Defense Islands in Bacterial and Archaeal Genomes and Prediction of Novel Defense Systems

Kira S. Makarova, 1 Yuri I. Wolf, 1 Sagi Snir, 2 and Eugene V. Koonin 1*

National Center for Biotechnology Information, NLM, National Institutes of Health, Bethesda, Maryland 20894, 1 and Department of Evolutionary and Environmental Biology and The Institute of Evolution, University of Haifa Mount Carmel, Haifa 31905, Israel 2

Received 13 June 2011/Accepted 28 August 2011

The arms race between cellular life forms and viruses is a major driving force of evolution. A substantial fraction of bacterial and archaeal genomes is dedicated to antiviral defense. We analyzed the distribution of defense genes and typical mobilome components (such as viral and transposon genes) in bacterial and archaeal genomes and demonstrated statistically significant clustering of antiviral defense systems and mobile genes and elements in genomic islands. The defense islands are enriched in putative operons and contain numerous overrepresented gene families. A detailed sequence analysis of the proteins encoded by genes in these families shows that many of them are diverged variants of known defense system components, whereas others show features, such as characteristic operonic organization, that are suggestive of novel defense systems. Thus, genomic islands provide abundant material for the experimental study of bacterial and archaeal antiviral defense. Except for the CRISPR-Cas systems, different classes of defense systems, in particular toxin-antitoxin and restriction-modification systems, show nonrandom clustering in defense islands. It remains unclear to what extent these associations reflect functional cooperation between different defense systems and to what extent the islands are genomic “sinks” that accumulate diverse nonessential genes, particularly those acquired via horizontal gene transfer. The characteristics of defense islands resemble those of mobilome islands. Defense and mobilome genes are nonrandomly associated in islands, suggesting nonadaptive evolution of the islands via a preferential attachment-like mechanism underpinned by the addictive properties of defense systems such as toxin-antitoxins and an important role of horizontal mobility in the evolution of these islands.

Theoretical modeling of the evolution of replicator systems shows that virus-like genomic parasites inevitably emerge as soon as the distinction between the genotype and the phenotype is as established (19, 84). Indeed, viruses and virus-like selfish elements are virtually ubiquitous parasites of cellular organisms, with the only apparent exception of intracellular parasitic organisms that do not seem to harbor their own viruses. The arms race between parasites, especially genomic parasites such as viruses, and host organisms is a key driving force of the evolution of all life forms (14, 21, 32, 67, 79). An intrinsic part of this arms race is the evolution of multiple, diverse antiviral defense systems in all cellular life forms and the evolution of counterdefense systems in many viruses. The Red Queen evolutionary dynamics cause rapid evolution of both viruses and defense systems (9, 71). Textbook examples include the rapid antigenic drift of the major human pathogens influenza viruses and HIV, a phenomenon that is of central importance for the epidemiology of these viruses (30).

Metagenomics has produced startling results showing that bacterial viruses (bacteriophages) are by far the most abundant and genetically diverse biological entities on Earth, at least in marine habitats (18, 47, 82). Apparently, bacteria and archaea are subject to a constant barrage by diverse viruses, which inevitably triggers the evolution of multiple diverse defense systems (48, 79). The major mechanisms of defense system variability fueled by the virus-host arms race include rapid sequence evolution, extensive gene duplication (amplification), and horizontal gene transfer (HGT), which is often mediated by plasmids carrying the respective defense genes (14, 23, 27, 43, 89). The consequences are the characteristic patchy phyletic distributions of most of these systems (bacterial and archaeal strains that are otherwise closely related often differ in the content of defense systems) and the extreme divergence of the protein sequences of orthologous defense genes, which make the identification of defense systems by computational methods a nontrivial task (49, 53, 55, 70).

The antivirus defense systems function on one of the two general principles, (i) self-nonself discrimination, whereby a defense mechanism recognizes and destroys foreign (e.g., viral) genomes whereas the host genome is protected, and (ii) programmed cell suicide or dormancy induced by infection (48). The self-nonself discrimination principle is employed in particular by the restriction-modification (RM) systems, which are probably the best-characterized defense systems in prokaryotes, to a large extent because restriction endonucleases are essential experimental tools of molecular biology (46, 70, 92). Methylase subunits of RM systems methylate specific sites in the host DNA, whereas nonmethylated foreign DNA is
cleaved by the endonuclease subunits. Another type of defense machinery that is also based on self-nonself discrimination is represented by the CRISPR (clustered regularly interspaced short palindromic repeats)-Cas (CRISPR-associated genes) systems that are encoded in the genomes of the great majority of archaea and many bacteria (31, 38, 54, 88). Unlike RM systems, which generically distinguish between modified and unmodified recognition sites in DNA, the CRISPR-Cas systems function via a bona fide adaptive immunity mechanism that targets specific sequences in plasmid or viral genomes. Through a series of reactions catalyzed by Cas proteins and not yet fully characterized, the CRISPR loci incorporate fragments of viral or plasmid genomes that are then transcribed into guide RNAs. The guide RNAs are incorporated into the Cascade complex built of multiple Cas protein subunits and used to recognize and cleave the cognate alien DNA. Occasional incorporation of host DNA sequences into CRISPR repeat loci leading to autoimmunity has been reported (1, 78), and the mechanisms of self-nonself discrimination by the CRISPR-Cas systems remain unclear. The toxin-antitoxin (TA) systems that are extremely widespread in bacteria and archaea function through stress-induced cell suicide or dormancy (6, 25, 55, 90). The TA systems consist of two genes which encode, respectively, a toxin and an antitoxin. Under normal conditions, the toxin, which in most TA system is either a protein making holes in the cell membrane (holin) or an endonuclease cleaving ribosome-associated mRNA (RNA interferase), is maintained in an inactive state via interaction with the antitoxin gene product, either a small RNA that prevents toxin gene translation or a protein that forms an inactive complex with the toxin. Various stresses, including virus infection, inactivate the antitoxin and so unleash the toxin, which either kills the infected cell or induces dormancy, thus restricting the impact of the infection (11). The abortive infection (Abi) systems, also known as phage exclusion systems, represent another widespread group of defense mechanisms that abrogate virus infection at different stages and cause concomitant cell death (41). Thus, the Abi systems are effectively a variant of the TA systems (20). It should be noted that defense systems that employ the self-nonself discrimination mechanism have the potential to evolve into systems causing cell suicide or dormancy, as shown for RM systems (22, 33).

Genomes of free-living bacteria and archaea typically encode multiple defense systems of one or more classes, e.g., one or more CRISPR loci often coexist with multiple RM and/or TA systems within the same genome. Furthermore, although numerous and remarkably diverse defense systems have been identified, there is little doubt that bacterial and archaeal genomes harbor numerous uncharacterized variants of the RM, Abi, and TA systems and might encode novel classes of defense systems as well. In fact, the discovery and subsequent molecular characterization of CRISPR-Cas is an excellent exhibit for the likely existence of novel types of defense systems in bacteria and archaea. Although the CRISPRs have been known for about 2 decades and the arrays of cas genes and their predicted activities were described in 2002 (51), there was no inkling that this could be a defense system until 2005, when CRISPR spacers homologous to phage DNA were discovered (8, 60). Only then was the hypothesis proposed that CRISPR-Cas could be an immune system functioning on the RNA interference (RNAi) principle, and molecular evidence in support of this hypothesis was subsequently obtained (5, 53). A comprehensive genomic survey of TA systems has revealed numerous pairs of putative toxin and antitoxin genes, some of which, upon extensive analysis of the encoded protein sequences, show distant similarity to known toxins and antitoxins, whereas others share the generic features of TA systems but are likely to function via distinct mechanisms (55, 86). More recently, a comparative genomic study of the archaeal and bacterial homologs of the Argonaute proteins involved in eukaryotic RNAi has shown that the prokaryotic Argonaute-encoding genes, some of which encode characterized or predicted nucleases, are located in genomic neighborhoods enriched in other known and predicted genes involved in defense functions (56). We designated such neighborhoods defense islands (DIs), in a rough analogy to the pathogenicity and symbiosis islands identified in many bacteria (12, 36). Similar clustering of defense genes has been observed in other studies; in particular, diverse genomic islands carrying RM systems have been described (39). Furthermore, involvement of transposable elements in the horizontal transfer of defense systems has been demonstrated (23, 36, 83). An additional twist in the relationships between transposable elements and defense systems is added by the demonstration that mobile elements, in particular retroelements, can function as Abi systems and are also associated with CRISPR-Cas systems (14, 45, 53). More generally, RM and TA systems that are traditionally associated with defense functions themselves possess properties of selfish elements (23, 43, 89, 90). These elements are often present in genomes of plasmids and viruses and confer addictive properties on the carrier genomes; for example, the loss of a plasmid carrying a TA leads to the death of the respective bacterial cell due to the destructive effect of the stable toxin that is no longer kept inactive by the unstable antitoxin (88). Thus, these systems contribute to the maintenance of plasmids, proviruses, and genomic islands and also to competition between mobile elements so that the selfish behavior and defense functions of these elements are intertwined.

We sought to investigate DIs in detail with two major goals, (i) to assess the statistical significance of the clustering of defense genes in all major groups of bacteria and archaea, i.e., determine whether the existence of DIs could be demonstrated objectively, compare lineage-specific trends, and investigate the evolutionary dynamics of the DIs, and (ii) (provided the DIs prove to be objectively definable) to perform an exhaustive investigation of gene families overrepresented in DIs in an attempt to predict potential novel defense systems.

MATERIALS AND METHODS

Genomic data. For this analysis, 1,985 genome partitions (chromosomes, megaplasmids, and plasmids) representing 1,055 completely sequenced prokaryotic genomes (978 bacterial and 77 archaeal genomes) were downloaded from the NCBI FTP site (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/) in March 2010; this set was fixed for subsequent analysis (see Table S1 in the supplemental material). Protein sequences were assigned to NCBI clusters of orthologous groups (COGs) (85) using PSI-BLAST (2) searches with COG-derived position-specific scoring matrices; in 2,659,594 out of the 3,468,563 analyzed proteins (77%), at least one COG domain was identified. Proteins without COG assignments were run against the Pfam section of the NCBI conserved domain database (CDD) (57); additional domain assignments were obtained for 201,977 proteins. The available completely sequenced genomes sample the microbial diversity in an extremely nonuniform manner. For example, the set of 1,055 genomes con-
Clustering of additional protein families. All proteins currently not assigned to a family were extracted from the genomic set into a separate database. Proteins from positive islands that were not assigned to a protein family were used as queries in a BLAST (2) search against this database; hits with e values below 0.001 were extracted from the database and subjected to two-round clustering using the blastclust program (ftp://ftp.ncbi.nlm.nih.gov/blast/documents/blastclust.html). In the first round (score density threshold of 1.5 bits/position, coverage threshold of 0.8, one-way coverage), highly similar fragments were clustered together with the longer sequences. In the second round (score density threshold of 0.7 bit/position, coverage threshold of 0.7, two-way coverage), representative sequences of the first-round clusters formed families of homologous proteins.

Refinement of the positive and negative data sets. We started our analysis with an initial set of 136 COGs (see Table S3 in the supplemental material) from bacteria and archaea that are definitively or tentatively implicated in antivirus defense (positive set), such as TA, Abi, and RM system components, and 3,061 COGs, most of which are known to be involved in housekeeping functions or metabolic pathways and hence are unlikely to function in defense (see Table S4 in the supplemental material) (negative set).

In the course of the subsequent analyses, we included in the positive set three additional families that are likely to be involved in antivirus defense, removed one family that was prone to produce false positives due to the presence of a ubiquitous helix-turn-helix (HTH) domain, and moved six families of predominantly housekeeping methyltransferases to the negative list, bringing the number of positive COGs to 132 and the number of negative COGs to 3,067 (see Tables S5 and S6 in the supplemental material). As the first step in the DI analysis, we mapped the members of these COGs to chromosomes and plasmids of 383 representative genomes described above and collected five adjacent genes upstream and downstream of each gene from the positive set (Fig. 1). For all of the annotated gene families in these neighborhoods, we analyzed the statistics of occurrence in the vicinity of the positives. Families for which the Bonferroni-corrected P value was below 0.05 were considered overrepresented. We identified a total of 416 COGs that were overrepresented in the vicinity of known defense genes, of which 56 belonged to the original negative set. After the subtraction of these COGs, the final negative set was pruned to 3,011 COGs (see Table S7 in the supplemental material). As the second step in the DI analysis, we further defined a positive island as an island that contained at least one gene from the positive set. All genes in the positive islands that were not assigned to a COG or a Pfam family were collected and clustered into families (see above). We analyzed the family statistics of genes found in the positive islands in the same set of 383 representative genomes and identified 422 overrepresented...
families (Bonferroni-corrected $P$ value, < 0.05). This list contained an additional 128 families of known phage defense-related genes that were not part of the COG database (e.g., pfam01945, CcdB, a component of a TA system or pfam09504 or BpRl, a restriction endonuclease), bringing the total number of positive families in the final positive set to 260 (see Table S8 in the supplemental material).

Statistical analysis of scaling of gene classes. Statistical analysis was performed using the R software package (http://www.r-project.org). In particular, scaling of the number of genes in phage defense and mobilome categories relative to the genome size was performed using the power law model for the expectation and negative binomial distribution for the error function (7). Models with estimated and fixed exponents were compared using the Akaike information criterion corrected for small samples (10). Linear models with nonnumerical factors were analyzed using the Akaike information criterion and analysis of variance; stepwise reduction of fully defined models (those including all factors and all combinations of factors) was performed until further elimination of factors led to a statistically significant loss of the model’s explanatory power.

Analysis of proximity between sets of genes in prokaryotic genomes. To investigate whether there is a tendency for two gene sets to stay close to each other on microbial chromosomes, the following approach was employed (Fig. 2A). Genes from both sets were identified in each genome partition (assumed circular); for each gene from the first set, the closest neighbor from the second set was identified and vice versa. Distances to the closest neighbor from the second set were averaged across all genes in both sets; the mean was used as the measure of the proximity of the two sets in this genomic partition. Then, a random arrangement of both sets on the chromosome was generated such that the distribution of distances between the genes in each set was retained. To this end, for each set, the partition was split into fragments representing stretches of genes in each set. The reconstructed random arrangement of both sets was identical to a strict subset of the random arrangement of both sets.

Now consider a pair of genomes, $A$ and $B$, where each genome is characterized by a set of islands $\{A_i\}$ and $\{B_j\}$. Comparing each island from genome $A$ to all of the islands in genome $B$, we find the most similar counterpart in that genome as follows: $M_{ab} = \max(S_{ab})$. The median value of the combined set $\{M_{ab}, M_{ba}\}$ gives a symmetrical measure of the evolutionary conservation of islands in genomes $A$ and $B$.

We chose 30 pairs of closely related completely sequenced genomes (see Table S9 in the supplemental material) from the ATGC database (68) using the following informal criteria: (i) “fair” coverage of prokaryotic clades (i.e., we avoided including many pairs from closely related ATGCs, such as Escherichia and Salmonella) and (ii) appropriate distance between genomes, allowing for some diversity in gene order and content, but not to the point where long syntenic blocks are mostly nonexistent (68). Within this set of closely related genomes, the evolutionary conservation of DIs and mobilome islands with a minimum length of 4 genes was measured. To obtain the proper scale of the observed similarity of each of these pairs of genomes, we generated 100 sets of randomly selected nonnegative families (mock positives) with the same number of representatives as the defense (mobilome) genes, identified pseudoislands based on these families, and measured the evolutionary conservation of these sets. For each pair of genomes, the conservation of defense (mobilome) islands was characterized as the percentile within the corresponding distribution of pseudoislands based on randomly chosen families.

Distant protein sequence similarity analysis. Protein sequence database searches were performed using PSI-BLAST (2) with an inclusion threshold $e$ value of 0.01 and no composition-based statistical correction. In addition, distant similarity detection approaches were applied, namely, the CDD search (57) and the HHpred search, which is based on the comparison of protein family profiles using the hidden Markov model technique (76). Alignments of multiple protein sequences were constructed by using the MUSCLE program (17), followed, when necessary, by a minimal manual correction on the basis of local alignments obtained using PSI-BLAST and HHpred programs. Protein secondary structure was predicted using the Jpred program (15), and these results were used to improve the alignment between families within a superfamily. Structural comparisons were performed using the DALI server (29). Membrane topology was predicted using the TMHMM program (77).

FIG. 2. Analysis of proximity between sets of genes in prokaryotic genomes.
RESULTS AND DISCUSSION

Clustering of antivirus defense genes. Statistical analysis of the clustering of defense genes in bacterial and archaeal genomes was performed with 136 COGs in the initial positive set (see Table S3 in the supplemental material) and 3,061 COGs in the initial negative set (see Table S4 in the supplemental material). From the representative set of 774 bacterial and archaeal genomes (an amended set of genomes from our previous study [55]), we selected those genomes that contained at least 10 positive genes and analyzed the distribution of genomic distances between the nearest positives in each of these genomes (the distance was calculated as the number of nonpositive protein-coding genes separating the neighboring positive genes). If the positives are distributed randomly, these distances are expected to approximately follow an exponential distribution with a mean of \((N - n)/n\), where \(N\) is the total number of genes in the genome and \(n\) is the number of defense genes (positives). By binning the observed distribution of distances (we used bins of approximately equal sizes) and computing the expected number of distances in each bin from the exponential approximation, one can compare the observed and expected distributions by using the chi-squared test.

In the analyzed set of 774 genomes, 114 genomes showed significant deviation from the expected distribution with a \(\chi^2\) value of \(<0.05\). In all of these 114 genomes, the pattern of deviation from the expectation was the same, with the shortest-distance and longest-distance bins containing an excess of data points. This result demonstrates nonrandom clustering of known and predicted antivirus defense genes.

As a control, the following experiment was performed. From the set of 3,061 negatives, 136 random mock-positive COGs were chosen and the above-described procedure was applied to this data set. Repeating this procedure 100 times, we obtained the distribution of the number of genomes that deviated from the expected exponential distribution (Fig. 3). The highest observed number of deviating genomes was 15, and in 95% of the replicates, the number of deviations was \(\leq5\). Thus, the fraction of significantly deviating genomes obtained for the defense genes was highly unexpected, indicating that in a considerable fraction of the bacterial and archaeal genomes, the defense genes are nonrandomly clustered. These results show that the DIs represent a bona fide genomic feature. We further explored in detail the statistical and biological properties of the DIs.

Delineation of DIs. We identified positive islands in completely sequenced genomes of Bacteria and Archaea using the final sets of positive and negative families (see Materials and Methods). The 33,582 identified islands were spread across 1,026 genomes and included 190,444 genes, of which 53,419 belonged to known defense systems. Genes previously unassigned to families were clustered to form new families; the statistics of the distribution of these families were analyzed in the set of 383 representative genomes. Altogether, we identified 473 overrepresented families (see Table S10 in the sup-
(supplemental material) of which 211 belonged to the positive set; the remaining 262 families were the subject of the subsequent detailed analysis described below.

**Mobilome islands.** At least some defense systems show extreme mobility across diverse bacterial and archaeal genomes; i.e., these genes are subject to frequent HGT (53, 55, 66). To compare the distribution of the defense genes with the typical mobilome genes (various transposon and prophage components), we identified islands bounded by genes from the same negative set and containing at least one gene from the mobilome positive set (289 families; see Table S11 in the supplemental material). This analysis yielded 47,274 mobilome islands (a slightly larger set compared to the DIs) spread across 1,044 genomes and containing 275,743 genes of which 91,373 belonged to the mobilome.

**Length and density of DIs and mobilome islands.** We statistically investigated the properties of DIs and mobilome islands within the set of 383 representative genomes. The mean lengths of DIs and mobilome islands were, respectively, 5.7 and 6.1 genes; both values are substantially greater than the mean length of a generic island (a genome fragment consisting of nonnegative genes and bounded by negative genes, i.e., without the requirement for the presence of a positive gene), which was 2.1 genes. The fact that DIs and mobilome islands do not represent an unbiased sample of generic islands is particularly obvious from the island length distributions (Fig. 4). Longer islands were significantly overrepresented among both the DIs and mobilome islands compared to the generic islands (Kolmogorov-Smirnov test *P* values, <10^{-10}). The length distributions of the DIs and mobilome islands were highly similar to each other but strikingly different from the generic island distribution. Notably, most of the longest identified islands contained both defense and mobilome genes.

Functionally linked genes in prokaryotes tend to form operons. Operons for different functional systems differ substantially in length, from typically two-component TA systems (55) to CRISPR-Cas system operons that often consist of 4 to 8 genes (54), whereas components of other functional systems are encoded in one-gene operons. To assess the degree of gene “operonization” in the DIs and mobilome islands, we counted the positive directons, which are stretches of adjacent codirectional genes belonging to the positive set (74, 93). The mean positive directon densities of DIs and mobilome islands were 1.193 and 1.401 per island, respectively. For comparison, we computed the positive directon density of random mock-positive sets drawn from nonnegative families such that the total number of genes in these families was close to that in the defense and mobilome sets. Repeating this procedure 1,000 times (the mean size of the random set was 28,100 ± 500 genes, close to the target value of 27,700 genes, the geometric mean of the defense and mobilome set sizes in representative genomes) yielded a mean positive directon density of 1.081 ± 0.014, which is significantly lower than the direction density in the DIs and mobilome islands (*P* values under a normal approximation, <10^{-10}). Thus, the DIs and mobilome islands are significantly enriched in putative operons encoding components of defense systems and typical mobilome components, respectively.

**Factors defining the abundance of phage defense and mobilome genes in bacterial and archaeal genomes.** We investigated the abundance of the defense and mobilome genes in the 383 representative bacterial and archaeal genomes (see Table

---

**FIG. 4.** Distribution of island lengths. Defense, DIs; Mobilome, mobilome islands; All, all islands of nonnegative genes.
S2 in the supplemental material). The abundance of both the defense and mobilome genes strongly correlated with the total number of protein-coding genes in the respective genomes (Spearman rank correlation coefficients were 0.58 and 0.68 for defense and mobilome genes, respectively; \(P\) values, \(< 1 \times 10^{-20}\)). Both categories follow the power law scaling that is typical of functional classes of genes in prokaryotes (61, 91) (Fig. 5A): the number of defense genes grows linearly with the genome size (the exponent is statistically indistinguishable from 1; \(P\) value, 0.5), whereas the number of mobilome genes shows significant superlinear scaling (exponent of 1.44; \(P\) value, \(4 \times 10^{-7}\)). This difference might indicate that the defense
genes have a substantial impact on the fitness of bacteria and archaea, and therefore their propagation in the host genome is more tightly constrained by selection, compared to the selfish propagation of the mobilome genes. However, the overall linear scaling of defense genes with the genome size is a mixture of significantly different trends: the scaling exponents for TA, Abi, RM, and CRISPR-Cas systems were 1.30, 1, 0.77, and 0, respectively (Fig. 5B). Connecting these distinct exponents to the biological features of the respective systems remains an interesting challenge for further study.

Taking into account the genome size explains ~44% of the original variance in the abundance of the defense genes (on a log scale). To investigate other factors that might affect the abundance of defense genes, we computed the expected number of defense genes in each genome using the scaling parameters estimated as described above and calculated the log ratio of the observed and expected values (using a value of 0.5 where the observed number of phage defense genes was 0). We then estimated the effects of taxonomic affiliation (Archaea versus Bacteria), growth temperature (mesophiles versus thermophiles), and ecological niche (free-living, host-associated, and endosymbiont). All three factors affected the abundance of defense genes significantly ($P < 0.05$) and independently of each other. Specifically, archaea tend to possess more defense genes than bacteria by a factor of ~1.5; compared to mesophiles, thermophiles possess ~1.4 times more defense genes; host-associated bacteria have ~1.3 times fewer defense genes than free-living prokaryotes, whereas for endosymbionts, the reduction factor is ~3.8. Overall, taking into account these three factors removed a further 10% of the original variance in the abundance of phage defense genes. Restricting the analysis to 257 free-living organisms has no noticeable effect on the contributions of taxonomic affiliation and growth temperature; both factors remain significant and independent.

A similar analysis of the abundance of mobilome genes produced notably different results. The dependence on genome size was somewhat more pronounced for mobilome genes (51% of the original variance is removed by power law scaling), whereas the other factors combined additionally explained only about 3% of the variance. The ecological niche does not significantly affect the abundance of mobilome genes; bacteria possess ~1.6 times more mobilome genes than archaea, and thermophiles have fewer mobilome genes than mesophiles by a factor of ~1.3 (in the latter case, the direction of the effect is the opposite of that with the defense genes). The effects of taxonomy and temperature preference are significant ($P < 0.05$) and independent of each other.

**Association between different defense systems and mobilome genes.** We further analyzed the tendency of genes belonging to different defense systems and the mobilome to cluster on microbial chromosomes by comparing the observed distributions of the genes with simulated random arrangements (see Materials and Methods). Genes of three of the four classes of defense systems (TA, Abi, and RM) showed significant nonrandom colocalization between each other and with mobilome genes (Fig. 2B). In contrast, for cas genes, no association with other defense or mobilome genes was demonstrated, probably because of the tendency of the CRISPR-Cas operons to occupy only a few loci in the genome, depriving the approach of statistical power.

The association of TA, Abi, and RM systems with each other and with mobilome components could be partly explained by the addictive (selfish) properties of these defense systems. In a sense, the DIs enriched in these systems probably should be considered “addiction islands” that are retained in bacterial genomes in part owing to the deleterious effect of the loss of the addictive elements. Moreover, some of the DIs, in particular, those that include integrase genes, are likely to represent active or inactivated mobile elements such as integrons and superintegrons; the role of TA in integron stabilization is a recognized phenomenon (12). We examined the potential association between DIs and integrons quantitatively. There are 1,095 predicted integrons (identified by the presence of the XerD/IntI/COG4974 integrase gene, the marker of integrons [12]) and 14,288 DIs in the 383 representative genomes. Of these DIs, only 196 (~1.4%) are associated with integrases. Thus, although there is a highly significant association ($P$ value, $\sim 10^{-60}$) between XerD integrases (and hence putative integrons) and defense genes, this association is far too weak to explain any trends in DI distribution.

**Evolutionary conservation of genomic islands.** We compared the evolutionary conservation of DIs and mobilome islands in 30 pairs of closely related organisms (see Table S9 in the supplemental material) with the distribution of pseudoislands defined by randomly selected sets of families. On average, DIs ranked in the 25th percentile and mobilome islands ranked in the 20th percentile of the distribution (see Tables S12 and S13 in the supplemental material). Both sets of islands are significantly less conserved than randomly selected islands ($P$ values, according to the Stouffer Z score omnibus test [80], $8 \times 10^{-6}$ and $6 \times 10^{-10}$, respectively).

**Gene families overrepresented in DIs and putative novel defense systems.** We performed a detailed, case-by-case analysis of the 262 families of protein-coding genes that are statistically significantly overrepresented in DIs (see Materials and Methods; see also Table S10 in the supplemental material). Of these families, 101 were found to belong to various mobile elements (i.e., are part of the mobilome). This observation provides additional support for the association between defense systems and the mobilome in the DI. Of the remaining 161 families, 57 are predicted to be involved in defense, 6 families possess predicted functions unrelated to defense (although 5 of these are related to arsenic resistance, falling into a general stress resistance category), and for the remaining 98 families, no clear prediction could be made. Among these 57 families of newly predicted defense genes, 14 belong to diverged subfamilies of known components of TAS, RM, Abi, or CRISPR-Cas; 1 family, pfam11194, is a result of an annotation error (CRISPR repeat region translated); and the rest might represent novel defense systems. We selected several such examples to discuss in greater detail.

One predicted defense system centers around a family of proteins containing the PglZ domain, which belongs to the alkaline phosphatase superfamily (pfam08665) (94). Although this system (here, the Pgl system) is poorly characterized and hence eluded our original positive set, in retrospect, it became clear that the PglZ (phage growth limitation) protein is a component of the PglWXYZ system that confers protection against the temperate bacteriophage phiC31 in *Streptomyces coelicolor* A3 (2) (13, 81). This system also includes the P-loop...
ATPase domain-containing protein PglY, the methylase PglW, and the serine-threonine kinase PglX (the latter two proteins are encoded in a different locus). The Pgl phenotype is characterized by the ability of PgL hosts to support a phage burst upon initial infection, but subsequent phage growth cycles are severely restricted (81). The molecular mechanism of the Pgl system is not known, but it has been hypothesized that it modifies (methylates) the DNA of the phage progeny rather than the host DNA. The modification is not lethal to the phage, and a few infected cells might undergo lysis but reinfection of the remaining cells in the same Streptomyces colony is thought to activate the system inhibiting phage growth (13, 81). Thus, this unusual defense system could function through a mechanism that is the reverse of the RM mechanism in that the Pgl system modifies the virus DNA to tag it for destruction rather than modifying the host DNA, even at the expense of a few initially infected and lysed cells. Thus, the Pgl system seems to combine the self-nonself discrimination and virus-induced cell death modes of antivirus defense in a novel defense strategy.

The Pgl system was described in an early study (13) but seems to have escaped the attention of subsequent antivirus defense research, so we rediscovered it in the course of the present analysis of DIs. Altogether, we identified 81 pglZ genes in 77 genomes from a variety of bacteria and several mesophiles and archaea. The neighborhood analysis of the PglZ family revealed a complicated patchwork of Pgl systems with variable compositions. The PglW and PglX proteins are among the 13 families that are encoded in these neighborhoods in six or more genomes (Table 1); the neighborhoods also contain numerous less common genes, including that for the PglY ATPase (see Table S14 in the supplemental material). Thus, it seems that the previously described four-gene pgwXYZ system of S. coelicolor is only the tip of the proverbial iceberg. Some of the operons containing the pglZ gene consist of up to 8 genes (Fig. 6). The remarkable complexity of this system could reflect an elaborate molecular mechanism of self-nonself discrimination and fine-tuned regulation.

Our approach revealed several families represented in the vicinity of pglZ genes that could be shared with regular RM systems. One of these families is COG1479 (or DUF262). A member of this family has been previously identified in the same locus as the type I RM system in Campylobacter jejuni (59). We identified the same association in numerous other bacterial genomes (Fig. 7A). However, in many cases, PglZ-associated genes were detected outside any operonic context. In particular, the COG1479 family is significantly expanded in Helicobacter (up to 8 copies per genome) but none of these genes is associated with RM-related genes. Thus, proteins of this family might be able to function independently or in trans. A sequence database search using the HHpred program (76) shows that the core of COG1479-like proteins is similar to the ParB-like nuclease fold (see Table S15 in the supplemental material). The most conserved sequence motif of the COG1479 family, DGQHR, is similar to that of the so-called DGQHR domain (TIGR03187 and TIGR03233) that also belongs to the ParB-like nuclease fold. The DGQHR domain-containing proteins have been shown to participate in sulfur modification of DNA (95). Proteins containing a COG1479 domain show several distinct domain arrangements (Fig. 7B).

Most often, the DGQHR domain is associated with a C-terminal HKH-type nuclease domain (DUF1524; see Table S15 in the supplemental material; Fig. 7B). All of these findings collectively suggest the involvement of this protein family in a variety of defense and perhaps other functional systems with a wide spectrum of potential molecular mechanisms.

Some families found in DIs are significantly overrepresented only in distinct clades of bacteria and archaea. We were par-
ticularly interested in those genes that are overrepresented in archaeal DIs because archaeal antivirus defense and TA systems remain poorly characterized. One such example includes COG3372 proteins, which contain a C-terminal PD-(D/E)XK nuclease domain and are often associated with DEXH helicases of COG1061 proteins (40). These two genes are found in most major archaeal lineages (Thermococci, Methanosarcinales, Halobacteriales, Archaeoglobales, Thermoplasmatales, Desulfurococcales, Thermoproteales, Korarchaeota) and a few bacteria from diverse lineages. These proteins have been implicated in nucleotide excision repair (NER) and have been recently characterized biochemically, supporting the predicted nuclease and helicase activities, although no evidence of involvement in NER in vivo has been demonstrated (73). The abundance of this two-gene array in archaeal DIs suggests that the proteins encoded by these genes are components of a novel defense system.

In the course of the present analysis, we detected links between defense systems and genes that are generally considered housekeeping genes, suggesting that some of these housekeeping genes could perform additional functions in defense. An example is the family of HepA/RapA-like helicases (COG0553) that are tightly associated with RNA polymerase and participate in polymerase recycling during transcription (58, 65, 75). We found that genes encoding these helicases are also strongly associated with DIs and, moreover, are often located in the RM operons and fused to RM domains, suggesting direct involvement in defense (see Table S16 in the supplemental material). Other examples in this category include helicases of COG1205 and COG1112, RecD family ATP-dependent exo-DNase (COG0507), RecT family single-strand annealing protein (pfam03837) (34), and DNA mismatch endonuclease Vsr (COG3727). All of these proteins are involved in various DNA repair pathways, and it cannot be ruled out that at least some of them are coregulated with defense system components targeting DNA, ensuring timely repair of potential collateral damage to host DNA.

Because toxins and antitoxins typically are small, fast-evolving proteins, many families initially identified as new could be assigned to known superfamilies of toxins and antitoxins when more sensitive methods of sequence analysis were applied (see Table S10 in the supplemental material). However, one pair of genes, DUF3532 and a family that is typified by MAE_01690 (xls000012) from Microcystis aeruginosa (Fig. 8A), probably represents a novel type II TA system. This prediction is based on the same criteria that we applied in the previous comparative genomic analysis of toxins-antitoxins (52, 55): a putative TA system is defined as an operon that encodes two small
proteins, shows a patchy distribution among bacteria and/or archaea, and is represented in DIs/mobilome islands (e.g., GI no. 164514538; see reference 44) and/or on plasmids (e.g., p49879_1p19 from a plasmid of *Leptospirillum ferrooxidans*).

Additional support for this prediction comes from the results of sequence analysis. For the DUF3532 family, several structures have been solved and analysis of these structures using DALI revealed another structurally similar family, DUF2442, which, in addition to the DUF3532-related domain, contains a HTH domain of the Xre family that is found in many antitoxins (55). According to SCOP, the DUF3532 family represents a novel fold (NE0471 N-terminal domain-like). For the MAE_01690 family, no structures are available but analysis of a multiple-sequence alignment reveals a pattern of conserved amino acid residues suggestive of a metal-dependent enzyme, possibly a nuclease (Fig. 8B). Taken together, these observations imply that MAE_01690 could be the toxin and the DUF3532 family could be the antitoxin of a novel TA system.

Another family, DUF1814, is strongly associated with DIs and appears to form a two-component system with COG5340, a predicted transcriptional regulator. The majority of the DUF3532 family are annotated as hypothetical, but a few are described as Abi proteins. Indeed, an HHpred search initiated with any of the DUF1814 proteins identifies significant similarity to COG5340 (DUF2204), which includes one of the experimentally characterized Abi proteins; this family was present in our positive set along with two other related but poorly characterized families, COG4849 and COG4914. This experimentally characterized protein belongs to the AbiG family, and the gene that encodes it was first identified on lactococcal plasmid pCL750, which is involved in the inhibition of early and/or late transcription of several phages (69). Another
protein in this family, AbiE, is encoded by lactococcal plasmid pNP40 (24). Both systems appear to act at the stage of phage DNA replication, but their molecular mechanisms remain unknown (14). Sequence analysis using HHpred revealed significant similarity of DUF1814 proteins with nucleotidyltransferases, including the signature DxD motif (e.g., for query protein Ksed_22780 from Kytococcus sedentarius, HHpred found similarity to a putative nucleotidyltransferase with Protein Data Bank [PDB] code 2FCL with a probability of 96.26).

The DUF1814 family, along with the related COG5340 family, is abundant in bacteria and is also present in several archaea. Involvement in TA systems has been previously proposed for FIG. 8. A predicted novel TA system. (A) Two distinct operons containing genes coding for members of the protein family with an HxH motif. The structures of representatives of the DUF3532 and DUF2442 families are shown. The PDB codes are indicated. The ribbon diagrams were generated using the Jmol server (http://www.jmol.org/). For the DUF2442 structure, two folds of the two domains are denoted according to the fold assignment in the SCOP database (3, 64). (B) Multiple-sequence alignment of the HxH motif-containing family. Secondary structure prediction is shown beneath the alignment as follows: H, α-helix; E, extended conformation (β-strand). The sequences are denoted by their GI numbers and species names. The conserved amino acids are in bold. The coloring is based on the consensus shown at the bottom of the alignment; h, hydrophobic residues (WFYMLIVACTH); p, polar residues (EDKRNQHTS); s, small residues (ACDGNPSTV); a, aromatic residues (YWF).

protein Ksed_22780 from *Kytococcus sedentarius*, HHpred found similarity to a putative nucleotidyltransferase with Protein Data Bank [PDB] code 2FCL with a probability of 96.26). The DUF1814 family, along with the related COG5340 family, is abundant in bacteria and is also present in several archaea.

Involvement in TA systems has been previously proposed for...
another family of nucleotidyltransferases that is extremely widespread in archaea (55). The putative DUF1814-COG5340 TA system might act via a similar mechanism that remains to be characterized experimentally.

The genes discussed above appear to be promising candidates for components of novel defense systems. However, the strong association of some other gene families with DIs is puzzling. For example, we identified several families of membrane proteins (pfam12553, COG4854, COG5658, and a few others) which are abundant in DIs. A small protein with one transmembrane helix from the DUF3742 (pfam12553) family shows the strongest linkage to DIs. This protein is often encoded in predicted operons with toxins or TA systems and two genes, \( \text{traG} \) and \( \text{traU} \), that are involved in the assembly of conjugal-transfer pili (62). The association with \( \text{traG} \) and \( \text{traU} \) suggests that the respective DIs might represent unrecognized ICE elements or conjugative plasmids (87); alternatively, a distinct mechanism of conjugal transfer of TA systems might exist (see Table S17 in the supplemental material). The association with a number of membrane protein families has been noticed in both recent attempts to identify new TA components employing the “guilt-by-association” approach, but no experimental evidence is available to shed light on the functional implications of this connection (35, 52).

**Analysis of selected islands with a high density of defense genes.** The procedure described above, by design, can detect only genes overrepresented in DIs, i.e., common gene families. Examination of individual islands has the potential to reveal additional, rare defense systems that nevertheless could be potential targets for experimental study. To address this possibility, we selected for an exhaustive analysis 17 DIs with the maximum density of defense directives from different bacterial and archaeal lineages (see Table S18 in the supplemental material). The domain compositions of proteins encoded in these DIs suggest the existence of numerous novel, relatively rare defense systems. For example, in a DI from *Nostoc* sp., we found a pair of genes (alr0507 and alr0508) which encode an uncharacterized predicted ATPase related to COG4637 and a putative TOPRIM domain-containing nuclease of the OLD family (4), respectively. The COG4637 ATPase is closely related to the COG1106 family, which also includes the gene for AbiL. The gene for the AbiL system has been detected on a plasmid in *Lactococcus lactis diacetylactis* (16). The molecular mechanism of this system is unknown. The second component of the lactococcal AbiL system, the \( \text{abiLii} \) gene, is homologous to \( \text{rloB} \), the gene found in the same locus with a type I RM system in *C. jejuni* (59). Sequence analysis using HHpred confirms the presence of the TOPRIM domain in the proteins of this family (e.g., for RloB from *C. jejuni*, HHpred detects profile cd01026, TOPRIM_OLD, with a probability of 94.44). Taken together, these findings lead to the prediction that the alr0507-alr0508 system in *Nostoc* is an Abi system related to the AbiL system of *Lactococcus*.

Two uncharacterized genes, Shewana3_3791 and Shewana3_3792, in *Shewanella* sp. are located in a predicted operon together with several RM genes. Moreover, these genes are found in the same context in several additional genomes, and some homologs of this pair of genes, revealed by exhaustive PSI-BLAST, are present in other DIs (Fig. 9). Using the Shewana3_3792 protein as a query, HHpred detects a statistically significant similarity to the PF08378 profile, NERD nuclease-related domain (probability = 81.17), and other families of the PD-(D/E)XK nuclease superfAMILY. The signature motif PD-ExK is conserved in the majority of the Shewana3_3792 family proteins (see Fig. S1 in the supplemental material). We could not detect any similarity to known families for the Shewana3_3792 homologs. Nevertheless, taken together, these observations suggest that the two uncharacterized genes in

---

**FIG. 9.** Two examples of putative novel defense system components identified in high-density DIs. Designations are as in Fig. 6. The arrows corresponding to genes encoding subunits of RM systems are yellow; TA components are magenta. Arrows for genes encoding putative novel components of defense systems described in this work are outlined in red. Abbreviations: RM, restriction modification; HNH, a motif conserved in the corresponding family of nucleases; Fic, toxin of the corresponding family; PHD, antitoxin of the corresponding family; TOPRIM, conserved catalytic domain of topoisomerases, primases, and OLD family nucleases (4).
Shewanella represent a novel defense module that functions in conjunction with RM systems.

**DIs of *M. aeruginosa***. The genome of the common freshwater cyanobacterium *M. aeruginosa* NIES 843 is remarkable for the abundance and diversity of its identified defense systems. *M. aeruginosa* contains the largest number of identified defense genes (*n* = 492; Fig. 10A) among the 1,055 analyzed genomes, 80% more than the next highest number in *Cyanothecceum* PCC8802 and *Roseiflexus* RS-1 and 4.3 times more than expected from its genome size. Our analysis shows that 1,835 protein-coding genes (29%) of *M. aeruginosa* belong to DIs. We investigated the DIs in this remarkable genome in some detail. The defense systems of *M. aeruginosa* are poorly annotated in current databases. We were able to assign defense-related functions to more than 600 proteins found in DIs but annotated as hypothetical (ftp://ftp.ncbi.nih.gov/pub/wolf/_suppl/defense/). In addition, it appears that many components of TA systems are misannotated (wrong open reading frames [ORFs] predicted in the corresponding loci) and at least 19 ORFs correspond to translated CRISPR repeats.

First, some of the gene families present in the *M. aeruginosa* DIs show a notable expansion of paralogs. For example, COG4636 [PD-(D/E)XK nucleases], which was included in our original positive set because it has been predicted to be a stand-alone defense system related to TA systems (see reference 55), is represented by 98 paralogs, which is on a par with hundreds of copies of transposons of several families identified in *M. aeruginosa* (37). We identified another expanded family (DUF29) which is represented in *M. aeruginosa* by 43 paralogs (25 of these are located in DIs) and is also expanded in *Gloeobacter*, *Cyanothecce*, and *Synechocystis*. The structure of one family representative has been solved and revealed a distinct all-α-helical protein fold (PDB code 3FCN). A multiple-sequence alignment shows the presence of several conserved residues (see Fig. S2 in the supplemental material), including an aspartate, a glutamate, and a histidine, which might form a catalytic triad of a novel nuclease. Thus, it appears likely that this family is functionally similar to COG4636 and could be involved in a TA-like mechanism. Additional families with similar features were identified in DIs, e.g., various derivatives of the PD-(D/E)XK nuclease superfamily, especially COG5464 (42) (also known as YhgA-like family or transposase 31), which resemble transposable elements; however, the function and mechanism of these proteins remain obscure, and it cannot be ruled out that they function like a TA system (Table 2).

Among the known defense systems, the specific expansion of several is notable. For example, among the TA systems, only the HicA-HicB system (35, 52) is substantially expanded: there are at least 31 HicB antitoxins and 21 HicA toxins in *M. aeruginosa*. The expansion of this TA system in cyanobacteria has been described previously, along with the fact that HicB is encoded by stand-alone genes more often than HicA (52). Another notable expansion involves the antitoxin of the COG2442 family (20 paralogs). This protein has been recently identified as a new antitoxin for the PIN nuclease toxin and shown to adopt the DNA/RNA-binding 3-helical-bundle fold (55). In *M. aeruginosa*, it is encoded by multiple stand-alone genes but in three cases is associated with a new family of putative toxins (e.g., MAE_32580) that appear to be specific to cyanobacteria. Although we could not identify reliable sequence similarity for this family, secondary structure prediction and the conservation of an aspartate after the N-terminal beta-strand suggested that this protein is a highly diverged PIN nuclease domain (data not shown). Thus, the analysis of this family in *M. aeruginosa* suggests a potential mechanism of action in *trans* for COG2442 and reveals a new family of putative toxins affected by this antitoxin. (55). The largest expanded family related to RM systems in *M. aeruginosa*, COG1002, with 7 paralogs, belongs to type III RM systems that have only recently been studied in detail (63). In particular, it has been shown that most of these enzymes combine endonuclease and methyltransferase activities in a single polypeptide and modify a conserved adenine on only one DNA strand for host protection (63). Typically, the gene encoding this enzyme is a stand-alone gene and appears to be functionally self-sufficient. All 7 paralogs in *M. aeruginosa* are highly divergent and accordingly are expected to recognize unique DNA sequences.

Numerous additional new components of defense systems were predicted among the genes that constitute the DIs of *M. aeruginosa*. The most interesting cases, which include a rare CRISPR-Cas system variant, several new components of TA and Abi systems, and several families of nucleases, some of which combine features of transposable elements and TA systems, are listed in Table 2, and several selected DIs are shown in Fig. 10B (see ftp://ftp.ncbi.nih.gov/pub/wolf/_suppl/defense/ for additional details).

**Conclusions**. The extensive comparative genomic analysis reported here demonstrates statistically significant clustering of antivirus defense systems, as well as mobilome components, in genomic islands in bacterial and archaeal genomes. The density of DIs in bacterial and archaeal genomes varies over an extremely broad range: some genomes contain only a few DIs, whereas others, for example, the genome of the freshwater cyanobacterium *M. aeruginosa*, are “chock-full” of DIs. We observed a substantial excess of DIs in thermophiles versus mesophiles, in archaea versus bacteria, and in free-living versus host-associated bacteria. The latter difference at least suggests the intuitively expected positive correlation between virus diversity in an environment and the abundance of defense systems. Nevertheless, the functional underpinning(s) of the relationship between DI abundance and the lifestyle of organisms remains a challenge for further study.

The DIs are enriched in putative operons and contain numerous overrepresented gene families. The detailed follow-up sequence analysis of the proteins encoded by genes that are overrepresented in DIs shows that many of them are diverged variants of known defense system components whereas others possess features, such as characteristic operonic organization, that are suggestive of novel defense systems. Examination of the expansive but poorly characterized Pgl defense system indicates that novel mechanisms of antivirus defense in prokaryotes remain to be discovered, and the DIs present abundant material for experimental study of bacterial and archaeal antivirus defense.

With the exception of the CRISPR-Cas systems, different classes of defense systems show nonrandom colocalization in
FIG. 10. The extreme abundance of defense systems in the *M. aeruginosa* genome. (A) Distribution of defense genes in the *M. aeruginosa* NIES-843 and *Paenibacillus* sp. strain JDR-2 genomes. (B) Examples of DIs containing putative novel defense system components. Designations are as in Fig. 6. The arrows denoting cas genes are blue, TA system components are magenta, and transposable elements are black. PIN, RHH, PHD, HTH, RelE, HicA, HicB, MNT, HEPT, and COG4636 are predicted TA system components described in detail previously (55). FS, frameshift; VPEP, family of proteins containing a VPEP motif (28). The cas gene nomenclature is given in accordance with the recently proposed classification (54). Proteins annotated as hypothetical are indicated by asterisks.
| System and representative component | No. of homologs in DI (no. in genome) | Detected similarity(ies) | Gene context | Comment |
|-------------------------------------|--------------------------------------|-------------------------|--------------|---------|
| Putative ABI or TA system           |                                      |                         |              |         |
| MAE_32580                           | 3 (4)                                | None                    | COG2442      | The conserved motifs and predicted secondary structure of this family are consistent with PIN nuclease signature, a toxin often associated with COG2442 |
| MAE_49350                           | 1                                    | PD-(D/E)XK family       | COG1106 (ATPase; AbiL) | Also associated with another ATPase (COG4938); potential substitution of TOPRIM domain function |
| MAE_28990                           | 2 (1)                                | AbiF-like protein family | No           |         |
| MAE_24760                           | 3                                    | RHH (C terminus),       | COG1403 (McrA/HNH) | Two-component system |
| MAE_09110                           | 2 (3)                                | None                    | MAE_09120    | Two-component system; both proteins are predicted to be alpha-helical; abundant in cyanobacteria |
| MAE_04340                           | 2                                    | None                    | MAE_04350    | Two-component system; MAE_04340 family also abundant in Firmicutes (e.g., yozC in Bacillus subtilis) |
| MAE_04400                           | 3                                    | None                    | MAE_04390    | Two-component system of two small proteins specific for cyanobacteria |
| Putative CRISPR-Cas system (cas gene), MAE_30780 | 1                                    | Cas5 (COG1688)          | Predicted Cas7 (MAE_30770), Cas3 (MAE_30790) | Divergent rare variant of CRISPR-Cas system, so far present only in Cyanobacte sp. strain PCC 8802 and two species of Thermus thermophilus |
| Uncharacterized defense system      |                                      |                         |              |         |
| MAE_55110                           | 5                                    | Uncharacterized family  | COG2929/DUF497, which belongs to RelE toxin family | Often stand-alone gene; fused to Era GTPase in Chloroflexi; COG5428 is similar to N-terminal domains of some OB-fold-containing proteins (e.g., PDB code 3bu2) |
| MAE_01610                           | 1                                    | Plasmid stabilization protein SpbA family | No | SpbA family identified on plasmid pH71030 of Bacillus thuringiensis and shown to be responsible for plasmid maintenance (50) |
| MAE_11480                           | 25                                   | Uncharacterized family  | No           | See text |
| MAE_57710                           | 5 (14)                               | PD-(D/E)XK nuclease     | No           | Distant homologs of COG5493/DUF1626 family for which the nuclease domain was identified before (42); also expanded in Pyrobaculum aerophilum (called paRep7 family) |
| MAE_00370                           | 27 (46)                              | PD-(D/E)XK nuclease; COG5493 | No | Often expanded in genomes with many identical copies, thus considered to be transposable element |
| MAE_11990                           | 1 (1)                                | PD-(D/E)XK nuclease; COG1637 (DUF91) | Encoded in operon with RM systems in many other genomes | N-terminal nuclease is distantly related to NucS endonuclease from archaea, which interacts with the replication clamp PCNA (72) |
DIs. It remains unclear to what extent these associations reflect functional cooperation between different defense systems and to what extent the islands are genomic “sinks” that accumulate diverse nonessential genes, particularly those acquired via HGT. The findings that the characteristics of DIs resemble those of mobile islands and that defense and mobile islands are intermixed within many islands argue for the latter possibility, given that the mobile island components exert little or no fitness effect. The islands are likely to evolve by nonselective “preferential attachment” whereby the probability of fixation of a newly acquired defense or mobile island gene increases with the increase in island size. The addictive properties of the defense systems, in particular TA, are likely to be a major factor behind this preferential attachment.

ACKNOWLEDGMENTS

K.S.M., Y.I.W., and E.V.K. are supported by the Department of Health and Human Services intramural program (NIH, National Library of Medicine). S.S. is supported by the Yoshaya Horowitz Association through the Center for Complexity Science.

REFERENCES

1. Akuljarov, M., and D. R. Lovley. 2010. Interference with histidyl-tRNA synthetase by a CRISPR spacer sequence as a factor in the evolution of Pelo-
baric bacteriocins. BMC Evol. Biol. 10:230.
2. Altschul, S. F., et al. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25:3389–3402.
3. Andreeva, A., et al. 2008. Data growth and its impact on the SCOP database: new developments. Nucleic Acids Res. 36:D419–D425.
4. Aravind, L., D. D. Leipe, and E. V. Koonin. 2002. Extensive domain shuffling in transcription regulators of DNA viruses and implications for the origin of fungal APSES transcription factors. Genome Biol. BResearch00102.
5. Bolotin, A., B. Quinquis, A. Sorokin, and S. D. Ehrlich. 2005. Genome comparison and context of novel restriction endonuclease-like fold families among hypothet-
ical genome islands: tools of bacterial horizontal gene transfer and evolution. FEBS Microbiol. Rev. 33:376–393.
6. Cambray, G., A. M. Guerout, and D. Mazel. 2010. Integrons. Annu. Rev. Genet. 44:141–166.
7. Chinenova, T. A., N. M. Mikkurinen, and N. D. Lomovskaya. 1982. Genetic characteristics of a new phage resistance trait in Streptomyces coelicolor A3(2). Genetika 18:1945–1952.
8. Chopin, M. C., A. Chopin, and E. Bidnenko. 2005. Phage abortive infection in lacticocci: variations on a theme. Curr. Opin. Microbiol. 8:473–479.
9. Cuff, J. A., M. E. Clamp, A. S. Siddiqui, M. Finlay, and G. J. Barton. 1998. JPro: a consensus secondary structure prediction server. Bioinformatics 14:892–893.
10. Deng, Y. M., C. Q. Liu, and N. W. Dunn. 1999. Genetic organization and functional analysis of a novel phage abortive infection system, AbiII, from Lactococcus lactis. J. Bacteriol. 67:135–149.
11. Edgar, R. C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32:1792–1797.
12. Edwards, R. A., and F. Rohwer. 2005. Viral metagenomics. Nat. Rev. Mi-
crobol. 3:504–510.
13. Eigen, M. 1971. Selforganization of matter and the evolution of biological macromolecules. Naturwissenschaften 58:465–523.
14. Fineran, P. C., et al. 2009. The phage abortive infection system, ToxIN, functions as a phage RNA-toxin-antitoxin pair. Proc. Natl. Acad. Sci. U. S. A. 106:894–899.
15. Furutai, K., A. Ake, and I. Kobayashi. 2010. Genome comparison and context analysis reveals putative mobile forms of restriction-modification systems and related rearrangements. Nucleic Acids Res. 38:2428–2443.
16. Furuta, Y., K. Abe, and I. Kobayashi. 2009. Activation and transfer of the chromosomal phage resistance mechanism AbiV in Lactococcus lactis. Int. J. Bacteriol. 9:R123.
17. Geider, K. S., K. Christensen, and A. Lohner-Olesen. 2005. Prokaryotic toxin-antitoxin stress response loci. Nat. Rev. Microbiol. 3:371–382.
18. Gogol, M., and A. Cain. 2004. NERD: a DNA processing-related domain present in the anthrax virulence plasmid, pX01. Trends Biochem. Sci. 29:106–110.
19. Haaber, J., S. Moineau, and K. Hammer. 2009. Activation and transfer of the chromosomal phage resistance mechanism AbiV in Lactococcus lactis. Appl. Environ. Microbiol. 75:3358–3361.
20. Haft, D. H., I. T. Paulsen, N. Ward, and J. D. Selengut. 2006. Exopolysac-
charide-associated protein sorting in environmental organisms: the PEP-
CRISPR/EpsH system. Application of a novel phylogenetic profiling heuris-
tic. BMC Biol. 4:29.
21. Holm, L., and P. Rosenstrom. 2010. DalI server: conservation mapping in BD. Nucleic Acids Res. 38:W545–W549.
22. Hoyle, R. J., et al. 2007. Hypervariation and phase variation in the bacteriophage ‘resistome’. Curr. Opin. Microbiol. 10:356–400.
23. Ishikawa, K., E. Fukuda, and I. Kobayashi. 2010. Conflicts targeting epige-
etic systems and their resolution by cell death: novel concepts for methyl-
specific and other restriction systems. DNA Res. 17:325–342.
24. Iyer, L. M., E. V. Koonin, and L. Aravind. 2002. Extensive domain shuffling in transcription regulators of DNA viruses and implications for the origin of fungal APSES transcription factors. Genome Biol. BResearch00102.
25. Jørgensen, M. G., D. P. Pandey, M. Jaskolska, and K. Gerdes. 2009. HisA of Escherichia coli defines a novel family of translation-independent microRNA interferases in bacteria and archaea. J. Bacteriol. 191:1191–1199.
26. Junhas, M., et al. 2009. Genomic islands: tools of bacterial horizontal gene transfer and evolution. FEMS Microbiol. Rev. 33:376–393.
27. Kojima, K. K., and M. Kanehisa. 2008. The evolution and emergence of RNA viruses. Oxford University Press, Oxford, United Kingdom.
28. Kogure, T., et al. 2011. Defense islands in bacterial and archaeal genomes. Nat. Rev. Microbiol. 8:1196–1207.
29. Kojima, K. K., M. Kanehisa, and M. Kojima. 2009. Balancing at survival’s edge: the structure and adaptive benefits of prokaryotic toxin-antitoxin par-
tens. Curr. Opin. Struct. Biol. 21:109–118.
30. Kolb, B. M. 2008. Ecological models and data in R. Princeton University Press, Princeton, NJ.
31. Kojima, K. K., M. Kanehisa, and M. Kojima. 2009. Balancing at survival’s edge: the structure and adaptive benefits of prokaryotic toxin-antitoxin par-
tens. Curr. Opin. Struct. Biol. 21:109–118.
32. Kool, E., A. Prieto, J. M. Bujnicki, and I. Kobayashi. 2007. Hypervariation and phase variation in the bacteriophage ‘resistome’. Curr. Opin. Microbiol. 10:356–400.
33. Kool, E., A. Prieto, J. M. Bujnicki, and I. Kobayashi. 2007. Hypervariation and phase variation in the bacteriophage ‘resistome’. Curr. Opin. Microbiol. 10:356–400.
34. Kool, E., A. Prieto, J. M. Bujnicki, and I. Kobayashi. 2007. Hypervariation and phase variation in the bacteriophage ‘resistome’. Curr. Opin. Microbiol. 10:356–400.
35. Kool, E., A. Prieto, J. M. Bujnicki, and I. Kobayashi. 2007. Hypervariation and phase variation in the bacteriophage ‘resistome’. Curr. Opin. Microbiol. 10:356–400.
stability of pTH1030, a novel type of gram-positive replicon. Mol. Microbiol. 63:5–46.

51. Makarova, K. S., L. Aravind, N. V. Grishin, I. B. Rogozin, and E. V. Koonin. 2002. A DNA repair system specific for thermophilic Archaea and bacteria from marine environments. J. Mol. Biol. 318:482–496.

52. Makarova, K. S., N. V. Grishin, and E. V. Koonin. 2006. The HicAB cassette, a putative novel, RNA-targeting toxin-antitoxin system in archaea and bacteria. Bioinformatics 22:2581–2584.

53. Makarova, K. S., N. V. Grishin, S. A. Shahalina, Y. I. Wolf, and E. V. Koonin. 2006. A putative RNA-interference-based immune system in prokaryotes: computational analysis of the predicted enzymatic machinery, functional analogies with eukaryotic RNAs, and hypothetical mechanisms of action. Biol. Direct 4:29.

54. Makarova, K. S., et al. 2011. Evolution and classification of the CRISPR-Cas systems. Nat. Rev. Microbiol. 4:97–477.

55. Makarova, K. S., Y. I. Wolf, and E. V. Koonin. 2009. Comprehensive comparative-genomic analysis of type 2 toxin-antitoxin systems and related mobile stress response systems in prokaryotes. Biol. Direct 4:19.

56. Makarova, K. S., Y. I. Wolf, J. van der Oost, and E. V. Koonin. 2009. Prokaryotic homologs of Argonaute proteins are predicted to function as key components of a novel system of defense against mobile genetic elements. Biol. Direct 4:29.

57. Marchler-Bauer, A., et al. 2009. CDD: specific functional annotation with the Conserved Domain Database. Nucleic Acids Res. 37:D205–D210.

58. McKinley, B. A., and M. V. Sukhodolets. 2003. Escherichia coli RNA polymerase-associated SWI/SNF protein RapA: evidence for RNA-directed binding and remodeling activity. Nucleic Acids Res. 35:7044–7060.

59. Miller, W. G., et al. 2005. Diversity within the Campylobacter jejuni type I restriction-modification loci. Microbiology 151:337–351.

60. Mozica, F. J., C. Diez-Villanueva, J. Garcia-Martinez, and E. Soria. 2005. Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. J. Mol. Evol. 60:174–182.

61. Molina, N., and E. van Nimwegen. 2009. Scaling laws in functional genome content across prokaryotic clades and lifestyles. Trends Genet. 25:243–247.

62. Moore, D., et al. 1990. Characterization of the F-plasmid conjugative transfer gene traU. J. Bacteriol. 172:4263–4270.

63. Morgan, R. D., E. A. Dwinnell, T. K. Bhatia, E. M. Lang, and Y. A. Luyten. 2001. The MmeI family: type II restriction-modification enzymes that employ single-strand modification for host protection. Nucleic Acids Res. 37:5208–5221.

64. Murzin, A. G., S. E. Brenner, T. Hubbard, and C. Chothia. 1995. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J. Mol. Biol. 247:536–540.

65. Muzzin, O., et al. 1998. Disruption of Escherichia coli hepA, an RNA polymerase-associated protein, causes UV sensitivity. J. Biol. Chem. 273:15157–15161.

66. Naderer, M., J. R. Brust, D. Knowle, and R. M. Blumenthal. 2002. Mobility of a restriction-modification system revealed by its genetic contexts in three hosts. J. Bacteriol. 184:2411–2419.

67. Nechaev, S., and K. Severinov. 2008. The elusive object of desire—interactions of bacteriophages and their hosts. Curr. Opin. Microbiol. 11:186–193.

68. Novichkov, P. S., I. Ratnere, Y. I. Wolf, E. V. Koonin, and I. Dubchak. 2000. ATGC: a database of orthologous genes from closely related prokaryotic genomes. Nucleic Acids Res. 28(Database issue):D448–D454.

69. O’Connor, L., M. Tanguay, and G. F. Fitzgerald. 1999. Expression, regulation, and mode of action of the AbiG abortive infection system of Lactococcus lactis subsp. cremoris UC653. Appl. Environ. Microbiol. 65:330–335.

70. Orlowski, J., and J. M. Bujnicki. 2008. Structural and evolutionary classification of type II restriction enzymes based on theoretical and experimental analyses. Nucleic Acids Res. 36:3552–3569.

71. Paterson, S., et al. 2010. Antagonistic coevolution accelerates molecular evolution. Nature 464:275–278.

72. Ren, B., et al. 2009. Structure and function of a novel endonuclease acting on branched DNA substrates. EMBO J. 28:2479–2489.

73. Salgado, H., G. Moreno-Hagelsieb, T. F. Smith, and J. Collado-Vides. 2000. CORUM in Escherichia coli: genome analyses and predictions. Proc. Natl. Acad. Sci. U. S. A. 97:6652–6657.

74. Shaw, G., et al. 2008. Structure of RapA, a Swi2/Snf2 protein that recycles RNA polymerase during transcription. Structure 16:1417–1427.

75. Singh, J., A. Bieger, and A. N. Lupas. 2005. The Hi1pred interactive server for protein homology detection and structure prediction. Nucleic Acids Res. 33:W244–W248.

76. Sonnhammer, E. L., G. von Heijne, and A. Krogh. 1998. A hidden Markov model for predicting transmembrane helices in protein sequences. Proc. Int. Conf. Intell. Syst. Mol. Biol. 6:175–182.

77. Stern, A. L., K. Rogen, W. Wurstel, G. Amitai, and R. Sorek. 2010. Self-targeting by CRISPR: gene regulation or autoimmunity? Trends Genet. 26:335–340.

78. Stern, A. R., and S. Oreg. 2011. The phage-host arms race: shaping the evolution of microbes. Bioessays 33:43–51.

79. Stouffer, S. A., E. A. Suchman, L. C. DeVincenzy, S. A. Star, and R. M. J. Williams. 1949. The American soldier, vol. 1. Adjustment during army life. Princeton University Press, Princeton, NJ.

80. Sturny, P., and M. C. Smith. 2005. Diversity within the Campylobacter jejuni type I restriction-modification loci. Trends Genet. 21:479–484.

81. Sumby, P., and M. C. Smith. 2002. Genetics of the phage growth limitation (Pgl) system of Streptomyces coelicolor A3(2). Mol. Microbiol. 44:489–500.

82. Suttle, C. A. 2007. Marine viruses—major players in the global ecosystem. Nat. Rev. Microbiol. 5:801–812.

83. Tatusov, R. L., et al. 2003. The COG database: an updated version includes eukaryotes. BMC Bioinformatics 4:41.

84. Tesfazgi Mebrhatu, M., et al. 2011. Evidence for an evolutionary antagonism between Mrr and type III modification systems. Nucleic Acids Res. 39:5991–6001.

85. Toleman, M. A., and T. R. Walsh. 2011. Combinatorial events of insertion sequences and ICE in Gram-negative bacteria. FEMS Microbiol. Rev. 35:912–935.

86. van der Oost, J., M. M. Jore, E. R. Westra, M. Lundgren, and S. J. Brons. 2009. CRISPR-based adaptive and heritable immunity in prokaryotes. Trends Biochem. Sci. 34:401–407.

87. Van Melderen, L. 2010. Toxin-antitoxin systems: why so many, what for? Curr. Opin. Microbiol. 13:781–785.

88. Van Melderen, L., and M. Saavedra De Bast. 2009. Bacterial toxin-antitoxin systems: more than selfish entities? PLoS Genet. 5:e1000437.

89. Van Nimwegen, E. 2003. Scaling laws in the functional content of genomes. Trends Genet. 19:479–484.

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

91. Wolf, Y. I., I. B. Rogozin, A. S. Kondrashov, and E. V. Koonin. 2001. Genome alignment, evolution of prokaryotic genome organization, and prediction of gene function using genomic context. Genome Res. 11:356–372.

92. Zalatan, J. G., T. D. Fenn, and D. Herschlag. 2009. The phage-host arms race: shaping the evolution of microbes. Bioessays 33:43–51.

93. Zhou, X., et al. 2005. A novel DNA modification by sulphur. Mol. Microbiol. 57:1428–1438.