Systematic prediction of genes functionally linked to CRISPR-Cas systems by gene neighborhood analysis

Sergey A. Shmakov[a,b], Kira S. Makarova[b], Yuri I. Wolf[b], Konstantin V. Severinov[a,c,d], and Eugene V. Koonin[b,1]

[a]Center for Data-Intensive Biomedicine and Biotechnology, Skolkovo Institute of Science and Technology, 143025 Skolkovo, Russia; [b]National Center for Biotechnology Information, National Library of Medicine, Bethesda, MD 20894; [c]Waksman Institute for Microbiology, Rutgers, The State University of New Jersey, Piscataway, NJ 08854; and [d]Laboratory of Genetic Regulation of Prokaryotic Mobile Genetic Elements, Institute of Molecular Genetics, Russian Academy of Sciences, 123182 Moscow, Russia

Contributed by Eugene V. Koonin, April 18, 2018 (sent for review March 1, 2018; reviewed by Lennart Randau, Virginijus Siksnys, and Rotem Sorek)

The CRISPR-Cas systems of bacterial and archaeal adaptive immunity consist of direct repeat arrays separated by unique spacers and multiple CRISPR-associated (*cas*) genes encoding proteins that mediate all stages of the CRISPR response. In addition to the relatively small set of core *cas* genes that are typically present in all CRISPR-Cas systems of a given (sub)type and are essential for the defense function, numerous genes occur in CRISPR-*cas* loci only sporadically. Some of these have been shown to perform various ancillary roles in CRISPR response, but the functional relevance of most remains unknown. We developed a computational strategy for systematically detecting genes that are likely to be functionally linked to CRISPR-Cas. The approach is based on a "CRISPRicity" metric that measures the strength of CRISPR association for all protein-coding genes from sequenced bacterial and archaeal genomes. Uncharacterized genes with CRISPRicity values comparable to those of *cas* genes are considered candidate CRISPR-linked genes. We describe additional criteria to predict functionally relevance for genes in the candidate set and identify 79 genes as strong candidates for functional association with CRISPR-Cas systems. A substantial majority of these CRISPR-linked genes reside in type III CRISPR-*cas* loci, which implies exceptional functional versatility of type III systems. Numerous candidate CRISPR-linked genes encode integral membrane proteins suggestive of tight membrane association of CRISPR-Cas systems, whereas many others encode proteins implicated in various signal transduction pathways. These predictions provide ample material for improving annotation of CRISPR-*cas* loci and experimental characterization of previously unsuspected aspects of CRISPR-Cas system functionality.

CRISPR-Cas | signaling | membrane proteins | computational genomics | gene neighborhoods

Driven largely by the exceptional recent success of Cas9, Cas12, and Cas13 RNA-guided nucleases as the new generation of genome and transcriptome editing tools, comparative genomics, structures, biochemical activities, and biological functions of CRISPR-Cas systems and individual Cas proteins have been studied in exquisite detail (1–5). The CRISPR-Cas immune response is conventionally described in terms of three distinct stages: (*i*) adaptation, (*ii*) expression and maturation of CRISPR (cr) RNA, and (*iii*) interference. At the adaptation stage, a distinct complex of Cas proteins binds to a target DNA and, typically after recognizing a short (2–4 bp) motif known as protospacer-adjacent motif (PAM), excises a portion of the target DNA (protospacer), and inserts it into the CRISPR array (most often, at the beginning of the array, downstream of the leader sequence) as a spacer (6, 7). The adaptation process creates immune memory, that is, "vaccinates" a bacterium or archaeon against subsequent infection with the memorized agent. At the expression-maturation stage, the CRISPR array is typically transcribed into precrRNA that is then processed into mature crRNAs, each consisting of a spacer and a portion of an adjacent repeat, by a distinct complex of Cas proteins or a single, large Cas protein, or an external, non-Cas RNase (8). At the final, interference stage, the crRNA bound to Cas proteins is employed as the guide to recognize the protospacer or a closely similar sequence in an invading genome of a virus or plasmid that is then cleaved and inactivated by Cas nuclease(s) (9, 10).

Under the current classification, the CRISPR-Cas systems are divided into two classes, which radically differ with respect to the composition and structure of the effector modules that are responsible for the interference and, in most of the CRISPR-Cas types, also the processing stages (11, 12). In class 1 systems, which include types I, III, and IV, the effector module is a complex of several Cas proteins that perform a tightly coordinated sequence of reactions, from precrRNA processing to target cleavage. In class 2 systems, including types II, V, and VI, all activities of the effector module reside in a single, large multidomain protein such as Cas9 in type II, the programmable endonuclease that is most widely used for genome-editing applications (2, 3, 13).

The biochemical activities and biological functions of the 13 families of core Cas proteins that are essential for each of the three stages of the CRISPR immune response in different types of CRISPR-Cas have been extensively studied, although some notable gaps in knowledge remain (4, 14, 15). Specifically, Cas1 and Cas2 form the adaptation complex that is universal to all autonomous CRISPR-Cas systems (many type III loci and a few in other types lack the adaptation module and apparently rely on the adaptation machinery of other CRISPR-Cas systems in the same organism which they recruit *in trans*) (6, 7, 16–22). Cas3 is a helicase that typically also contains a nuclease domain

Significance

The CRISPR-Cas systems that mediate adaptive immunity in bacteria and archaea encompass a small set of core *cas* genes that are essential in diverse systems belonging to the same subtype, type, or class. However, a much greater number of genes only sporadically co-occur with CRISPR-Cas, and for most of these, involvement in CRISPR-Cas functions has not been demonstrated. We developed a computational strategy that provides for systematic identification of CRISPR-linked proteins and prediction of their functional association with CRISPR-Cas systems, and employed it to identify 79 previously undetected, putative CRISPR-accessory proteins. A large fraction of these proteins are predicted to be membrane-associated, revealing a potentially unknown side of CRISPR biology.

Author contributions: K.S.M., Y.I.W., and E.V.K. designed research; S.A.S. and K.S.M. performed research; S.A.S., K.S.M., Y.I.W., K.V.S., and E.V.K. analyzed data; and E.V.K. wrote the paper.

Reviewers: L.R., Max Planck Institute for Terrestrial Microbiology; V.S., Vilnius University; and R.S., Weizmann Institute of Science.

The authors declare no conflict of interest.

[1]To whom correspondence should be addressed. Email: koonin@ncbi.nlm.nih.gov.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1803440115/-/DCSupplemental.

Published online May 21, 2018.

and is involved in target cleavage in type I systems (23–27). Cas4 is an endonuclease that is required for adaptation in many CRISPR-Cas variants, although its exact functions remain unknown (28–30). Cas5, Cas6, and Cas7 are distantly related members of the so-called RAMP domain superfamily (31), which includes RNases involved in precrRNA processing in types I and III (Cas6 and some variants of Cas5), as well as enzymatically inactive RNA-binding proteins that form the backbone of type I and type III effector complexes (9, 32–38). Cas8 is the enzymatically inactive large subunit of type I effector complexes (39). Cas9 is the type II effector nuclease that also contributes to spacer acquisition (15, 22, 40–43). Cas10 is the large subunit of the type III effector complexes that contains a Palm domain homologous to those of DNA polymerases and nucleotide cyclases (44–46), and possesses an oligoA synthetase activity (47, 48). Cas11 is the small subunit of the type I and type III effector complexes (49). Cas12 is the effector endonuclease of type V (12, 50, 51). Finally, Cas13 is the effector RNase of type VI (50–55).

In addition to the core proteins included in the formal Cas nomenclature, numerous proteins are found in various subsets of CRISPR-Cas systems and are generally thought of as performing accessory, in particular, regulatory functions in the CRISPR response (4, 11, 56). Admittedly, the separation of the proteins encoded in the CRISPR loci into Cas and accessory groups is somewhat arbitrary because some of the "accessory" ones play important roles in the functionality of the system. As a case in point, many type III systems employ an alternative mechanism of adaptation, namely, spacer acquisition from RNA via reverse transcription by a reverse transcriptase (RT) that is encoded in the CRISPR-*cas* locus, often being fused to the Cas1 protein (57, 58). Another striking case of an "accessory" protein turning out to be an essential component of CRISPR-Cas systems is the Csm6 protein that is also common among type III systems and consists of a nucleotide ligand-binding CARF domain fused to a HEPN RNase domain (59). It has been shown in two independent studies that upon target recognition by Csm6-containing type III systems, Cas10 is induced to synthesize oligoA molecules that are bound by the CARF domain of Csm6 resulting in stimulation of the RNase activity of the HEPN domain, which then initiates the immune response (47, 48). However, for most of the putative accessory proteins encoded in CRISPR-*cas* loci, the functions and the very relevance for the CRISPR response remain unknown. The question of functional relevance is far from being moot because defense islands in microbial genomes appear to be "genomic junkyards" that often accommodate functionally irrelevant genes (60).

We sought to develop a computational strategy for systematic identification of proteins that are nonrandomly associated with CRISPR-Cas systems and assess their functional relevance. The approach is based on the CRISPRicity metric that measures the strength of CRISPR association for individual genes. Genes with CRISPRicity values similar to those of *cas* genes were considered candidates for previously undetected accessory genes, and the encoded proteins were examined in depth using sensitive methods for domain identification and coevolution analysis. As a result, 79 genes encoding diverse proteins were predicted to be involved in CRISPR-Cas functions, although, mostly, not connected to CRISPR-Cas previously. A substantial majority of the detected putative CRISPR-accessory proteins are encoded in type III CRISPR-*cas* loci, and many of these are implicated in membrane association of these systems and/or various signaling pathways.

Results and Discussion

CRISPRicity: A Measure for Predicting Functionally Relevant Connections of Genes to CRISPR-Cas Systems. In bacteria and archaea, functionally linked genes are often organized into operons, that is, arrays of codirected, cotranscribed, and cotranslated genes (61–63). The evolutionary dynamics of operons is such that, for a group of genes that are involved in the same pathway or functional system, different microbial genomes often contain partially overlapping operons encoding subsets of proteins involved in the respective process (64). Comparative analysis of gene neighborhoods in multiple genomes often provides for a complete delineation of the components of functional systems. Indeed, precisely this type of analysis resulted in the description of the group of functionally linked proteins that later became known as Cas (44).

We sought to expand the set of proteins that are functionally linked to the CRISPR-mediated immunity by enumerating all genes in the CRISPR neighborhoods and attempting to distinguish functionally relevant genes from spurious ones. To this end, a dedicated computational pipeline was constructed (Fig. 1). Briefly, we identified the union of all genes located in the vicinity (±10 kb) of CRISPR arrays, *cas1* genes (representing the CRISPR-Cas adaptation module) or CRISPR effector modules (for details, see *Materials and Methods*; *SI Appendix*, *Supporting Information File 1*), hereinafter, CRISPR-linked genes. The union of all of these genes was taken so as to maximize the likelihood of detection of CRISPR-linked genes because certain CRISPR-*cas* loci might lack one or even two of these key elements. All genes in the CRISPR-cas neighborhoods were represented as

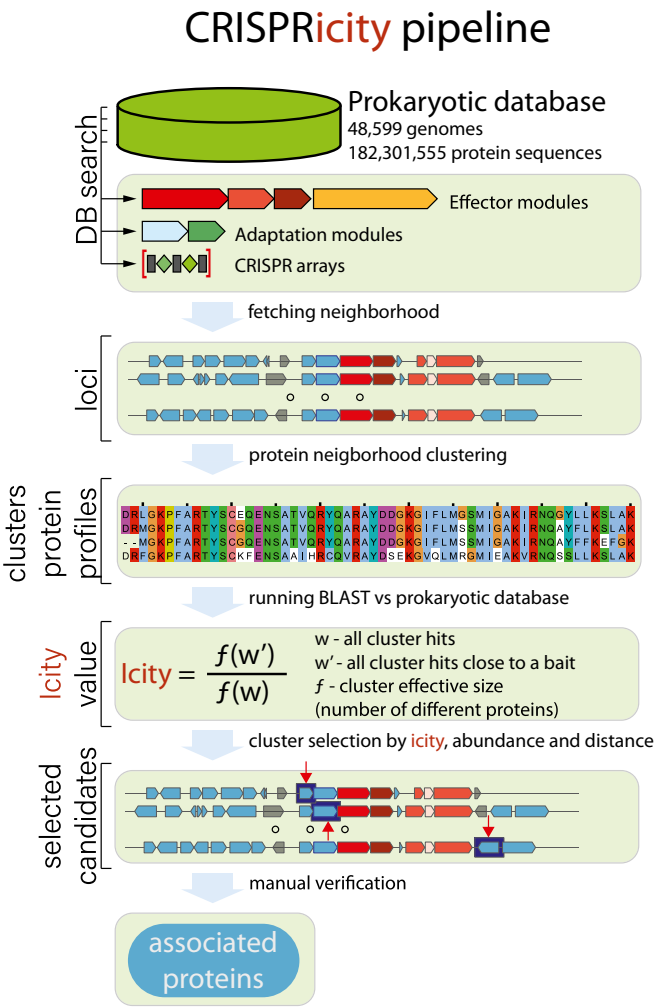

Fig. 1. The computational pipeline for the analysis of the CRISPR-linked gene space.

points in three dimensions defined by the following: (*i*) CRISPRicity, that is, the ratio of the number of CRISPR-linked occurrences of the given gene to its total number of occurrences in the analyzed genomes; (*ii*) abundance (total number of occurrences); and (*iii*) distance from the respective genomic anchor (CRISPR array, *cas1* or effector). All of the genes in the neighborhoods were classified into (*i*) CRISPR-associated (including both *cas* genes and previously identified accessory genes), (*ii*) non-CRISPR (i.e., genes with well-characterized functions unrelated to the CRISPR immunity), and (*iii*) unknowns. The counts of the genes in each of the three classes were obtained for each voxel in the space of the three coordinates defined above, and the ratio of the probability mass of *cas* genes jointly with the recognized CRISPR-accessory genes (D_c) to that of the non-CRISPR genes (D_n) was calculated (hereinafter, CRISPR index; see *Materials and Methods* for details). The "unknown" and "non-CRISPR" genes with $D_c/D_n > 2$ were considered candidates for previously undetected CRISPR-linked genes. This threshold was chosen to select genes with similar characteristics to genes known to be functionally linked to CRISPR-Cas systems. Indeed, the range of CRISPR index values for the candidate accessory proteins identified by this procedure overlapped that for Cas proteins and functionally characterized accessory proteins, which is compatible with their functional relevance of the previously undetected candidates (Fig. 2*A*). Clearly, the CRISPR index cutoff can be adjusted to search for stricter or looser associations.

Altogether, the above procedure yielded 468 putative distinct candidate CRISPR-linked genes (i.e., clusters of related sequences) (Fig. 2 *A* and *B* and Dataset S1). The 360 (predicted) clusters that included homologous proteins with sequences longer than 100 aa were thoroughly explored case by case using methods for sequence profile comparison and phylogenetic analysis (see *Materials and Methods* for details). From this initial list of candidates, 67 proteins were found to be previously unannotated, diverged variants of known Cas and accessory proteins; 81 were classified as likely previously undetected CRISPR-linked proteins; and the remaining 212 were inferred to be, most likely, irrelevant for the CRISPR function based on the genomic context analysis (poorly conserved genes represented only in

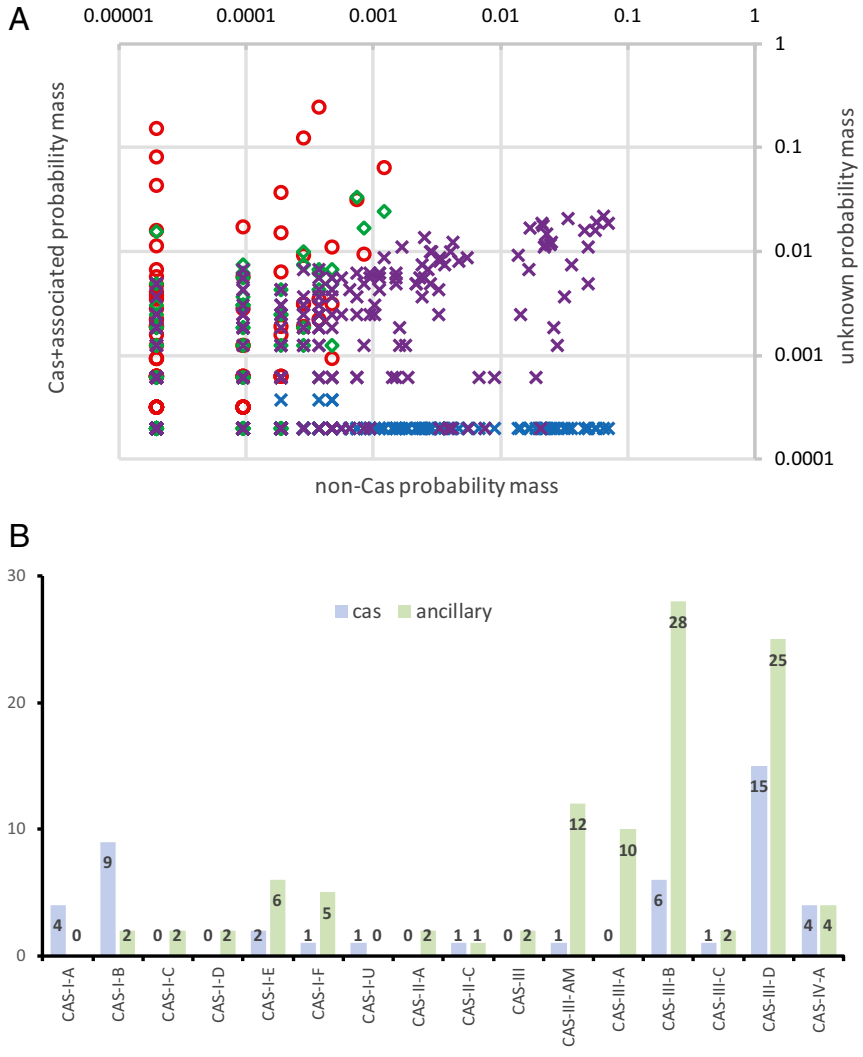

Fig. 2. Protein clusters in CRISPR-*cas* neighborhoods. (*A*) Distribution of voxels in the CRISPRicity-abundance-distance space. Red circles: probability mass distribution for the union of *cas* and previously identified CRISPR-associated genes, voxels with CRISPR index $I > 2$; blue crosses: probability mass distribution for the union of *cas* and previously identified CRISPR-associated genes, voxels with CRISPR index $I < 2$; green diamonds: probability mass distribution for unknown genes, voxels with CRISPR index $I > 2$ (candidate CRISPR-linked genes); purple crosses: probability mass distribution for unknown genes, voxels with CRISPR index $I < 2$. (*B*) Breakdown of previously undetected CRISPR-linked protein clusters by the CRISPR-Cas types and subtypes.

closely related genomes, common components of defense islands that only occasionally co-occur with *cas* genes, gene fragments, and ORFs overlapping with CRISPR arrays; Dataset S1). Previously missed members of Cas protein families known for their fast evolution, in particular, both the large and the small subunits of type I effector complexes and the type III small subunit, were identified in CRISPR-*cas* loci of most subtypes of both types (Fig. 2*B* and Dataset S1). Additionally, detection of previously unidentified Cas proteins was common for type IV systems that are typically carried by plasmids and appear to be the most diverged among the known CRISPR-Cas systems (Fig. 2*B* and Dataset S1). Strikingly, the distribution of the predicted accessory proteins among CRISPR-Cas types was far more skewed than the distribution of previously undetected Cas proteins: the great majority of candidate accessory genes were detected in type III loci (Fig. 2*B* and Dataset S1). To assist future annotation of CRISPR-Cas loci, we constructed sequence alignments and the corresponding position-specific scoring matrices (profiles) for CRISPR-linked proteins detected in the present work (*SI Appendix*, *Supporting Information Files 2* and *3*). When this work

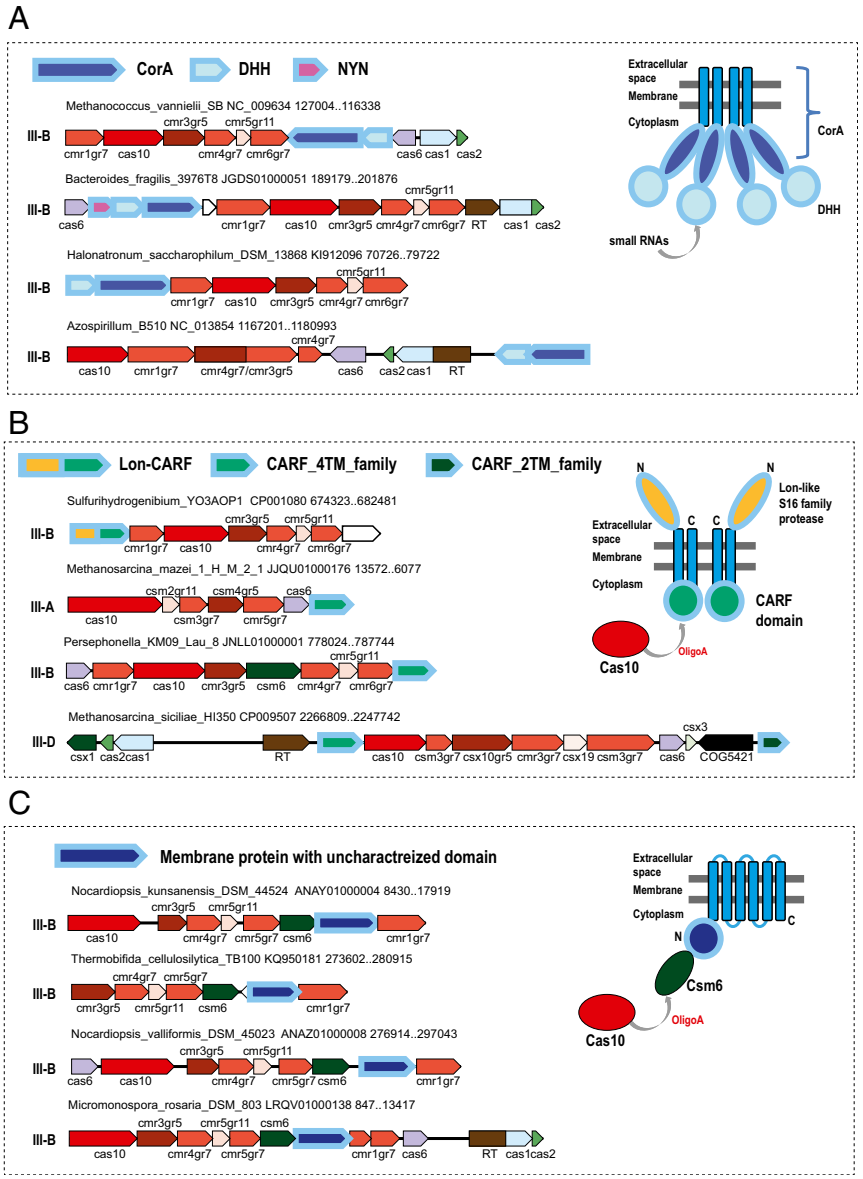

Fig. 3. Locus organization of type III CRISPR-Cas systems containing predicted CRISPR-linked genes encoding membrane proteins. (*A*) CorA, divalent cation membrane channel encoded in type III-B CRISPR-*cas* loci along with two distinct nucleases. (*B*) Membrane-associated CARF domain-containing proteins. (*C*) Uncharacterized membrane protein family in diverse type III loci. For each locus, species name, genome accession number, and the respective nucleotide coordinates and CRISPR-Cas system subtype are indicated. The genes in a representative locus are shown by block arrows, which show the transcription direction. The scale of an arrow is roughly proportional to the respective gene length. Homologous genes and domains are color-coded; empty arrows show predicted genes without detectable homologs. On the *Right*, models of the membrane topology of the predicted CRISPR-linked membrane proteins protein are shown according to the TMHMM predictions. Hypothetical interactions of the identified CRISPR-linked proteins with CRISPR-Cas system components are also depicted (see *Predicted CRISPR-Linked Proteins: Membrane Connections and Signal Transduction*). The *cas* gene names follow the current nomenclature (11); for several core *cas* genes, an extension specifies the gene group (gr5, gr6, gr7, groups 5, 6, and 7 of the RAMP superfamily, respectively; gr8, large subunit of the effector complex; gr11, small subunit of the effector complex). Abbreviations and other gene names: CARF, CRISPR-associated Rossmann fold domain; COG5421, transposase of COG5421 family; DHH, DHH family nuclease; Lon, Lon family protease; NYN, NYN family nuclease; RT, reverse transcriptase; TM, transmembrane helix.

was being prepared for submission, a preprint describing comparative analysis of type III CRISPR-*cas* loci aimed at identification of CRISPR-linked genes has been posted (65); the lists of the predictions from the two studies partially overlap.

Predicted CRISPR-Linked Proteins: Membrane Connections and Signal Transduction. The set of predicted CRISPR-accessory proteins included those that have been shown to stably co-occur or even function jointly with specific CRISPR-Cas variants but have not been formally listed as CRISPR-linked. These included RT (57, 58), Argonaute family proteins (66, 67), and transposon-encoded proteins of the TniQ family (68). However, the majority of the predicted CRISPR-linked proteins have not been reported previously or at least have not been explored in detail. Below, we discuss several examples of previously undetected CRISPR-linked genes that are representative of the findings made through the CRISPRicity analysis. A major theme that is emerging from the case-by-case examination of the predicted CRISPR accessory genes is the potential membrane connection of CRISPR-Cas systems, especially, those of type III (Dataset S1). The most abundant of these genes is *corA*, which encodes a widespread divalent cation channel in bacteria and archaea where it provides the primary route for electrophoretic Mg^{2+} uptake (69) (Fig. 3*A*). The CorA protein is encoded in numerous loci of subtype III-B; some cases of this association have been noticed in previous analyses (70, 71). These CorA-encoding type III loci show considerable diversity of genome architectures, and in many of them, the *corA* gene is adjacent to a gene coding for a DHH family of nucleases (72) or even fused to it (e.g., EDN71418.1 from *Beggiatoa* sp. PS). Moreover, some of these loci also contain a gene for another predicted nuclease (RNase), one of the NYN family (73) (Fig. 3*A*). Structural analysis of CorA has shown that the protein is a pentamer, with each subunit contributing two transmembrane (TM) helices and containing a bulky cytosolic part (74, 75). The stable association of CorA with type III-B CRISPR-Cas strongly suggests a functional connection, and more specifically, a link between the CRISPR-mediated defense and membrane processes. Moreover, in a previous analysis, we have serendipitously detected evidence of recombination within III-B loci of *Clostridium botulinum* where *corA* segregated with the effector genes, suggestive of coordinated functions (71). However, predicting the CRISPR-linked functions of CorA more precisely is difficult. The nuclease connection suggests that CorA might regulate additional cleavage of the target and/or its transcripts DNA during the CRISPR response, but more generally, it seems likely that CorA was exapted for membrane tethering of the respective CRISPR-Cas systems.

Numerous type III systems of subtypes A, B, and D encompass genes that encode previously undetected, highly diverged proteins containing a CARF domain (76) and two predicted transmembrane helices; additionally, some of these proteins contain a fused, diverged Lon family (77) protease domain (Fig. 3*B*). This variety of CARF domains has been independently described previously as SAVED domains (78). Membrane topology prediction suggests that the CARF domain faces the cytosol, whereas the Lon protease is extracellular (Fig. 3*B*). Given that the respective CRISPR-*cas* loci do not encode any other CARF domains but possess Cas10 proteins with predicted nucleotide polymerization activity, it appears most likely that the membrane-bound CARFs recognize signaling oligoA molecules synthesized by Cas10. However, the nature of the effector that could be activated by such binding remains unclear. The Lon family protease might fulfill this function in the case of the respective fusion proteins, but we cannot currently predict the specific mechanism; in other cases, the protease or another effector could be recruited *in trans*. In other cases, predicted membrane proteins are stably associated with type III CRISPR-Cas but contain no identifiable soluble domains that would provide for a specific functional prediction; nevertheless, it appears likely that such proteins anchor the CRISPR-Cas machinery in the bacterial membrane (Fig. 3*C*). A membrane association of at least some CRISPR-Cas system reverberates with the previous reports on activation of the *Escherichia coli* type I-E CRISPR-Cas system by envelope stress (79) and on enhancement of bacterial envelope integrity by type II-B CRISPR-Cas system of *Francisella novicida* (80).

A distinct variant of apparently degenerate I-E systems that lack Cas1, Cas2, and Cas3, and accordingly, cannot be active in either adaptation or target cleavage encode a predicted NTPase of the STAND superfamily (81) that, in addition to the P-loop NTPase domain, contains a cassette of TPR repeats (Fig. 4*A*). The STAND NTPases that typically contain protein–protein interaction domains, such as TPR, are involved in various signal transduction networks that are poorly characterized in prokaryotes but in eukaryotes contribute to various forms of programmed cell death (81, 82). Phylogenetic analysis of STAND NTPases shows that the CRISPR-associated ones form a strongly supported branch (Fig. 4*A* and *SI Appendix*, *Supporting Information File 4*). Given the considerable diversity of gene arrangements in these loci, the monophyly of the NTPases implies their long-term association with CRISPR-Cas systems. Notably, in many cases, the gene adjacent to the NTPase gene encodes a predicted small membrane protein (Fig. 4*A*), suggesting, once again, membrane association of CRISPR-Cas. This particular variant of subtype I-E is likely to perform a nondefense, probably, regulatory function. Conceivably, the STAND NTPase connects these CRISPR-Cas systems to signaling pathways that remain to be identified; involvement in programmed cell death or dormancy induction seems a plausible possibility.

Many subtype IV-B loci that are typically located on plasmids or predicted prophages (Fig. 4*B*) encode a predicted enzyme of the CysH family, which belongs to the adenosine 5′-phosphosulfate (PAPS) reductase family (83). Some of these loci also encode a predicted enzyme of the ADP ribosyltransferase family (84) (Fig. 4*B*). Similarly to the subtype I-E systems discussed above as well as "minimal" I-F systems encoded by Tn7-like transposons (68), type IV systems lack nucleases that could cleave the target DNA, and therefore can be predicted to perform nondefense functions similarly to transposon-encoded CRISPR-Cas systems (68). Analogously to the case of the STAND NTPases, the CRISPR-associated CysH homologs comprise a well-supported clade in the phylogenetic tree of the CysH protein family (Fig. 4*B* and *SI Appendix*, *Supporting Information File 5*). As with other predicted CRISPR accessory genes, the CysH-like enzyme and the associated proteins might play a role in a signal transduction pathway connecting CRISPR-Cas with cellular regulatory networks and perhaps stabilizing the prophages and plasmids in the host bacteria.

Coevolution of Predicted CRISPR-Linked Genes with Bona Fide Cas Genes. The predicted CRISPR-linked genes recur in multiple CRISPR-*cas* loci and accordingly can be predicted to contribute to the functions of the respective CRISPR-Cas systems. We sought to explore the linkage between these genes and CRISPR-Cas at a deeper level. To this end, we performed an analysis of potential coevolution between those of the predicted accessory genes that are sufficiently widespread and well-conserved with signature *cas* genes. Phylogenetic trees were constructed for the analyzed CRISPR-linked genes and compared with trees of *cas* genes and, as a control, to those for 16S RNA (a proxy for the species tree of the respective microbes), and the evolutionary distances between the respective genes from different organisms extracted from the trees were compared (see *Materials and Methods* for details). Notably, the evolutionary distances for the *corA* genes were strongly, positively correlated with the distances

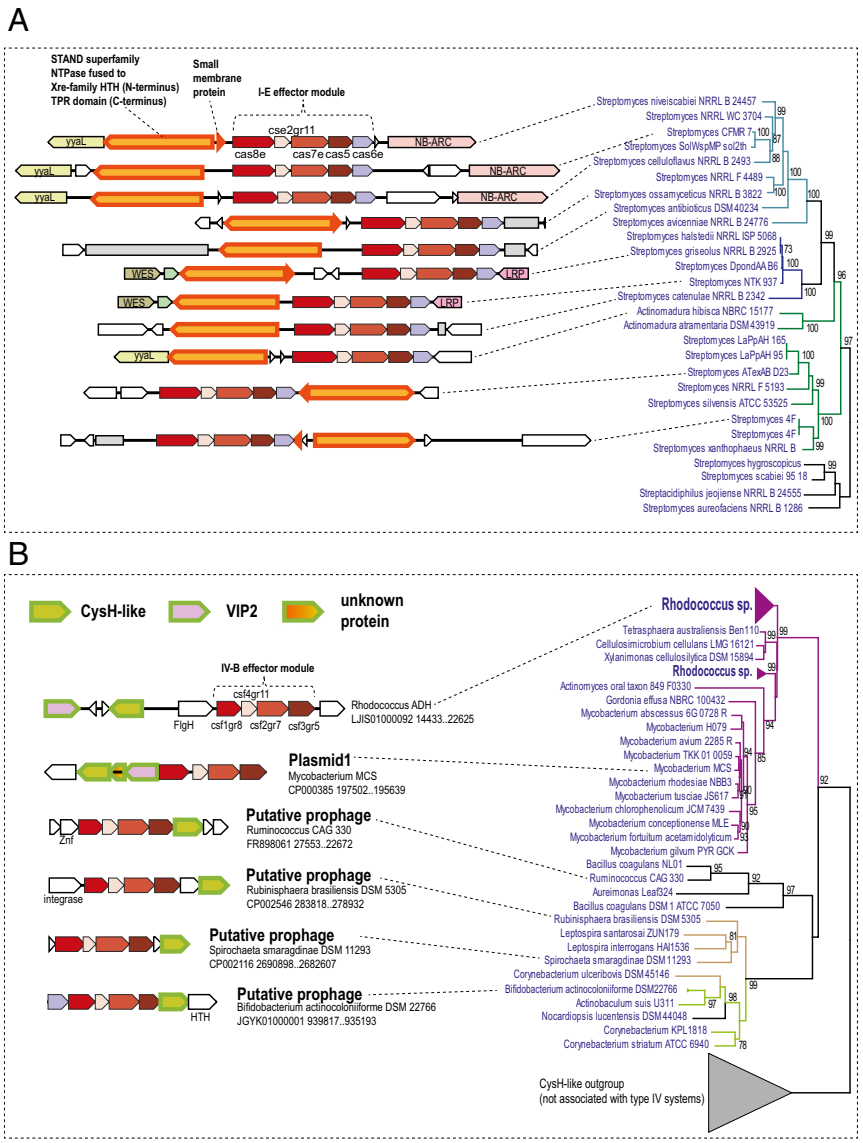

Fig. 4. Locus organization of type I-E and type IV CRISPR-Cas systems containing predicted CRISPR-linked genes. (*A*) STAND family NTPases encoded in minimal type I-E loci. The clade of STAND NTPases associated with type I-E systems is shown on the *Right* (complete tree is available at ftp://ftp.ncbi.nlm.nih.gov/pub/wolf/_suppl/CRISPRicity). Genomes in which the STAND NTPases gene is linked to a type I-E locus are shown in blue, and genomes in which there is no such link are shown in black. Colored branches denote three subfamilies (clusters) identified in this work. Support values greater than 70% are indicated for the respective branches. (*B*) CysH family PAPS reductases encoded in type IV-B loci. The clade of CysH family enzymes associated with type IV CRISPR-Cas systems is shown on the *Right* (complete tree is available at ftp://ftp.ncbi.nlm.nih.gov/pub/wolf/_suppl/CRISPRicity). Genomes in which *cysH*-like genes are linked to type IV-B loci are shown in blue, and genomes in which there is no such link are shown in black. Colored branches denote three subfamilies (clusters) identified in this work. Support values greater than 70% are indicated for the respective branches. The designations are as in Fig. 3. CRISPR arrays are shown by gray boxes. Additional abbreviations and gene names: ADP-PRT, ADP phosphoribosyltransferase; FlhG, MinD-like ATPase involved in chromosome partitioning or flagellar assembly; HTH, helix-turn-helix DNA-binding domain; LRP, LRP family transcriptional regulator; N6-MTase, N6 adenosine methylase; NB_ARC, STAND NTPase fused to TPR-repeats (distinct from the predicted CRISPR-linked STAND NTPase); SSB, single-stranded DNA-binding protein.

for *cas10*, whereas none of these genes showed comparative correlation with 16S RNA (Fig. 5 and *SI Appendix, Supporting Information Files 8–10*), suggesting that the genes within the CRISPR-*cas* loci including *corA* coevolve but the loci themselves spread largely via horizontal transfer. In the case of the membrane-associated CARF-domain proteins, highly significant correlation was observed between the trees for these proteins and both Cas10 and 16S RNA (*SI Appendix, Supporting Information Files 7, 9, and 10*), suggesting that the evolution of the respective subset of type III CRISPR-Cas systems, including the accessory proteins predicted here, involved a major vertical component. In contrast, in the case of RT associated with type III systems, the correlations between the RT trees and those for both Cas10 and 16 S RNA were relatively weak (*SI Appendix, Supporting Information Files 6, 9, and 10*), in agreement with the previous conclusion that the RT-containing adaptation modules largely behaved as distinct evolutionary units (57). Thus, comparative analysis of phylogenetic trees highlights distinct patterns of evolution among CRISPR-Cas systems but, on the whole, presents strong evidence of coevolution and implies tight functional association between (at least) the most common of the predicted CRISPR accessory genes and the effector modules of the respective systems.

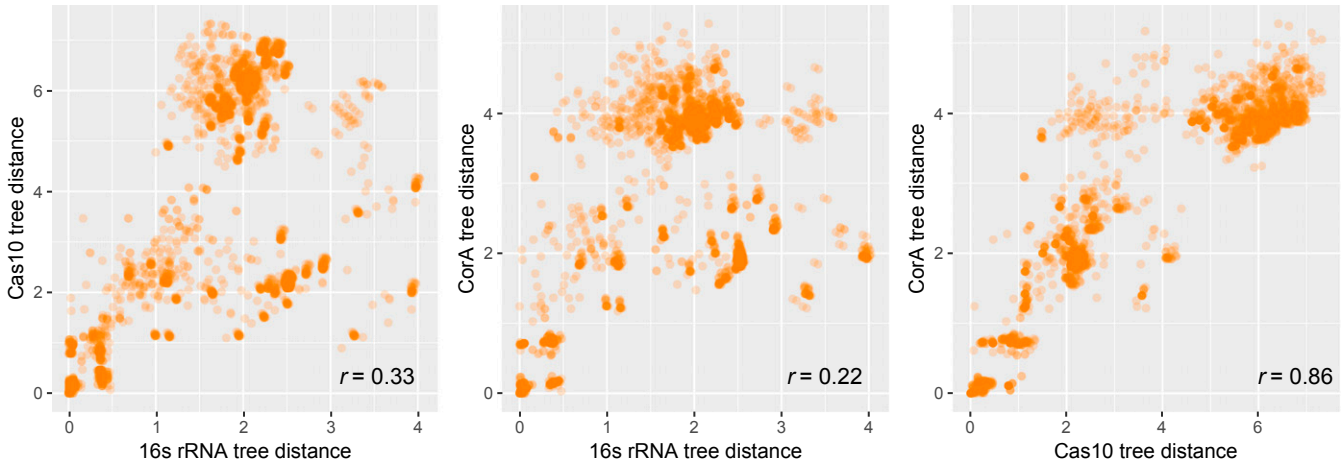

Fig. 5. Coevolution of predicted CRISPR-linked genes with signature *cas* genes. The panels show plots of pairwise distances between predicted the CRISPR-linked *corA* gene product, Cas10 and 16S rRNA estimated from the respective phylogenetic trees. The Spearman rank correlation coefficient is indicated on each plot.

Concluding Remarks

We developed a computational strategy to predict genes that are enriched in the CRISPR-*cas* genomic neighborhood and functionally linked to CRISPR-Cas systems. Exhaustive case-by-case analysis of the detected CRISPR-linked genes shows that, despite the absence of rigorous statistical framework, this CRISPRicity strategy yields sets of genes that are substantially enriched in confident predictions of functional association. Clearly, this approach can be readily generalized beyond CRISPR-Cas, to systematically explore functional links for any other systems encoded in microbial genomes. In biological terms, the CRISPRicity analysis reveals remarkable functional complexity of type III CRISPR-Cas systems that seems to substantially exceed that of other CRISPR-Cas types. A limitation of this approach is that we analyzed only genes that are encoded within or in the vicinity of CRISPR-*cas* loci. As demonstrated by the well-characterized involvement of RNase III in type II pre-crRNA processing (8, 85, 86), proteins encoded elsewhere in microbial genomes can contribute to the CRISPR-Cas function. One approach to address potential contributions of proteins supplied *in trans* could be analysis of co-occurrence patterns of various bacterial and archaeal genes with different CRISPR-Cas types and subtypes (87–89). However, this methodology has its own limitations as illustrated by the same case of RNase III, a protein with a near-ubiquitous presence among bacteria and, accordingly, an uninformative phyletic pattern.

Major themes among the previously unnoticed CRISPR-linked genes identified here are the predicted membrane association and connections to signal transduction pathways. It appears likely that membrane association of CRISPR-Cas systems ensures rapid recognition of viral DNA while it is being injected into the host cell, thus facilitating both adaptation and integration. Indeed, spacer acquisition immediately following phage DNA injection has been demonstrated experimentally, albeit by a type II CRISPR-Cas system that lack components implicated in membrane association (90). It remains to be investigated what biological features of certain CRISPR-Cas systems, in particular, those of type III, might specifically favor the membrane connections as well as apparent links to signal transduction pathways. It might not be a sheer coincidence that many type III systems lack adaptation modules (11): conceivably, type III systems could respond to infection and perhaps other forms of stress in ways different from straightforward adaptive immunity. Taken together, the findings described here imply that entire layers of CRISPR-Cas biology remain unexplored and open up many experimental directions.

Materials and Methods

Prokaryotic Genome Database. A prokaryotic database that consisted of 4,961 completely assembled genomes and 43,599 partial genomes, or 6,342,452 nucleotide sequences altogether (genome partitions, such as chromosomes and plasmids, and contigs), was assembled from archaeal and bacterial genomic sequences downloaded from the National Center for Biotechnology Information (NCBI) FTP database (ftp://ftp.ncbi.nlm.nih.gov/genomes/all/; *SI Appendix*) in March 2016. Default ORF annotation available on the FTP site was used for well annotated genomes (coding density, >0.6 coding sequences per kilobase), and the rest of the genomes were annotated with Meta-GeneMark (91) using the standard model MetaGeneMark_v1.mod (heuristic model for genetic code 11 and GC 30).

CRISPR Array Detection and Annotation. The CRISPR arrays were identified as previously described (92). Briefly, the Prokaryotic Genome Database was scanned with CRISPRFinder (93) and PILER-CR (94) using default parameters. The search identified 61,581 and 49,817 CRISPR arrays, respectively. The union of the search results with the two methods were taken as the set of 65,194 predicted CRISPR arrays; the CRISPRFinder prediction was accepted in cases of overlap. To eliminate spurious CRISPR array predictions, arrays of unknown type that did not produce reliable BLASTN hits (90% identity and 90% coverage) into CRISPR arrays of known type were discarded. This filtering resulted in 42,352 CRISPR arrays that were taken as the final prediction for the subsequent analyses.

Detection and Annotation of CRISPR-Cas Proteins. The translated prokaryotic database was searched with PSI-BLAST (95) using the previously described CRISPR-Cas protein profiles (11, 51, 96) with an e-value cutoff of 10^{-4} and effective database size set to 2×10^7.

Bait Islands. For the purpose of identification of previously undetected CRISPR-linked genes, three groups of "baits" were selected: (*i*) all CRISPR arrays from the final set of predictions; (*ii*) effector modules, that is, all interference-related genes detected using Cas protein profiles; and (*iii*) adaptation modules, that is, *cas1* genes located within 10 kb of a *cas2* and/or *cas4* gene (this additional criterion was adopted because of the existence of non–CRISPR-associated *cas1* homologs).

Bait islands, a data structure describing genomic neighborhoods of the above three classes of baits, were constructed by annotating all ORFs within 10 kb upstream and downstream of the baits. The ORFs were annotated using 30,953 cluster of orthologous groups (COG), pfam, and cd protein profiles from the NCBI CDD database (97) and 217 custom CRISPR-Cas profiles (96) using PSI-BLAST profile search with the same parameters as above.

Construction of Protein Clusters. Clusters of homologous proteins were constructed for all ORFs detected in the bait islands using the following iterative procedure:

i) All proteins were clustered using UCLUST (98) with the sequence similarity threshold of 0.3, yielding permissive cluster set.
ii) For each permissive cluster with three or more proteins, its members were clustered using UCLUST with sequence similarity threshold of 0.9, yielding stringent clusters within the permissive cluster. The number of stringent clusters within a permissive cluster was taken to represent the effective number of sequences in the given permissive cluster.
iii) For each permissive cluster, representatives of all constituent strict clusters were aligned using MAFFT (99). The resulting alignments were used as queries to initiate a PSI-BLAST search against all sequences within the permissive cluster. Sequences that did not produce a significant hit (e-value cutoff of 10^{-4}) were removed. This step was repeated until convergence.
iv) Three iterations of steps *ii* and *iii* were repeated with the updated set of the permissive clusters.
v) At the final clustering step, UCLUST with sequence similarity threshold of 0.5 was employed to cluster all of the remaining singleton sequences.

This two-stage iterative procedure was devised to classify proteins from the CRISPR neighborhoods into a relatively small number of inclusive families of homologs while excluding false positives that accumulate when only permissive clustering is performed. In particular, step *iv* in our procedure eliminates such false positives while retaining distant homologs.

The 16,433 protein clusters with effective size of 3 and greater that were constructed using this procedure were used for further analysis.

Measures of CRISPR-Cas Association. Alignments of the permissive clusters were used to initiate a PSI-BLAST search against the prokaryotic sequence database with an e-value cutoff of 10^{-4} and effective database size set to 2×10^7. All hits covering less than 40% of the query profile were discarded. When multiple queries produced overlapping hits (using the overlap threshold of 25%) to the same target sequence, the corresponding target segment was assigned to the highest-scoring query.

All target segments assigned to the same query were, once again, clustered using UCLUST with the sequence similarity threshold of 0.3. Clusters that did not contain any sequences from the query profiles were removed.

For the sets of hits assigned to a particular query profile, the effective number of sequences was calculated as described above for the following: (*i*) all sequences retrieved from the Prokaryotic Genome Database retrieved by the given profile; (*ii*) sequences from all bait islands; (*iii*) sequences from the CRISPR array islands; (*iv*) sequences from the effector module islands; and (*v*) sequences from the adaptation module islands

The aggregate measure of CRISPR association (CRISPRicity) was calculated for each permissive cluster as the ratio of the effective number of sequences in the bait islands to the effective number of sequences in the entire database that were associated with the given cluster.

Each gene in a specific bait island can be characterized by its genomic distance from the respective bait, that is, the number of genes between the given gene and the bait (genes directly adjacent to the bait and the baits themselves were assigned the distance of 0). The median of the distance to the closest bait across all representatives was used to characterize the permissive cluster.

Selection of Candidate CRISPR-Linked Protein Clusters. The CRISPRicity-abundance-distance space, embedding all permissive clusters, was partitioned into 1,000 voxels (volume elements) as follows. The CRISPRicity range (from 0 to 1) was split into 10 equal intervals. The abundance (the effective number of sequences in bait islands) range was split into 10 intervals in log space with a step of 0.3 decimal log units (factor of ~2.0), starting from 1; all clusters with abundance of $502 = 10^{0.3 \times 9}$ or greater were assigned to the last interval. The distance range was split into 10 intervals with a step of 1, starting from the distance of 0 (adjacent to the bait); clusters with distances of 9 and above were assigned to the last interval. This $10 \times 10 \times 10$ grid formed the 1,000 voxels. Within each voxel, the known Cas (together with Cas-associated) clusters and non-Cas clusters were counted; their probability masses (D_c and D_n, respectively) were calculated as the counts divided by the total number of such genes. Voxels with the CRISPR index (ratios of densities) $I = D_c/D_n > 2$ were selected for further analysis.

Phylogenetic Analysis. The 16S rRNA tree was constructed for all organisms from the prokaryotic database where both 16S SSU rRNA and CRISPR-Cas type III system were found. The 16S rRNA genes were identified using BLASTN (95) search with the *Pyrococcus* sp. NA2 16S rRNA as the query for archaeal genomes and the *Escherichia coli* K-12 ER3413 16S rRNA as a query for bacterial genomes (word size of 8 and dust filtering off). The best scoring BLASTN hits with at least 80% coverage and 70% identity were taken for each organism, 16S rRNA sequences were aligned using MAFFT (99), and a phylogenetic tree was constructed using FastTree (100) with gamma-distributed site rates and GTR evolutionary model. Protein sequences were iteratively aligned using hhalign (101) starting from MUSCLE (102) alignments of UCLUST clusters (similarity cutoff of 0.5). Approximate maximum-likelihood trees were built from these alignments using FastTree with gamma-distributed site rates and WAG evolutionary model.

Case-by-Case Analysis and Annotation of Permissive Clusters of Putative CRISPR-Linked Proteins. PSI-BLAST searches with COG, pfam, and cd protein profiles from NCBI CDD database (97) and with the custom CRISPR-Cas profiles (96) were run against a database made from consensus sequences (103) of the permissive clusters with an e-value cutoff of 10^{-4} and effective database size set to 2×10^7.

Iterative profile searches using PSI-BLAST (95), with a cutoff e-value of 0.01, and composition based-statistics and low complexity filtering turned off, were used to search for distantly similar sequences in NCBI's nonredundant (NR) database. Another sensitive method for remote sequence similarity detection, HHpred, was used with default parameters (101). Additionally, clusters were annotated using HHSearch (101) comparison between cluster-derived HMM profiles and CDD-derived HMM profiles. The results with an HHSearch probability score greater than 80% were recorded.

Protein secondary structure was predicted using Jpred (104). Transmembrane segments were predicted using the TMMHMM, version 2.0c, program with default parameters (105).

A fraction of the permissive protein clusters was found to comprise CRISPR arrays falsely annotated as protein-coding sequences. Clusters containing more than 10% of sequences overlapping with known CRISPR arrays or matching CRISPR repeats with 90% identity and 90% coverage in BLASTN search were identified and discarded from further analysis.

ACKNOWLEDGMENTS. This research was supported by the Intramural Funds of the US Department of Health and Human Services (E.V.K.), the Ministry of Education and Science of the Russian Federation Subsidy Agreement 14.606.21.0006 (Project identifier RFMEFI60617X0006; to S.A.S. and K.V.S.), and an NIH R01 Grant GM10407 (to K.V.S.).

1. Sorek R, Lawrence CM, Wiedenheft B (2013) CRISPR-mediated adaptive immune systems in bacteria and archaea. *Annu Rev Biochem* 82:237–266.
2. Wright AV, Nuñez JK, Doudna JA (2016) Biology and applications of CRISPR systems: Harnessing nature's toolbox for genome engineering. *Cell* 164:29–44.
3. Komor AC, Badran AH, Liu DR (2017) CRISPR-based technologies for the manipulation of eukaryotic genomes. *Cell* 168:20–36.
4. Mohanraju P, et al. (2016) Diverse evolutionary roots and mechanistic variations of the CRISPR-Cas systems. *Science* 353:aad5147.
5. Barrangou R, Horvath P (2017) A decade of discovery: CRISPR functions and applications. *Nat Microbiol* 2:17092.
6. Amitai G, Sorek R (2016) CRISPR-Cas adaptation: Insights into the mechanism of action. *Nat Rev Microbiol* 14:67–76.
7. Sternberg SH, Richter H, Charpentier E, Qimron U (2016) Adaptation in CRISPR-Cas systems. *Mol Cell* 61:797–808.
8. Charpentier E, Richter H, van der Oost J, White MF (2015) Biogenesis pathways of RNA guides in archaeal and bacterial CRISPR-Cas adaptive immunity. *FEMS Microbiol Rev* 39:428–441.
9. Plagens A, Richter H, Charpentier E, Randau L (2015) DNA and RNA interference mechanisms by CRISPR-Cas surveillance complexes. *FEMS Microbiol Rev* 39: 442–463.
10. Nishimasu H, Nureki O (2017) Structures and mechanisms of CRISPR RNA-guided effector nucleases. *Curr Opin Struct Biol* 43:68–78.
11. Makarova KS, et al. (2015) An updated evolutionary classification of CRISPR-Cas systems. *Nat Rev Microbiol* 13:722–736.
12. Koonin EV, Makarova KS, Zhang F (2017) Diversity, classification and evolution of CRISPR-Cas systems. *Curr Opin Microbiol* 37:67–78.
13. Doudna JA, Charpentier E (2014) Genome editing. The new frontier of genome engineering with CRISPR-Cas9. *Science* 346:1258096.
14. van der Oost J, Westra ER, Jackson RN, Wiedenheft B (2014) Unravelling the structural and mechanistic basis of CRISPR-Cas systems. *Nat Rev Microbiol* 12:479–492.
15. Jiang F, Doudna JA (2015) The structural biology of CRISPR-Cas systems. *Curr Opin Struct Biol* 30:100–111.
16. Wiedenheft B, et al. (2009) Structural basis for DNase activity of a conserved protein implicated in CRISPR-mediated genome defense. *Structure* 17:904–912.

17. Nuñez JK, et al. (2014) Cas1-Cas2 complex formation mediates spacer acquisition during CRISPR-Cas adaptive immunity. *Nat Struct Mol Biol* 21:528–534.
18. Nuñez JK, Lee AS, Engelman A, Doudna JA (2015) Integrase-mediated spacer acquisition during CRISPR-Cas adaptive immunity. *Nature* 519:193–198.
19. Rollie C, Schneider S, Brinkmann AS, Bolt EL, White MF (2015) Intrinsic sequence specificity of the Cas1 integrase directs new spacer acquisition. *eLife* 4:e08716.
20. Jackson SA, et al. (2017) CRISPR-Cas: Adapting to change. *Science* 356:eaal5056.
21. Wright AV, et al. (2017) Structures of the CRISPR genome integration complex. *Science* 357:1113–1118.
22. Xiao Y, Ng S, Nam KH, Ke A (2017) How type II CRISPR-Cas establish immunity through Cas1-Cas2-mediated spacer integration. *Nature* 550:137–141.
23. Sinkunas T, et al. (2011) Cas3 is a single-stranded DNA nuclease and ATP-dependent helicase in the CRISPR/Cas immune system. *EMBO J* 30:1335–1342.
24. Beloglazova N, et al. (2011) Structure and activity of the Cas3 HD nuclease MJ0384, an effector enzyme of the CRISPR interference. *EMBO J* 30:4616–4627.
25. Sinkunas T, et al. (2013) In vitro reconstitution of Cascade-mediated CRISPR immunity in *Streptococcus thermophilus*. *EMBO J* 32:385–394.
26. Hochstrasser ML, et al. (2014) CasA mediates Cas3-catalyzed target degradation during CRISPR RNA-guided interference. *Proc Natl Acad Sci USA* 111:6618–6623.
27. Künne T, et al. (2016) Cas3-derived target DNA degradation fragments fuel primed CRISPR adaptation. *Mol Cell* 63:852–864.
28. Zhang J, Kasciukovic T, White MF (2012) The CRISPR associated protein Cas4 Is a 5′ to 3′ DNA exonuclease with an iron-sulfur cluster. *PLoS One* 7:e47232.
29. Lemak S, et al. (2014) The CRISPR-associated Cas4 protein Pcal_0546 from *Pyrobaculum calidifontis* contains a [2Fe-2S] cluster: Crystal structure and nuclease activity. *Nucleic Acids Res* 42:11144–11155.
30. Hudaiberdiev S, et al. (2017) Phylogenomics of Cas4 family nucleases. *BMC Evol Biol* 17:232.
31. Makarova KS, Grishin NV, Shabalina SA, Wolf YI, Koonin EV (2006) A putative RNA-interference-based immune system in prokaryotes: Computational analysis of the predicted enzymatic machinery, functional analogies with eukaryotic RNAi, and hypothetical mechanisms of action. *Biol Direct* 1:7.
32. Brouns SJ, et al. (2008) Small CRISPR RNAs guide antiviral defense in prokaryotes. *Science* 321:960–964.
33. Rouillon C, et al. (2013) Structure of the CRISPR interference complex CSM reveals key similarities with cascade. *Mol Cell* 52:124–134.
34. Spilman M, et al. (2013) Structure of an RNA silencing complex of the CRISPR-Cas immune system. *Mol Cell* 52:146–152.
35. Staals RHJ, et al. (2013) Structure and activity of the RNA-targeting type III-B CRISPR-Cas complex of *Thermus thermophilus*. *Mol Cell* 52:135–145.
36. Tamulaitis G, et al. (2014) Programmable RNA shredding by the type III-A CRISPR-Cas system of *Streptococcus thermophilus*. *Mol Cell* 56:506–517.
37. Beloglazova N, et al. (2015) CRISPR RNA binding and DNA target recognition by purified cascade complexes from *Escherichia coli*. *Nucleic Acids Res* 43:530–543.
38. Hochstrasser ML, Taylor DW, Kornfeld JE, Nogales E, Doudna JA (2016) DNA targeting by a minimal CRISPR RNA-guided cascade. *Mol Cell* 63:840–851.
39. Cass SD, et al. (2015) The role of Cas8 in type I CRISPR interference. *Biosci Rep* 35: e00197.
40. Chylinski K, Le Rhun A, Charpentier E (2013) The tracrRNA and Cas9 families of type II CRISPR-Cas immunity systems. *RNA Biol* 10:726–737.
41. Chylinski K, Makarova KS, Charpentier E, Koonin EV (2014) Classification and evolution of type II CRISPR-Cas systems. *Nucleic Acids Res* 42:6091–6105.
42. Jiang F, Doudna JA (2017) CRISPR-Cas9 structures and mechanisms. *Annu Rev Biophys* 46:505–529.
43. Jiang F, et al. (2016) Structures of a CRISPR-Cas9 R-loop complex primed for DNA cleavage. *Science* 351:867–871.
44. Makarova KS, Aravind L, Grishin NV, Rogozin IB, Koonin EV (2002) A DNA repair system specific for thermophilic Archaea and bacteria predicted by genomic context analysis. *Nucleic Acids Res* 30:482–496.
45. Zhu X, Ye K (2012) Crystal structure of Cmr2 suggests a nucleotide cyclase-related enzyme in type III CRISPR-Cas systems. *FEBS Lett* 586:939–945.
46. Jung TY, et al. (2015) Crystal structure of the Csm1 subunit of the Csm complex and its single-stranded DNA-specific nuclease activity. *Structure* 23:782–790.
47. Kazlauskiene M, Kostiuk G, Venclovas Č, Tamulaitis G, Siksnys V (2017) A cyclic oligonucleotide signaling pathway in type III CRISPR-Cas systems. *Science* 357: 605–609.
48. Niewoehner O, et al. (2017) Type III CRISPR-Cas systems produce cyclic oligoadenylate second messengers. *Nature* 548:543–548.
49. Daume M, Plagens A, Randau L (2014) DNA binding properties of the small cascade subunit Csa5. *PLoS One* 9:e105716.
50. Shmakov S, et al. (2015) Discovery and functional characterization of diverse class 2 CRISPR-Cas systems. *Mol Cell* 60:385–397.
51. Shmakov S, et al. (2017) Diversity and evolution of class 2 CRISPR-Cas systems. *Nat Rev Microbiol* 15:169–182.
52. Smargon AA, et al. (2017) Cas13b is a type VI-B CRISPR-associated RNA-guided RNase differentially regulated by accessory proteins Csx27 and Csx28. *Mol Cell* 65:618–630.e7.
53. Abudayyeh OO, et al. (2016) C2c2 is a single-component programmable RNA-guided RNA-targeting CRISPR effector. *Science* 353:aaf5573.
54. Yan WX, et al. (2018) Cas13d is a compact RNA-targeting type VI CRISPR effector positively modulated by a WYL-domain-containing accessory protein. *Mol Cell* 70: 327–339.e5.
55. Konermann S, et al. (2018) Transcriptome engineering with RNA-targeting type VI-D CRISPR effectors. *Cell* 173:665–676.e14.
56. Makarova KS, Wolf YI, Koonin EV (2013) The basic building blocks and evolution of CRISPR-CAS systems. *Biochem Soc Trans* 41:1392–1400.
57. Silas S, et al. (2017) On the origin of reverse transcriptase-using CRISPR-Cas systems and their hyperdiverse, enigmatic spacer repertoires. *MBio* 8:e00897-17.
58. Silas S, et al. (2016) Direct CRISPR spacer acquisition from RNA by a natural reverse transcriptase-Cas1 fusion protein. *Science* 351:aad4234.
59. Anantharaman V, Makarova KS, Burroughs AM, Koonin EV, Aravind L (2013) Comprehensive analysis of the HEPN superfamily: Identification of novel roles in intra-genomic conflicts, defense, pathogenesis and RNA processing. *Biol Direct* 8:15.
60. Makarova KS, Wolf YI, Snir S, Koonin EV (2011) Defense islands in bacterial and archaeal genomes and prediction of novel defense systems. *J Bacteriol* 193: 6039–6056.
61. Jacob F, Monod J (1961) Genetic regulatory mechanisms in the synthesis of proteins. *J Mol Biol* 3:318–356.
62. Wolf YI, Rogozin IB, Kondrashov AS, Koonin EV (2001) Genome alignment, evolution of prokaryotic genome organization, and prediction of gene function using genomic context. *Genome Res* 11:356–372.
63. Beckwith J (2011) The operon as paradigm: Normal science and the beginning of biological complexity. *J Mol Biol* 409:7–13.
64. Rogozin IB, et al. (2002) Connected gene neighborhoods in prokaryotic genomes. *Nucleic Acids Res* 30:2212–2223.
65. Shah SA, et al. (2018) Conserved accessory proteins encoded with archaeal and bacterial type III CRISPR-Cas gene cassettes that may specifically modulate, complement or extend interference activity. bioRxiv:262675.
66. Swarts DC, et al. (2014) The evolutionary journey of Argonaute proteins. *Nat Struct Mol Biol* 21:743–753.
67. Kaya E, et al. (2016) A bacterial Argonaute with noncanonical guide RNA specificity. *Proc Natl Acad Sci USA* 113:4057–4062.
68. Peters JE, Makarova KS, Shmakov S, Koonin EV (2017) Recruitment of CRISPR-Cas systems by Tn7-like transposons. *Proc Natl Acad Sci USA* 114:E7358–E7366.
69. Maguire ME (2006) The structure of CorA: A Mg^{2+}-selective channel. *Curr Opin Struct Biol* 16:432–438.
70. Vestergaard G, Garrett RA, Shah SA (2014) CRISPR adaptive immune systems of Archaea. *RNA Biol* 11:156–167.
71. Puigbò P, Makarova KS, Kristensen DM, Wolf YI, Koonin EV (2017) Reconstruction of the evolution of microbial defense systems. *BMC Evol Biol* 17:94.
72. Aravind L, Koonin EV (1998) A novel family of predicted phosphoesterases includes *Drosophila* prune protein and bacterial RecJ exonuclease. *Trends Biochem Sci* 23: 17–19.
73. Anantharaman V, Aravind L (2006) The NYN domains: Novel predicted RNAses with a PIN domain-like fold. *RNA Biol* 3:18–27.
74. Matthies D, et al. (2016) Cryo-EM structures of the magnesium channel CorA reveal symmetry break upon gating. *Cell* 164:747–756.
75. Lerche M, Sandhu H, Flöckner L, Högbom M, Rapp M (2017) Structure and cooperativity of the cytosolic domain of the CorA Mg^{2+} channel from *Escherichia coli*. *Structure* 25:1175–1186.e4.
76. Makarova KS, Anantharaman V, Grishin NV, Koonin EV, Aravind L (2014) CARF and WYL domains: Ligand-binding regulators of prokaryotic defense systems. *Front Genet* 5:102.
77. Smith CK, Baker TA, Sauer RT (1999) Lon and Clp family proteases and chaperones share homologous substrate-recognition domains. *Proc Natl Acad Sci USA* 96: 6678–6682.
78. Burroughs AM, Zhang D, Schäffer DE, Iyer LM, Aravind L (2015) Comparative genomic analyses reveal a vast, novel network of nucleotide-centric systems in biological conflicts, immunity and signaling. *Nucleic Acids Res* 43: 10633–10654.
79. Perez-Rodriguez R, et al. (2011) Envelope stress is a trigger of CRISPR RNA-mediated DNA silencing in *Escherichia coli*. *Mol Microbiol* 79:584–599.
80. Sampson TR, et al. (2014) A CRISPR-Cas system enhances envelope integrity mediating antibiotic resistance and inflammasome evasion. *Proc Natl Acad Sci USA* 111: 11163–11168.
81. Leipe DD, Koonin EV, Aravind L (2004) STAND, a class of P-loop NTPases including animal and plant regulators of programmed cell death: Multiple, complex domain architectures, unusual phyletic patterns, and evolution by horizontal gene transfer. *J Mol Biol* 343:1–28.
82. Koonin EV, Aravind L (2002) Origin and evolution of eukaryotic apoptosis: The bacterial connection. *Cell Death Differ* 9:394–404.
83. Bick JA, Dennis JJ, Zylstra GJ, Nowack J, Leustek T (2000) Identification of a new class of 5′-adenylylsulfate (APS) reductases from sulfate-assimilating bacteria. *J Bacteriol* 182:135–142.
84. Jeong BR, et al. (2011) Structure function analysis of an ADP-ribosyltransferase type III effector and its RNA-binding target in plant immunity. *J Biol Chem* 286: 43272–43281.
85. Deltcheva E, et al. (2011) CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. *Nature* 471:602–607.
86. Court DL, et al. (2013) RNase III: Genetics and function; structure and mechanism. *Annu Rev Genet* 47:405–431.
87. Tatusov RL, Koonin EV, Lipman DJ (1997) A genomic perspective on protein families. *Science* 278:631–637.
88. Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO (1999) Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles. *Proc Natl Acad Sci USA* 96:4285–4288.
89. Galperin MY, Koonin EV (2000) Who's your neighbor? New computational approaches for functional genomics. *Nat Biotechnol* 18:609–613.

90. Modell JW, Jiang W, Marraffini LA (2017) CRISPR-Cas systems exploit viral DNA injection to establish and maintain adaptive immunity. *Nature* 544:101–104.
91. Besemer J, Lomsadze A, Borodovsky M (2001) GeneMarkS: A self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucleic Acids Res* 29:2607–2618.
92. Shmakov SA, et al. (2017) The CRISPR spacer space is dominated by sequences from species-specific mobilomes. *MBio* 8:e01397-17.
93. Grissa I, Vergnaud G, Pourcel C (2007) CRISPRFinder: A web tool to identify clustered regularly interspaced short palindromic repeats. *Nucleic Acids Res* 35: W52–W57.
94. Edgar RC (2007) PILER-CR: Fast and accurate identification of CRISPR repeats. *BMC Bioinformatics* 8:18.
95. Altschul SF, et al. (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res* 25:3389–3402.
96. Makarova KS, Koonin EV (2015) Annotation and classification of CRISPR-Cas systems. *Methods Mol Biol* 1311:47–75.
97. Marchler-Bauer A, et al. (2015) CDD: NCBI's conserved domain database. *Nucleic Acids Res* 43:D222–D226.
98. Edgar RC (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26:2460–2461.
99. Katoh K, Standley DM (2013) MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol Biol Evol* 30:772–780.
100. Price MN, Dehal PS, Arkin AP (2010) FastTree 2—approximately maximum-likelihood trees for large alignments. *PLoS One* 5:e9490.
101. Söding J (2005) Protein homology detection by HMM-HMM comparison. *Bioinformatics* 21:951–960.
102. Edgar RC (2004) MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32:1792–1797.
103. Yutin N, Makarova KS, Mekhedov SL, Wolf YI, Koonin EV (2008) The deep archaeal roots of eukaryotes. *Mol Biol Evol* 25:1619–1630.
104. Drozdetskiy A, Cole C, Procter J, Barton GJ (2015) JPred4: A protein secondary structure prediction server. *Nucleic Acids Res* 43:W389–W394.
105. Krogh A, Larsson B, von Heijne G, Sonnhammer EL (2001) Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes. *J Mol Biol* 305:567–580.