSCID 1L1), SPβ (BGSCID 1L5), SPR (BGSCID 1L56), phi105 (BGSCID 1L11), rho14 (BGSCID 1L15), and SPP1 (BGSCID 1P7). Phage phi29 was obtained from the DSMZ (DSM 5546). Phages SBSphiJ and SBSphiC were isolated by us from mixed soil and leaves samples on *B. subtilis* BEST7003. For this, soil and leaves samples were added to a log phase *B. subtilis* BEST7003 culture and incubated overnight to enrich for *B. subtilis* phages. The enriched sample was centrifuged and filtered through 0.2 μm filters, and the filtered supernatant was used to perform double layer plaque assays as described in

Kropinski et al. (59). Single plaques that appeared after overnight incubation were picked, re-isolated 3 times, and amplified as described below.

E. coli phages (T4, T7, lambda-vir) were kindly provided by U. Qimron. Phages SECphi17, SECphi18 and SECphi27 were isolated as described in Wommack et al. (60) from sewage samples on *E. coli* MG1655. 0.2 μm filtered concentrated sewage samples were used to perform double layer plaque assays, individual plaques were picked, reisolated 3 times, and amplified as described below.

All phages isolated by us were Illumina sequenced following a library prep using the Nextera protocol (61) and assembled using SPAdes v. 3.10.1 using the –careful and –cov-cutoff auto modifiers (62). Assembled genomes and raw reads were deposited in the European Nucleotide Archive (ENA) under study accession PRJEB23070. Phage classification was done according to sequence homology to the closest known similar phage. Phage SECphi17 (ENA ERS1981053) has a 5,538 bp genome and its closest relative is Coliphage WA3 (GenBank DQ079897.1, 66% coverage, 81% identity), indicating it is an ssDNA phage of the *Microviridae* family. Phage SECphi18 (ENA ERS1981054) has a 44,798 bp genome and its closest relative is *Escherichia* phage Gluttony (GenBank KX534336.1, 92% coverage, 93% identity), indicating it is a member of the *Siphoviridae* family. Phage SECphi27 (ENA ERS1981055) has a 51,811 bp genome, and its closest relative is *Escherichia* phage vB_Eco_swan01 (GenBank LT841304.1, 91% coverage, 98% identity), indicating it is a member of the *Siphoviridae* family. Phage SBSphiJ (ENA ERS1981056) has a 156,875 bp genome, and its closest relative is *Bacillus* phage Grass (GenBank KF669652.1, 91% coverage, 95% identity), indicating it is a member of the family *Myoviridae*. Phage SBSphiC (ENA ERS1981057) has a 144,651 bp genome, and its closest relative is Bacillus phage SP10 (GenBank AB605730.1, 94% coverage, 90% identity), indicating it is a member of the *Myoviridae* family. *Siphoviridae* and *Myoviridae* phage morphologies were verified by electron microscopy (EM). For the EM experiments, phage lysates were blotted onto copper grids, stained using uranyl acetate 2%, and visualized in FEI Tecnai T12 transmitting electron microscope.

Phages were propagated on either *E. coli* MG1655 or *B. subtilis* BEST7003 using the plate lysate method as previously described (63). Lysate titer was determined using the small drop plaque assay method as previously described (64). Bacteria were mixed with MMB agar (LB + 0.1 mM $MnCl_2$ + 5 mM $MgCl_2$ + 5 mM $CaCl_2$ + 0.5% agar), and serial dilutions of phage lysate in MMB were dropped on top of them. After the drops dried up, plates were incubated at room temperature overnight. Efficiency of plating (EOP) was measured by performing small drop plaque assay with the same phage lysate on control bacteria and bacteria containing the candidate defense system, and comparing the ratio of plaque formation.

To determine number of infective centers during infection with T7 phage of control bacteria and bacteria containing type I or type II Zorya, we used a modified version of the technique described in (65). Zorya-lacking *E.coli* MG1655 or Zorya-containing cells were infected with T7 phage at MOI 0.05 and incubated for 10 minutes at 37 °C to allow adsorption. Cells with adsorbed phages were then centrifuged (1 minute, 14000 rpm) at 4 °C, washed once with ice-cold MMB medium, and resuspended in 200 μl ice-cold MMB medium. Then, 100

µl aliquots of 10-fold dilutions of resuspended phage-infected cells were mixed with 100 µl of a Zorya-lacking *E.coli* MG1655 culture grown to O.D 0.3. The mixture was plated using the double agar overlay method and infection centers (plaques) were counted after overnight incubation in room temperature.

For the liquid culture infection with T7 phage, overnight cultures of Zorya-lacking *E.coli* MG1655 or Zorya-containing cells were diluted 1:100 in MMB medium. 180 µl volumes of the diluted culture were dispersed into wells in a 96-well plate and grown at 37 °C with vigorous shaking until early log phase ($O.D._{600}$ 0.3). 20 µl of T7 phage lysate were added at multiplicities of infection 0.05, 0.5 and 5 in three replicates. Optical density measurements at a wavelength of 600 nm were taken every 15 minutes using a TECAN Infinite 200 plate reader in a 96-well plate as previously described (9).

Transformation efficiency assay—Transformation was performed using the MC medium as described above. To test plasmid transformation efficiency, the episomal *Bacillus* plasmid pHCMC05 was used (66). Transformation efficiency was calculated by dividing the number of transformants that grew on LB plates containing 5 µg/ml chloramphenicol by the live count on LB plates.

DNA-seq and RNA-seq—DNA was extracted from bacteria using Qiagen DNeasy blood and tissue kit (Qiagen 69504). DNA libraries were constructed using the Nextera library preparation protocol as previously published (61). RNA-seq was performed with the NEBNext Ultra Directional RNA Library Prep Kit (NEB, E7420) according to the manufacturer's instructions with modifications as previously described (67). Prior to library preparation, equal amounts of extracted RNA from 3-7 strain samples were pooled together and processed as a single library. All libraries were sequenced using the Illumina NextSeq500. The sequencing reads were aligned to the reference genomes of *B. subtilis* BEST7003 (Genbank: AP012496) and *E. coli* MG1655 (Genbank: NC_000913), and to the plasmid sequence of each system, using Novoalign 3.02.02 (Novocraft Technologies Sdn Bhd, http://www.novocraft.com) with the default parameters and [-r Random]. The coverage along the reference genomes was calculated, to check if each system exists in the genome (DNA-seq) or expressed (RNA-seq). The pooled RNA library was sequenced to a depth of 5 million reads per sample and later aligned to the reference genomes as described.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We thank Hila Sberro, Omer Zuqert, Ofir Cohen, Sefi Mintzer, Rom Shenhav, Zohar Erez, Daniel Dar, Maya Voichek, and Nitzan Tal for useful discussion during the course of this study. We also thank Maya Voichek for assistance with RNA-seq, and Shahar Silverman for help with phage isolation. R.S. was supported, in part, by the Israel Science Foundation (personal grant 1360/16 and I-CORE grant 1796/12), the European Research Council (ERC) (grant ERC-CoG 681203), the Abisch-Frenkel foundation, the David and Fela Shapell Family Foundation, the Benoziyo Advancement of Science grant, the Minerva Foundation, and the Pasteur-Weizmann council. Assembled phage genomes and raw reads were deposited in the European Nucleotide Archive (ENA) under study accession PRJEB23070. Conflict of interest statement: R.S. is a scientific co-founder and advisor of BiomX Ltd.

S.D., S.M., G.A., A. Leavitt and R.S. are inventors on U.S. provisional patent application 62/586,911. S.D., S.M., G.O. and R.S. are inventors on U.S. provisional patent applications 62/512,216 and 62/512,219.

References

1. Labrie SJ, Samson JE, Moineau S. Bacteriophage resistance mechanisms. Nat Rev Microbiol. 2010; 8:317–27. [PubMed: 20348932]
2. Stern A, Sorek R. The phage-host arms race: shaping the evolution of microbes. Bioessays. 2011; 33:43–51. [PubMed: 20979102]
3. Dy RL, Richter C, Salmond GPC, Fineran PC. Remarkable Mechanisms in Microbes to Resist Phage Infections. Annu Rev Virol. 2014; 1:307–331. [PubMed: 26958724]
4. Tock MR, Dryden DT. The biology of restriction and anti-restriction. Curr Opin Microbiol. 2005; 8:466–472. [PubMed: 15979932]
5. Barrangou R, et al. CRISPR provides acquired resistance against viruses in prokaryotes. Science. 2007; 315:1709–12. [PubMed: 17379808]
6. Molineux IJ. Host-parasite interactions: recent developments in the genetics of abortive phage infections. New Biol. 1991; 3:230–236. [PubMed: 1831658]
7. Goldfarb T, et al. BREX is a novel phage resistance system widespread in microbial genomes. EMBO J. 2015; 34:169–83. [PubMed: 25452498]
8. Swarts DC, et al. DNA-guided DNA interference by a prokaryotic Argonaute. Nature. 2014; 507:258–61. [PubMed: 24531762]
9. Ofir G, et al. DISARM is a widespread bacterial defence system with broad anti-phage activities. Nat Microbiol. 2018; 3:90–98. [PubMed: 29085076]
10. Godde JS, Bickerton A. The repetitive DNA elements called CRISPRs and their associated genes: evidence of horizontal transfer among prokaryotes. J Mol Evol. 2006; 62:718–29. [PubMed: 16612537]
11. Kunin V, Sorek R, Hugenholtz P. Evolutionary conservation of sequence and secondary structures in CRISPR repeats. Genome Biol. 2007; 8:R61. [PubMed: 17442114]
12. Oliveira PH, Touchon M, Rocha EPC. The interplay of restriction-modification systems with mobile genetic elements and their prokaryotic hosts. Nucleic Acids Res. 2014; 42:10618–10631. [PubMed: 25120263]
13. Swarts DC, et al. The evolutionary journey of Argonaute proteins. Nat Struct Mol Biol. 2014; 21:743–753. [PubMed: 25192263]
14. Makarova KS, Wolf YI, Koonin EV. Comparative genomics of defense systems in archaea and bacteria. Nucleic Acids Res. 2013; 41:4360–77. [PubMed: 23470997]
15. Makarova KS, Wolf YI, Snir S, Koonin EV. Defense islands in bacterial and archaeal genomes and prediction of novel defense systems. J Bacteriol. 2011; 193:6039–56. [PubMed: 21908672]
16. Koonin EV, Makarova KS, Wolf YI. Evolutionary Genomics of Defense Systems in Archaea and Bacteria. Annu Rev Microbiol. 2017; 71:233–261. [PubMed: 28657885]
17. Depardieu F, et al. A Eukaryotic-like Serine/Threonine Kinase Protects Staphylococci against Phages. Cell Host Microbe. 2016; 20:471–481. [PubMed: 27667697]
18. Finn RD, et al. The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res. 2016; 44:D279–85. [PubMed: 26673716]
19. Makarova KS, et al. An updated evolutionary classification of CRISPR-Cas systems. Nat Rev Microbiol. 2015; 13:722–36. [PubMed: 26411297]
20. Yamaguchi Y, Park J-H, Inouye M. Toxin-Antitoxin Systems in Bacteria and Archaea. Annu Rev Genet. 2011; 45:61–79. [PubMed: 22060041]
21. Stern A, Keren L, Wurtzel O, Amitai G, Sorek R. Self-targeting by CRISPR: gene regulation or autoimmunity? Trends Genet. 2010; 26:335–40. [PubMed: 20598393]
22. Hoskisson PA, Smith MC. Hypervariation and phase variation in the bacteriophage “resistome”. Curr Opin Microbiol. 2007; 10:396–400. [PubMed: 17719266]

Europe PMC Funders Author Manuscripts

23. Anantharaman V, Iyer LM, Aravind L. Ter-dependent stress response systems: novel pathways related to metal sensing, production of a nucleoside-like metabolite, and DNA-processing. Mol Biosyst. 2012; 8:3142–65. [PubMed: 23044854]
24. Baker AE, O'Toole GA. Bacteria, Rev Your Engines: Stator Dynamics Regulate Flagellar Motility. J Bacteriol. 2017; 199:e00088–17. [PubMed: 28320878]
25. Blair DF, Berg HC. The MotA protein of *E. coli* is a proton-conducting component of the flagellar motor. Cell. 1990; 60:439–49. [PubMed: 2154333]
26. Hosking ER, Vogt C, Bakker EP, Manson MD. The *Escherichia coli* MotAB Proton Channel Unplugged. J Mol Biol. 2006; 364:921–937. [PubMed: 17052729]
27. González-García VA, et al. Conformational changes leading to T7 DNA delivery upon interaction with the bacterial receptor. J Biol Chem. 2015; 290:10038–44. [PubMed: 25697363]
28. Makarova KS, Wolf YI, van der Oost J, Koonin EV. Prokaryotic homologs of Argonaute proteins are predicted to function as key components of a novel system of defense against mobile genetic elements. Biol Direct. 2009; 4:29. [PubMed: 19706170]
29. Nimma S, Ve T, Williams SJ, Kobe B. Towards the structure of the TIR-domain signalosome. Curr Opin Struct Biol. 2017; 43:122–130. [PubMed: 28092811]
30. Cui H, Tsuda K, Parker JE. Effector-Triggered Immunity: From Pathogen Perception to Robust Defense. Annu Rev Plant Biol. 2015; 66:487–511. [PubMed: 25494461]
31. Burch-Smith TM, et al. A Novel Role for the TIR Domain in Association with Pathogen-Derived Elicitors. PLoS Biol. 2007; 5:e68. [PubMed: 17298188]
32. Caplan JL, Mamillapalli P, Burch-Smith TM, Czymmek K, Dinesh-Kumar SP. Chloroplastic protein NRIP1 mediates innate immune receptor recognition of a viral effector. Cell. 2008; 132:449–62. [PubMed: 18267075]
33. Murray NE. Type I restriction systems: sophisticated molecular machines (a legacy of Bertani and Weigle). Microbiol Mol Biol Rev. 2000; 64:412–34. [PubMed: 10839821]
34. Pancer Z, Cooper MD. The Evolution of Adaptive Immunity. Annu Rev Immunol. 2006; 24:497–518. [PubMed: 16551257]
35. Zhang Y, Xia R, Kuang H, Meyers BC. The Diversification of Plant *NBS-LRR* Defense Genes Directs the Evolution of MicroRNAs That Target Them. Mol Biol Evol. 2016; 33:2692–2705. [PubMed: 27512116]
36. Essuman K, et al. The SARM1 Toll/Interleukin-1 Receptor Domain Possesses Intrinsic NAD + Cleavage Activity that Promotes Pathological Axonal Degeneration. Neuron. 2017; 93:1334–1343.e5. [PubMed: 28334607]
37. Burroughs AM, Zhang D, Schäffer DE, Iyer LM, Aravind L. Comparative genomic analyses reveal a vast, novel network of nucleotide-centric systems in biological conflicts, immunity and signaling. Nucleic Acids Res. 2015; 43:10633–10654. [PubMed: 26590262]
38. Couillault C, et al. TLR-independent control of innate immunity in *Caenorhabditis elegans* by the TIR domain adaptor protein TIR-1, an ortholog of human SARM. Nat Immunol. 2004; 5:488–494. [PubMed: 15048112]
39. Fliegert R, Gasser A, Guse AH. Regulation of calcium signalling by adenine-based second messengers. Biochem Soc Trans. 2007; 35:109–14. [PubMed: 17233614]
40. Thiaville JJ, et al. Novel genomic island modifies DNA with 7-deazaguanine derivatives. Proc Natl Acad Sci U S A. 2016; 113:E1452–9. [PubMed: 26929322]
41. Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJE. The Phyre2 web portal for protein modeling, prediction and analysis. Nat Protoc. 2015; 10:845–858. [PubMed: 25950237]
42. Hirano T. SMC proteins and chromosome mechanics: from bacteria to humans. Philos Trans R Soc Lond B Biol Sci. 2005; 360:507–14. [PubMed: 15897176]
43. Danilova O, Reyes-Lamothe R, Pinskaya M, Sherratt D, Possoz C. MukB colocalizes with the oriC region and is required for organization of the two *Escherichia coli* chromosome arms into separate cell halves. Mol Microbiol. 2007; 65:1485–1492. [PubMed: 17824928]
44. Haering CH, Gruber S. SnapShot: SMC Protein Complexes Part I. Cell. 2016; 164:326–326.e1. [PubMed: 26771499]

45. Petrushenko ZM, She W, Rybenkov VV. A new family of bacterial condensins. Mol Microbiol. 2011; 81:881–896. [PubMed: 21752107]
46. Chen I, Christie PJ, Dubnau D. The ins and outs of DNA transfer in bacteria. Science. 2005; 310:1456–60. [PubMed: 16322448]
47. Krupovic M, Makarova KS, Forterre P, Prangishvili D, Koonin EV. Casposons: a new superfamily of self-synthesizing DNA transposons at the origin of prokaryotic CRISPR-Cas immunity. BMC Biol. 2014; 12:36. [PubMed: 24884953]
48. Markowitz VM, et al. IMG: the Integrated Microbial Genomes database and comparative analysis system. Nucleic Acids Res. 2012; 40:D115–22. [PubMed: 22194640]
49. Li L, Stoeckert CJ, Roos DS. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 2003; 13:2178–89. [PubMed: 12952885]
50. Enright AJ, Van Dongen S, Ouzounis CA. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res. 2002; 30:1575–84. [PubMed: 11917018]
51. van Dongen S, Abreu-Goodger C. Methods in molecular biology (Clifton, N.J.). 2012; 804:281–295.
52. Soding J, Biegert A, Lupas AN. The HHpred interactive server for protein homology detection and structure prediction. Nucleic Acids Res. 2005; 33:W244–W248. [PubMed: 15980461]
53. Alva V, et al. The MPI bioinformatics Toolkit as an integrative platform for advanced protein sequence and structure analysis. Nucleic Acids Res. 2016; 44:W410–W415. [PubMed: 27131380]
54. Altschul SF, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997; 25:3389–402. [PubMed: 9254694]
55. Marchler-Bauer A, et al. CDD: a Conserved Domain Database for the functional annotation of proteins. Nucleic Acids Res. 2011; 39:D225–D229. [PubMed: 21109532]
56. Camacho C, et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009; 10:421. [PubMed: 20003500]
57. Radeck J, et al. The *Bacillus* BioBrick Box: generation and evaluation of essential genetic building blocks for standardized work with *Bacillus subtilis*. J Biol Eng. 2013; 7:29. [PubMed: 24295448]
58. Wilson GA, Bott KF. Nutritional factors influencing the development of competence in the *Bacillus subtilis* transformation system. J Bacteriol. 1968; 95:1439–49. [PubMed: 4967198]
59. Kropinski, AM, Mazzocco, A, Waddell, TE, Lingohr, E, Johnson, RP. Bacteriophages: Methods and protocols. Clokie, MRJ, Kropinski, AM, editorsHumana Press; NY: 2009. 69–76.
60. Wommack, KE, Williamson, KE, Helton, RR, Bench, SR, Winget, DM. Bacteriophages: Methods and protocols. Clokie, MRJ, Kropinski, AM, editorsHumana Press; NY: 2009. 3–14.
61. Baym M, et al. Inexpensive Multiplexed Library Preparation for Megabase-Sized Genomes. PLoS One. 2015; 10:e0128036. [PubMed: 26000737]
62. Bankevich A, et al. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. J Comput Biol. 2012; 19:455–477. [PubMed: 22506599]
63. Fortier LC, Moineau S. Methods in molecular biology (Clifton, N.J.). 2009; 501:203–219.
64. Mazzocco A, Waddell TE, Lingohr E, Johnson RP. Methods in molecular biology (Clifton, N.J.). 2009; 501:81–85.
65. Sing WD, Klaenhammer TR. Characteristics of phage abortion conferred in lactococci by the conjugal plasmid pTR2030. J Gen Microbiol. 1990; 136:1807–1815.
66. Titok MA, et al. *Bacillus subtilis* soil isolates: plasmid replicon analysis and construction of a new theta-replicating vector. Plasmid. 2003; 49:53–62. [PubMed: 12584001]
67. Dar D, et al. Term-seq reveals abundant ribo-regulation of antibiotics resistance in bacteria. Science. 2016; 352:9822.

Europe PMC Funders Author Manuscripts

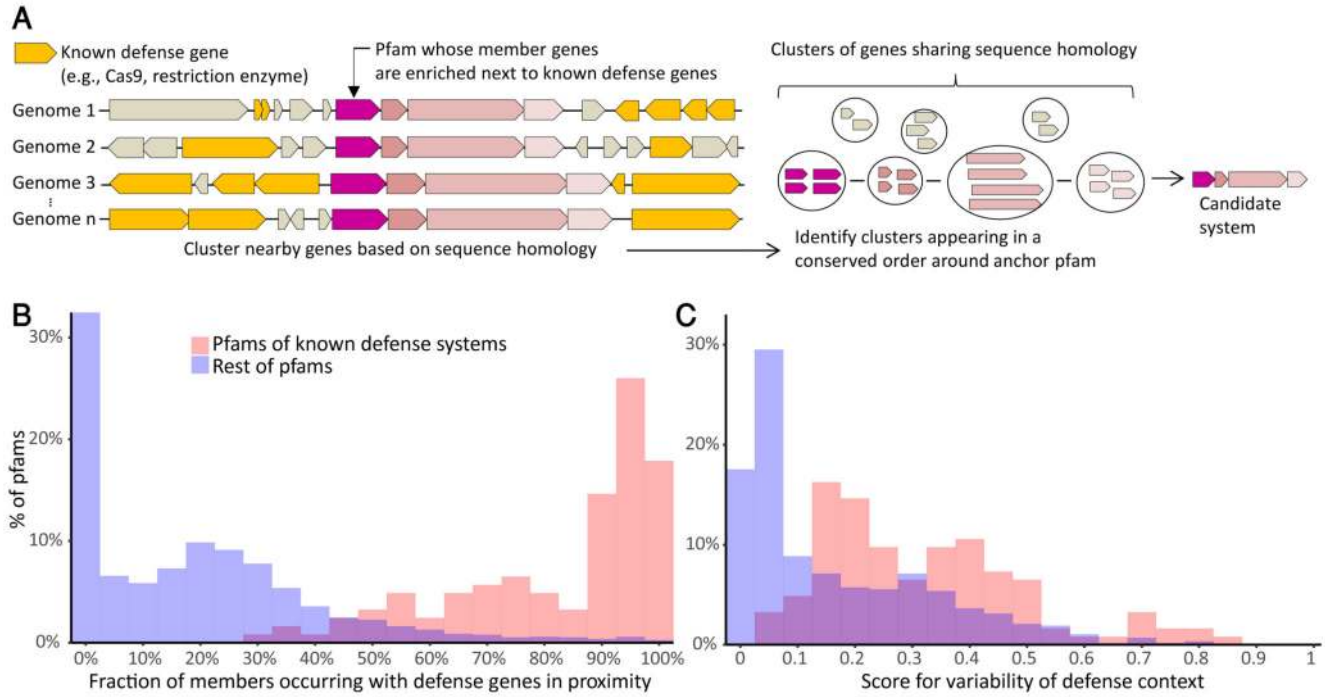

Figure 1. Discovery of new anti-phage defense systems in defense islands.

(A) Illustration of the computational analysis employed for each pfam found to be enriched in defense islands. Pfams that are enriched in the vicinity of known defense genes are identified, and their neighboring genes are clustered based on sequence homology to identify conserved cassettes that represent putative defense systems. (B) Tendency of protein families to occur next to defense genes. The genomic neighborhood for each member gene in each pfam is examined, and the fraction of member genes occurring in the vicinity (10 genes on each side) of one or more known defense genes is recorded. Pink, a set of 123 pfams known to participate in anti-phage defense ("positive set"); blue, the remaining 13,960 pfams analyzed in this study. (C) Neighborhood variability score for the analyzed pfams. Score represents the fraction of pfam members occurring in different defense neighborhoods out of total occurrences of pfam members (see Methods). Pink, the 123 positive pfams; blue, a set of 576 pfams that passed the 65% threshold for fraction of members occurring with defense genes in proximity.

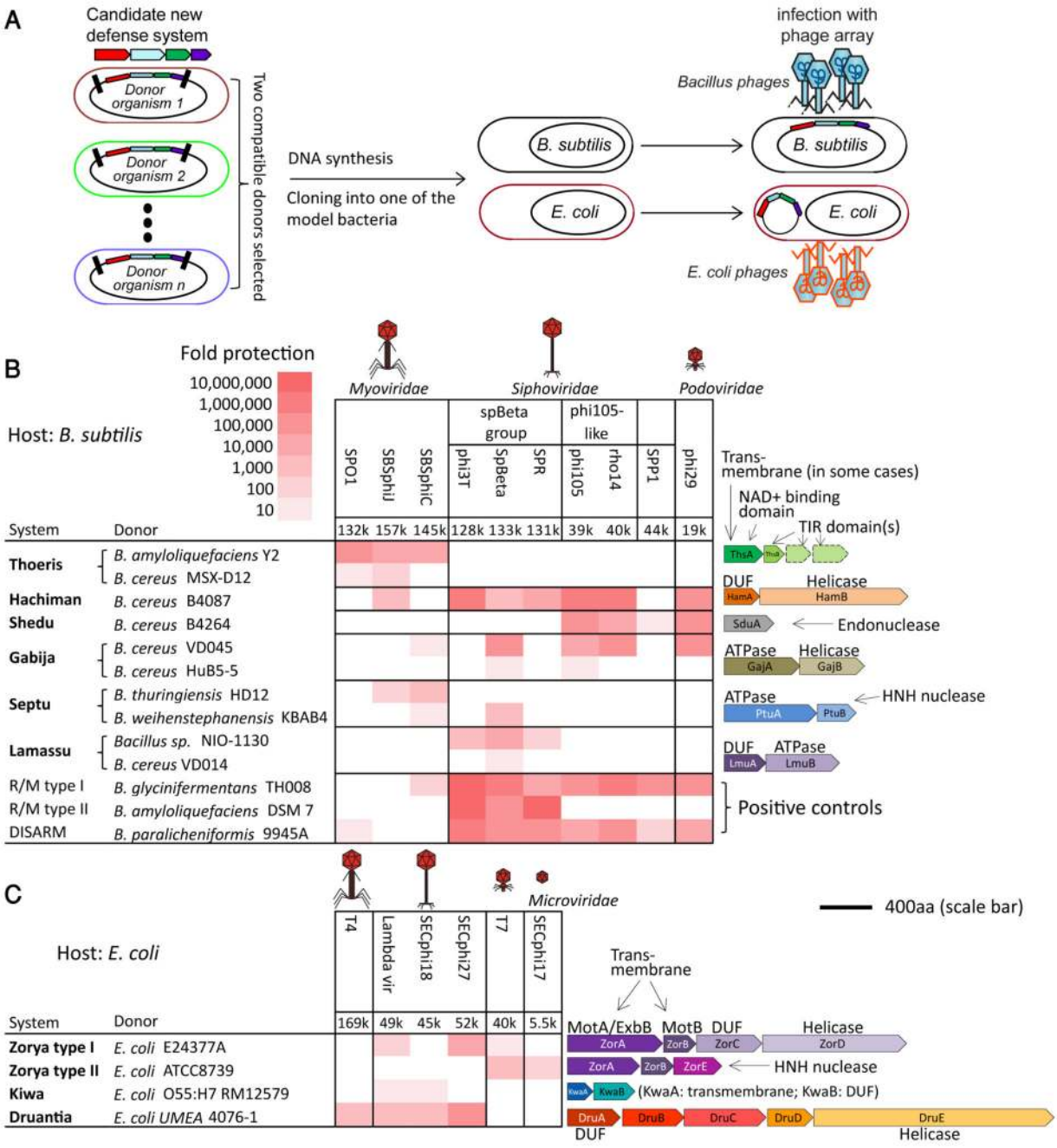

Figure 2. Experimentally verified defense systems.

(A) Flowchart of the experimental verification strategy. (B) Active defense systems cloned into *B. subtilis*. (C) Active defense systems cloned into *E. coli*. For B-C, fold protection was measured using serial dilution plaque assays, comparing the system-containing strain to a control strain that lacks the system and has an empty vector instead. Data represent average of 3 replicates, see Figures S2 and S3. Numbers below phage names represent phage genome size. On the right, gene organization of the defense systems, with identified domains

Europe PMC Funders Author Manuscripts

indicated (DUF, domain of unknown function). Gene sizes are drawn to scale; scale bar represents 400 amino acids.

Europe PMC Funders Author Manuscripts

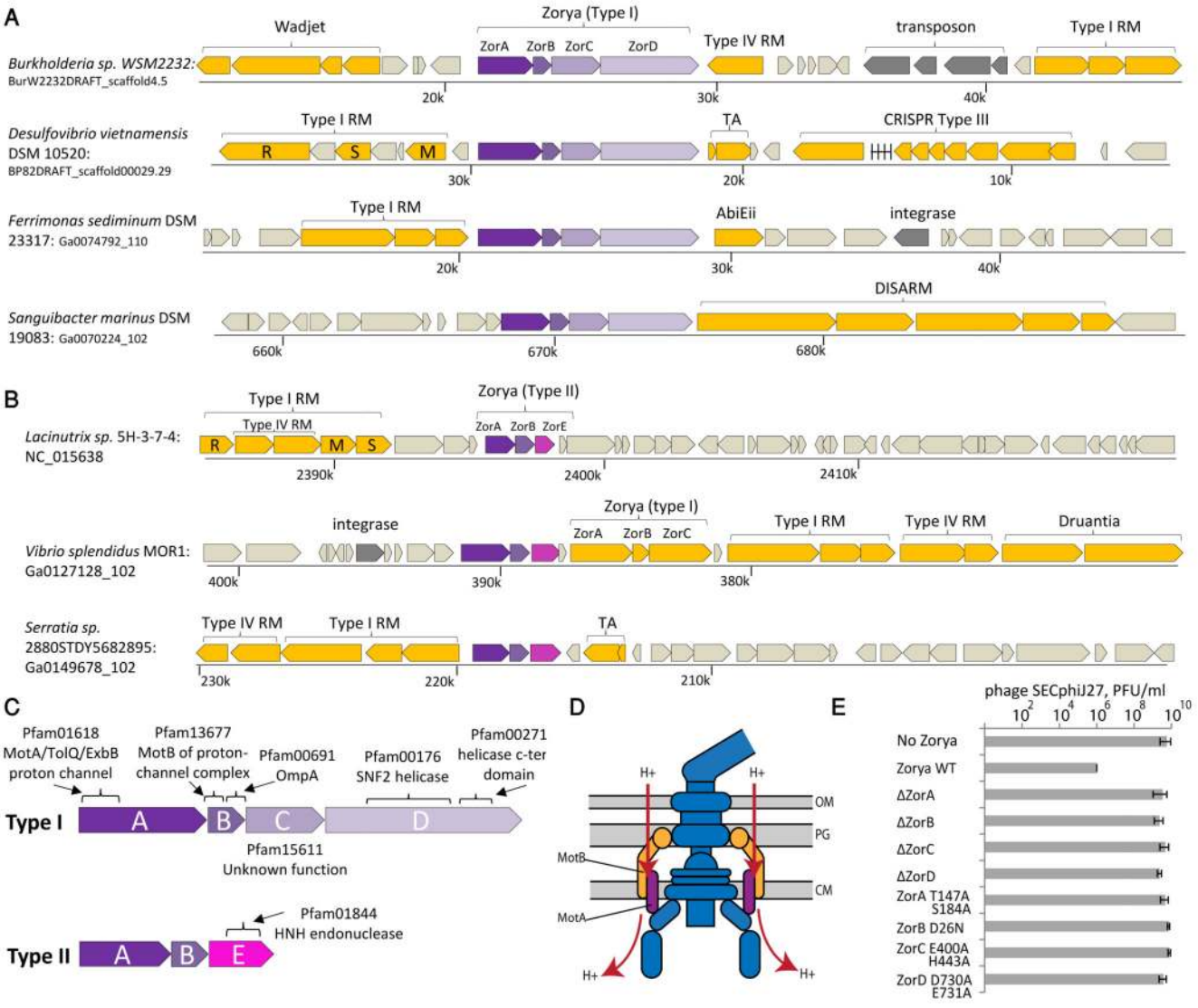

Figure 3. The Zorya system.

(A) Representative instances of the Type I Zorya system and their defense island context. Genes known to be involved in defense are orange. Mobilome genes are in dark grey. RM, restriction-modification; TA, toxin-antitoxin; Abi, abortive infection; Wadjet and Druantia are systems identified as defensive in this study (see below). (B) Representative instances of the Type II Zorya system. (C) Domain organization of the two types of Zorya. (D) Model of the flagellum base. The position of the MotAB complex is indicated. (E) Efficiency of plating (EOP) of phage SECphi27 infecting WT Type I Zorya, deletion strains, and strains with point mutations. Data represent PFU/ml, average of 3 replicates with error bars representing STD. ZorA:T147A/S184A and ZorB:D26N are predicted to abolish proton flux; ZorC:E400A/H443A are mutations in two conserved residues in pfam15611 ("EH domain") whose function is unknown (23); ZorD:D730A/E731A are mutations in the Walker B motif, predicted to abolish ATP hydrolysis.

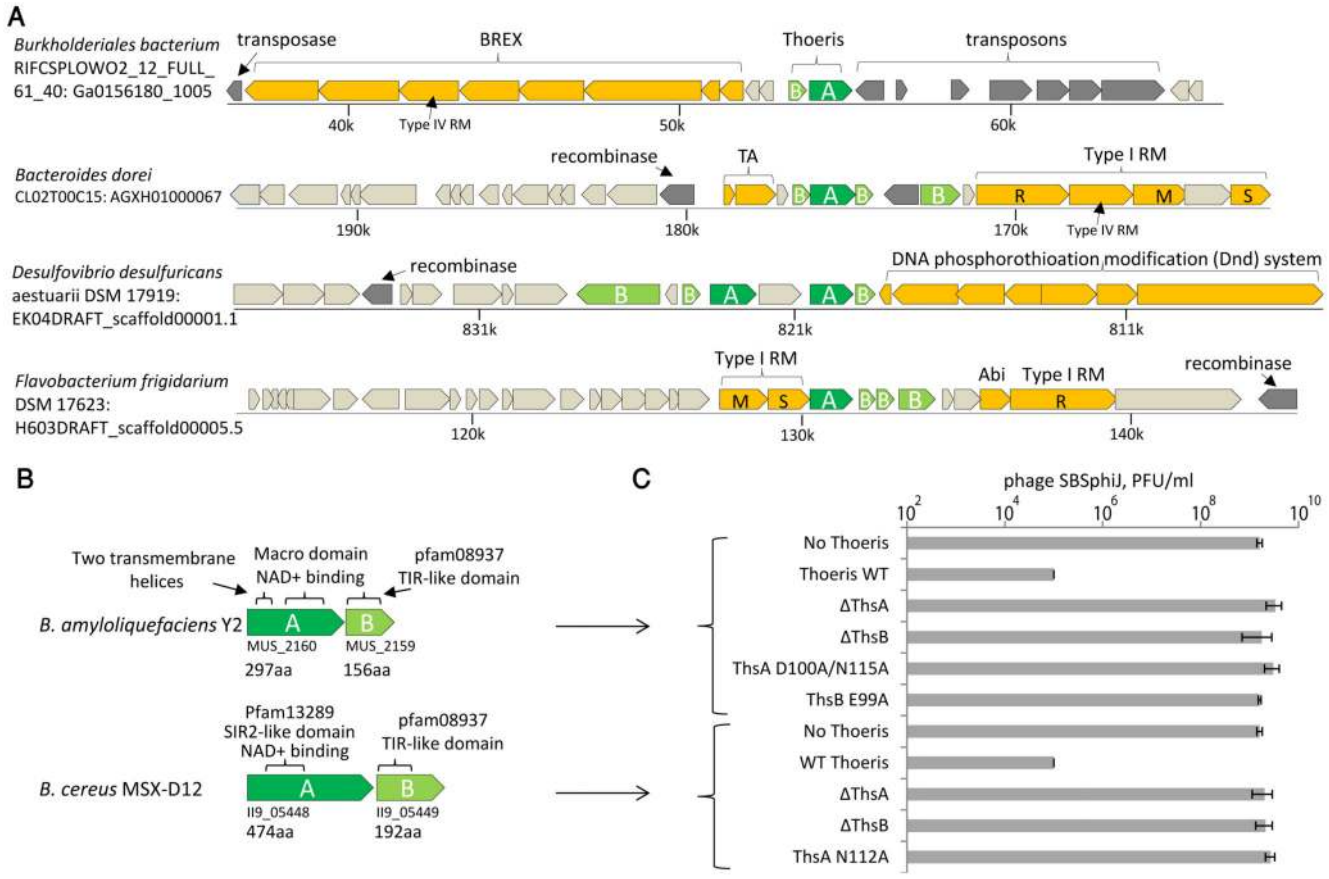

Figure 4. The Thoeris system.

(A) Representative instances of the Thoeris system and their defense island context. Thoeris genes *thsA* (containing NAD+ binding domain) and *thsB* (TIR domain) are marked dark and light green, respectively. Genes known to be involved in defense are orange. Mobilome genes are in dark grey. RM, restriction-modification; TA, toxin-antitoxin; Abi, abortive infection. (B) The two Thoeris systems shown in this study to protect against myophages. Locus tag accessions are indicated for the individual genes. (C) EOP of phage SBSphiJ infection with WT and mutated versions of the *B. amyloliquefaciens* Y2 Thoeris (top set) or *B. cereus* MSX-D12 Thoeris (bottom set) cloned into *B. subtilis* BEST7003. Average of 3 replicates, error bars represent STD.

Europe PMC Funders Author Manuscripts

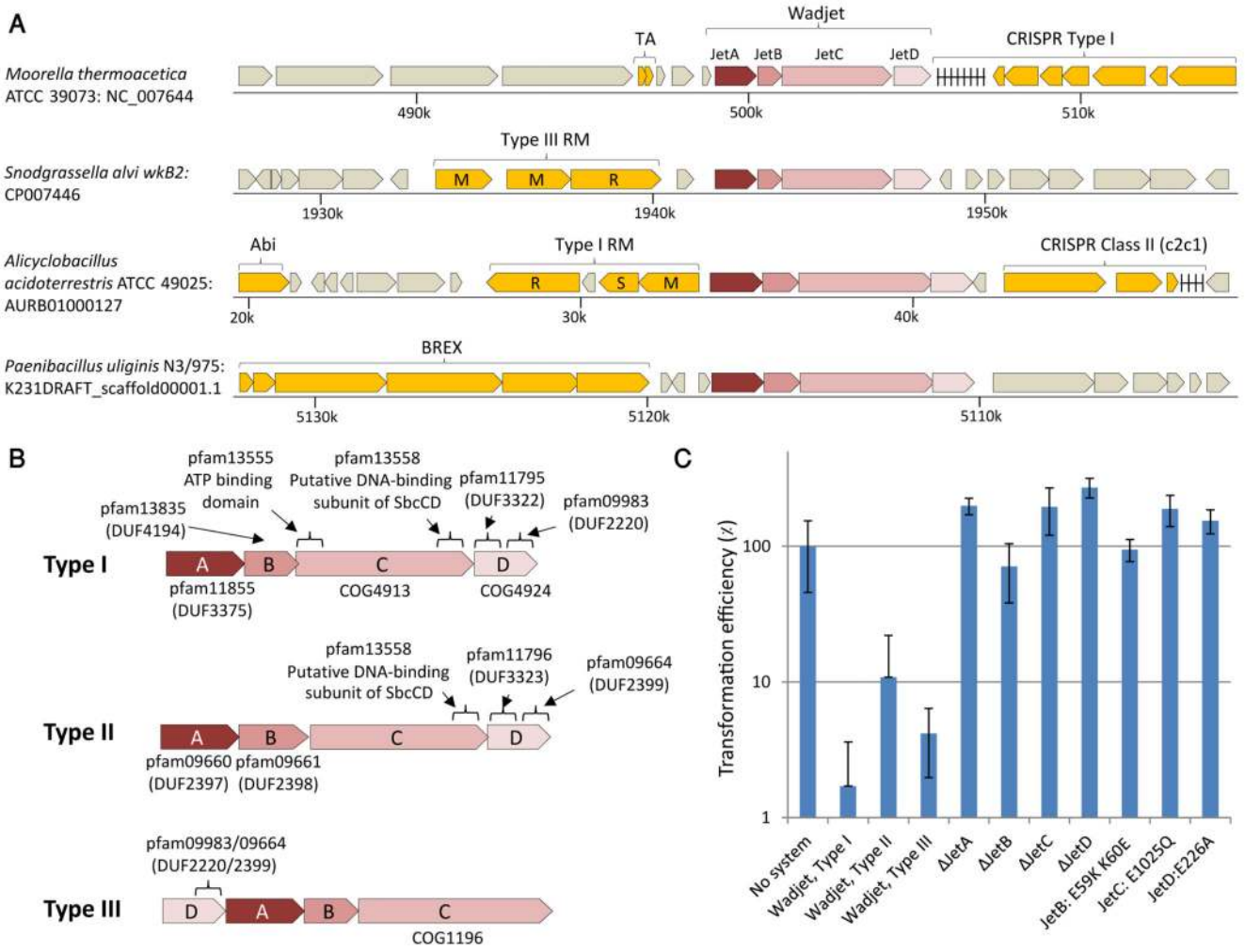

Figure 5. The Wadjet system provides protection against plasmid transformation in *B. subtilis*.

(A) Representative instances of the Wadjet system and their defense island context. Genes known to be involved in defense are orange. RM, restriction-modification; TA, toxin-antitoxin; Abi, abortive infection. (B) Domain organization of the three types of Wadjet. Pfam and COG domains were assigned according to the information in the IMG database (48). (C) Wadjet reduces plasmid transformation efficiency in *B. subtilis*. Instances of Wadjet systems were taken from *Bacillus cereus* Q1 (Type I), *Bacillus vireti* LMG 21834 (Type II) and *Bacillus thuringiensis* serovar finitimus YBT-020 (Type III) (Table S4) and cloned into *B. subtilis* BEST7003. Gene deletions and point mutations are of the *B. cereus* Q1 Type I Wadjet. Transformation efficiency of plasmid pHCMC05 into Wadjet-containing strains is presented as a percentage of the transformation efficiency to *B. subtilis* BEST7003 carrying an empty vector instead of the Wadjet system. Average of 3 replicates; error bars represent STD.

Europe PMC Funders Author Manuscripts

Table 1
Composition of defense systems reported in this study

System	Operon	Associated domains[a]	Domain annotations	# of instances detected within microbes	# (%) of genomes in which system is found	comments
Thoeris	ThsAB	pfam13289, pfam14519, pfam08937, pfam13676	SIR2, Macro domain, TIR domain	2,099	2,070 (4.0%)	Membrane associated (sometime)
Hachiman	HamAB	pfam08878, COG1204, pfam00270, pfam00271	Helicase	1,781	1,742 (3.4%)	
Shedu	SduA	pfam14082	Nuclease	1,246	1,191 (2.3%)	
Gabija	GajAB	pfam13175, COG3593, pfam00580, pfam13361, COG0210, pfam13245	ATPase, nuclease, helicase,	4,598	4,360 (8.5%)	
Septu	PtuAB	pfam13304, COG3950, pfam13395, pfam01844	ATPase, HNH nuclease	2,506	2,117 (4.1%)	
Lamassu	LmuAB	pfam14130, pfam02463	SMC ATPase N-terminal domain	697	682 (1.3%)	
Zorya (type I)	ZorABCD	pfam01618, pfam13677, pfam00691, COG1360, pfam15611, pfam00176, pfam00271, COG0553, pfam04471	MotA/ExbB, MotB, helicase, Mrr-like nuclease	1,173	1,055 (2.6%)	Membrane associated
Zorya (type II)	ZorABE	pfam01618, pfam13677, pfam00691, COG1360, COG3183, pfam01844	MotA/ExbB, MotB, HNH nuclease	656	655 (1.3%)	Membrane associated
Kiwa	KwaAB	pfam16162	No annotated domain	934	924 (1.8%)	Membrane associated
Druantia	DruABCDE (type I) DruMFGE (type II) DruHE (III)	pfam14236, pfam00270, pfam00271, pfam09369, COG1205, pfam00145, COG0270	Helicase, methylase	1,342	1,321 (2.6%)	
Wadjet	JetABCD	pfam11855, pfam09660, pfam13835, pfam09661, pfam13555, pfam13558, COG4913, COG1196, pfam11795, pfam09983, pfam11796, pfam09664, COG4924	MukBEF condensin, topoisomerase VI	3,173	2,880 (5.6%)	

[a]Pfam and COG domains were assigned according to the information in the IMG database (48) and supplemented using HHpred (52).