Applied and Environmental Microbiology, Apr. 2010, p. 2136–2144 Vol. 76, No. 7
0099-2240/10/$12.00 doi:10.1128/AEM.01985-09

Evolutionary Dynamics of Clustered Irregularly Interspaced Short Palindromic Repeat Systems in the Ocean Metagenome▽

Valery A. Sorokin,[2] Mikhail S. Gelfand,[2,3] and Irena I. Artamonova[1,3]*

N. I. Vavilov Institute of General Genetics, Russian Academy of Sciences, ul. Gubkina 3, Moscow 119991,[1] Faculty of Bioengineering and Bioinformatics, M. V. Lomonosov Moscow State University, Vorobievy Gory 1-73, Moscow 119992,[2] and A. A. Kharkevich Institute for Information Transmission Problems, Russian Academy of Sciences, Bolshoi Karetny Pereulok 19, Moscow 127994,[3] Russia

Received 18 August 2009/Accepted 25 January 2010

Clustered regularly interspaced short palindromic repeats (CRISPRs) form a recently characterized type of prokaryotic antiphage defense system. The phage-host interactions involving CRISPRs have been studied in experiments with selected bacterial or archaeal species and, computationally, in completely sequenced genomes. However, these studies do not allow one to take prokaryotic population diversity and phage-host interaction dynamics into account. This gap can be filled by using metagenomic data: in particular, the largest existing data set, generated from the *Sorcerer II* Global Ocean Sampling expedition. The application of three publicly available CRISPR recognition programs to the Global Ocean metagenome produced a large proportion of false-positive results. To address this problem, a filtering procedure was designed. It resulted in about 200 reliable CRISPR cassettes, which were then studied in detail. The repeat consensuses were clustered into several stable classes that differed from the existing classification. Short fragments of DNA similar to the cassette spacers were more frequently present in the same geographical location than in other locations (P, <0.0001). We developed a catalogue of elementary CRISPR-forming events and reconstructed the likely evolutionary history of cassettes that had common spacers. Metagenomic collections allow for relatively unbiased analysis of phage-host interactions and CRISPR evolution. The results of this study demonstrate that CRISPR cassettes retain the memory of the local virus population at a particular ocean location. CRISPR evolution may be described using a limited vocabulary of elementary events that have a natural biological interpretation.

Prokaryotes are highly diverse (33). One of the explanations of this diversity is the high extinction rate, due to genetic aggression, which leads to the clearance of ecological niches and, as a result, may allow new prokaryotic species to emerge. In the absence of host defense, viral infection of prokaryotic colonies results in colony extinction or the fixation of a fraction of the invader's genetic material in the host genome, profoundly affecting the life cycle of the host (32). Thus, bacteria and archaea have developed various kinds of defense mechanisms to resist this pressure; the best studied of these mechanisms is restriction-modification systems (4).

Along with well-known prokaryotic defense mechanisms, such as rapid evolution of cell receptors or the use of restriction-modification or toxin-antitoxin systems (see, e.g., references 6, 21, and 25), newly discovered clustered regularly interspaced palindromic repeat (CRISPR) systems seem to play an important role in protecting the cell from archaeal virus or bacteriophage assaults (reviewed in reference 36). A typical CRISPR system is a genetic locus comprising CRISPR-associated (*cas*) genes coding for proteins of several distinct functional classes (8, 19, 29) and a CRISPR cassette. A CRISPR cassette is formed by almost identical direct repeats with an average length of 32 nucleotides (nt), which are separated by similarly sized, unique spacers. A considerable proportion of spacers is similar to known phage or virus sequences, suggesting that the system is involved in antivirus defense (8, 29, 31). This involvement was experimentally demonstrated when a CRISPR system was shown to be essential for cell survival after invasion by foreign DNA (5). The mechanism is thought to be analogous to eukaryotic RNA interference (29), but it has not been characterized in detail yet.

CRISPR cassettes retain information that could be used to reveal the evolutionary history of individual systems. First, it has been shown that CRISPR-associated genes could be divided into eight subtypes according to operon organization and gene phylogeny (19). Second, the repeats of different CRISPR cassettes may be similar, which might indicate a common origin of such cassettes. The first attempt to cluster CRISPR cassettes by the similarity of repeat sequences resulted in 12 clusters (27). In that study, the cassettes were obtained by the application of PILER-CR to completely sequenced genomes. Third, pairwise comparison of spacers could also reveal the specific evolutionary history of individual CRISPR cassettes.

So far, most large-scale studies of CRISPR systems have been restricted to well-studied organisms with completely sequenced genomes (5, 9, 20, 28, 30). However, the dynamic interaction between viruses or phages and microorganisms in natural environments is of particular interest (2, 10, 15, 23, 35, 38, 40–42). It may be studied using CRISPRs in a metagenome, that is, sequenced DNA fragments collected in one geographical location and therefore representing one ecological niche with all its inhabitants. This approach is interesting

* Corresponding author. Mailing address: Bioinformatics Group, Vavilov Institute of General Genetics RAS, Gubkina 3, Moscow 119991, Russia. Phone: 7 916 9155809. Fax: 7 499 1328962. E-mail: irenart@vigg.ru.

▽ Published ahead of print on 29 January 2010.

for two reasons. First, metagenomic samples provide a common census of coexisting organisms, i.e., in many cases, both the infecting viruses and phages and their victims. Second, most bacteria and archaea from metagenomic samples cannot be cultivated, and hence little is known about their CRISPR systems.

To date, three studies have considered host-virus interactions in metagenomes. One study used two thermophilic *Synechococcus* isolates from microbial mats in hot springs at Yellowstone National Park to demonstrate fast coevolution of the host and phage genomes (22). Two studies described archaeal and bacterial interactions with viruses and phages, respectively, in acidophilic biofilms (2, 39). All environmental communities analyzed so far are extreme and are dominated by few species. Natural samples containing many diverse coexisting organisms may arguably be more interesting.

The largest available metagenome, produced by the *Sorcerer II* Global Ocean Sampling (GOS) expedition, comprises samples of genetic material collected from more than 50 geographical locations of the Pacific and Atlantic oceans (34). This variety provides an opportunity to study the evolution of phage-host interactions reflected in CRISPRs.

Three algorithms, PILER-CR (14), the CRISPR recognition tool (CRT) (7), and CRISPRFinder (18), have been developed as tools for the discovery of new CRISPR cassettes. All these algorithms define candidate CRISPR cassette sequences as short direct repeats separated by short unique spacers; they then use a variety of standard repeat-finding techniques. However, the implementation of specific details is different.

PILER-CR constructs local alignments of the input sequence to itself; each hit between two close regions is a candidate for an alignment of a repeat with its neighbor copy. In terms of dynamic programming, taking into account the repeat structure of a CRISPR cassette implies looking for hits only within a relatively narrow band around the main diagonal of the dot plot. This process is followed by several refinement steps.

CRT does not use alignments to identify candidate repeats; rather, it derives them directly from the analysis of an input sequence. It is based on finding series of short repeats of a specified length (searching for exact k-mer matches) and then extending these repeats (increasing k-mer length) while allowing for a certain level of mismatches.

Finally, CRISPRFinder is based on a suffix-tree-based algorithm for repeat discovery, again with additional refinement.

All three algorithms were used for the CRISPR cassette search in this study.

MATERIALS AND METHODS

The GOS data set was downloaded as "combined assemblies" in 3,081,849 contigs with a total length of 4.46 Gbp from the CAMERA website (http://web.camera.calit2.net/cameraweb/gwt/org.jcvi.camera.web.gwt.download.DownloadByPubPage/DownloadByPubPage.oa, file gos_combined_scaffolds_fasta.gz) in June 2007. These data were collected at 57 different locations of the global ocean.

CRT and PILER-CR were downloaded from the respective websites and were applied to the metagenome with the default parameters. The third procedure, CRISPRFinder, is available only as a Web service (http://crispr.u-psud.fr/Server/CRISPRfinder.php/) (18). Therefore, we designed a script that placed metagenomic contigs in the input form of the Web service (default parameters), collected outputs, parsed them, and reported identified CRISPR cassettes.

Each of these three algorithms is controlled by a set of parameters, such as the maximum and minimum repeat length, the maximum and minimum spacer length, and the maximum number of mismatches between the repeat copies. Although the default parameters are different for every program, we used the default settings rather than a common, standardized set of parameters. The reasoning behind this decision is that the variations in the approaches to repeat finding might require different settings to bring the programs to their top performance, and we assumed that the programs' authors used the best settings as the default options.

To identify *cas* genes, contigs containing predicted CRISPR cassettes were subjected to a BLASTX search (http://blast.ncbi.nlm.nih.gov/Blast.cgi) (1) against the nonredundant protein collection (nr) of GenBank (e-value threshold, 0.01). The output was parsed, and hits with description fields containing the words "cas" and "crispr" were collected for further manual analysis.

To construct the repeat clusters, the standard BLASTCLUST procedure (http://blast.ncbi.nlm.nih.gov/Blast.cgi) (1) was applied to the specified set of repeat consensus sequences (parameters: -L 0.5 -S 50 -e F -p F -W 15 [as explained in the reference site http://www.ncbi.nih.gov/IEB/ToolBox/C_DOC/lxr/source/doc/blast/blastclust.html]). The output was parsed, and all entries with more than two members were collected. We applied the standard MUSCLE procedure (13) to construct and visualize alignments for a given set of clusters (default parameters; reverse strand manually selected if necessary).

Flanks of CRISPR cassettes were compared with each other using standard BLASTN (http://blast.ncbi.nlm.nih.gov/Blast.cgi) (1) with an e-value threshold of 0.01. Flanks of 500 nt were considered when available. At contig termini, shorter flanks of the maximum possible length, but not less than 50 nt, were analyzed.

The spacers were subjected to the standard BLASTN search (e-value threshold, 0.01) against the viral, bacterial, prokaryotic, and eukaryotic subsections of GenBank and against the GOS data set. To estimate the significance of the observed frequency of similarities between spacers from a given set and a sequence database, we generated 10 randomized sets of "pseudospacers," where each spacer was replaced by a random fragment of the same length from the same contig. The same procedure was used for the search against marine viromes. The virome sequences were downloaded from the CAMERA web site (http://web.camera.calit2.net/cameraweb/gwt/org.jcvi.camera.web.gwt.download.DownloadByPubPage/DownloadByPubPage.oa?projectSymbol=CAM_PROJ_MarineVirome, files Arctic_fasta.gz, BBC_fasta.gz, GOM_fasta.gz, and SAR_fasta.gz) in March 2008.

The significance of the correlation between the origin (sample) of a spacer and metagenomic similarities to this spacer was estimated as follows. Each (proto)spacer was compared to all reads constituting the corresponding contig, and only reads including the complete spacer sequence were retained. If no such reads were identified, we considered reads identical to the (proto)spacer at the ends. The geographical locations of the GOS samples corresponding to the (proto)spacer reads formed the (proto)spacer "sample list".

A spacer-protospacer match was scored whenever two members of a pair had at least one sample (location) in common. The significance was estimated using a procedure that shuffled locations among protospacer sample lists. The occurrence rates for each sample were estimated from the set of protospacer sample lists. We then generated 10,000 random assignments of samples to the protospacers using the estimated occurrence rates and keeping the number of samples assigned to each protospacer fixed. For each shuffled data set, the number of sample coincidences between spacers and their protospacers was calculated.

Repeat clusters obtained by Kunin et al. (27) and the clusters constructed here were compared as follows. Profile hidden Markov models (HMM) were constructed for 12 clusters from reference 27 with HMMBUILD default parameters (http://hmmer.janelia.org/). Then we applied a calibration procedure to adjust the expected values (HMMCALIBRATE default parameters). Further, we estimated the e-value threshold by matching cluster members against 12 HMM profiles, while gradually strengthening the e-value threshold until only cluster members matched the cluster profile. Having estimated the threshold, we compared members of the metagenomic clusters to 12 profiles given the estimated threshold (HMMSEARCH parameters: -E 1e−07 -Z 1).

To determine the contig taxonomy, contigs were subjected to a BLASTX search against the nonredundant protein collection (nr) of GenBank (e-value threshold, 0.01). Taxonomy was assigned manually by analysis of the similarities obtained. In a nutshell, this analysis relied on the similarity levels and the taxonomical origins of the top hits, and specifically on the consistency of their taxonomy. More exactly, if all top hits were on about the same similarity level and belonged to one specific taxon, this taxon was assigned to the contig. If, on the contrary, the top hits demonstrated considerable diversity, only nonspecific taxa were assigned.

For the confirmation of phage origin, contigs matching the cassette spacers, but not their repeats, were subjected to a BLASTX search against the viral subsection of the nonredundant protein collection of GenBank (e-value thresh-

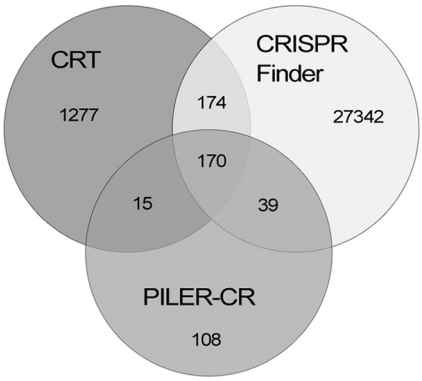


FIG. 1. Venn diagram of the numbers of CRISPR cassettes identified by the three programs.

old, 1e−10). A phage origin was assigned to contigs that had at least one hit to a phage protein.

In order to exclude the influence of assembly artifacts, we validated the elementary classes of evolutionary events at the read level, so that all spacer sequences were subjected to a BLASTN search against the set of metagenome reads. An elementary class of events was confirmed if both spacer combinations forming the event were observed in single reads for at least one event of the class.

RESULTS AND DISCUSSION

Compilation of the initial set of CRISPR cassettes. Three algorithms, CRT (7), PILER-CR (14), and CRISPRFinder (18), were used to identify CRISPR cassettes in the metagenome. This resulted in three poorly compatible sets of candidate CRISPR cassettes. The number of cassettes identified was 331 by PILER-CR, 1,636 by CRT, and 27,782 by CRISPRFinder. The subsets of cassettes identified by all combinations of these programs are shown in a Venn diagram (Fig. 1). Besides this incompatibility, each program produced a large number of false-positive results caused both by the large volume of the input data and the high degree of data fragmentation. For example, CRT and PILER-CR often treat genomic repeats and low-complexity regions as CRISPR cassettes, while CRISPRFinder reports too many short (repeat-spacer-repeat) "questionable" cassettes.

The metagenome consists of contigs formed by shotgun fragments linked by flank overlaps. A considerable proportion of the contigs includes a short stretch of N′s flanked by similar sequences with lengths of 20 to 40 nucleotides. It turns out that CRISPRFinder treats such regions as cassettes, assuming that flanking similarity indicates a pair of repeat copies and the N run represents a spacer. Such pseudocassettes found by CRISPRFinder accounted for 18,313 cassettes, which were excluded from further consideration.

To offset the remaining problem of numerous false-positive results, we designed a filtering technique, which processed the raw output cassettes and produced a set of the most reliable members. The method is represented schematically in Fig. 2 and is described in detail in the next section.

Prior to constructing the set of reliable CRISPR cassettes, we had to resolve conflicts arising from the fact that two algorithms, or even all three different algorithms, could identify cassettes in the same locus, but with slightly different boundaries.

Manual resolution of several such conflicts demonstrated that the prediction (e.g., boundaries, numbers of structural units) was most accurate in the CRT analysis, followed by

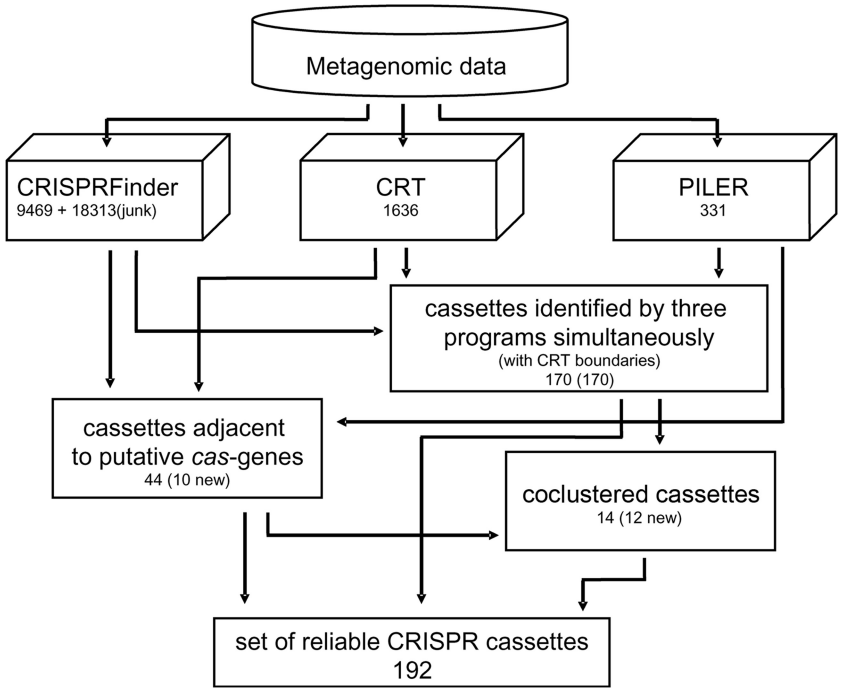


FIG. 2. Schematic representation of the data flow in the procedure developed for the identification of reliable CRISPR cassettes in metagenomic data. The number of CRISPR cassettes identified by each procedure is given. The number of cassettes added to the set of reliable cassettes at each step is shown in parentheses.

CRISPRFinder and PILER-CR. Hence, we relied on the program predictions in that order but assumed that the resulting cassette was supported by all programs that had initially identified any of its variants.

Construction and characteristics of a reliable set. We applied the three programs described above to the GOS data set. The core of the reliable set was formed by 170 CRISPR cassettes found by these programs simultaneously.

Then we took into account the fact that CRISPR cassettes are usually preceded by CRISPR-associated (*cas*) genes. While the short lengths of metagenomic contigs in most cases precluded the observation of *cas* genes, the occurrence of a *cas* gene in the same contig with a predicted CRISPR cassette was taken as evidence that the cassette was real. All contigs with at least one candidate CRISPR cassette were subjected to a similarity search against GenBank. If similarity to at least one *cas* gene was observed, the cassette was accepted. Ten cassettes were added at this step.

Finally, we supplemented the set with cassettes whose repeat consensus was similar to the consensus for a cassette already present in the set. This was motivated by the fact that CRISPR cassette repeats tend to form clusters in the sequence similarity space (27). We applied the standard BLASTCLUST procedure to build clusters of repeat consensuses (details are given under "Repeat clusters" below). We expected that the cassettes found simultaneously by three programs would be more likely to cluster with each other than other cassettes. Indeed, while 87% of the clusters did not contain such cassettes at all, 13% of the clusters consisted mainly of cassettes found by all three programs, thus validating other members of these clusters. Thus, if a cluster contained a repeat from an already accepted cassette, all cassettes with repeats from this cluster were accepted. This approach resulted in the addition of 12 cassettes to the reliable set.

The third natural feature that could indicate a reliable cassette was the similarity of cassette spacers to known archaeal virus and bacteriophage sequences. According to the existing paradigm, CRISPR cassette spacers stem from small fragments of foreign DNA. This hypothesis has been strongly supported by the demonstration of spacers similar to known virus, phage, and plasmid sequences (8, 29, 31). Therefore, we expected that the similarity search of CRISPR cassette spacers against GenBank would bring hits to archaeal virus and bacteriophage entries from the viral subsection and no hits in the eukaryotic subsection (used as a control). However, most similarities in the viral subsection were to eukaryotic viruses, and there were numerous hits to repeats and low-complexity regions of eukaryote genomes. As a control, we used randomly selected fragments of the same contigs (see Materials and Methods). The observed taxonomical distribution of hits was very similar (data not shown). The same situation obtained with a similarity search against marine viromes (3; also data not shown). This could reflect the fact that sequenced bacteriophages and archeal viruses poorly represent viral diversity in natural environments (26). Hence, we could not use similarity to known viruses and phages as a sign of cassette reliability.

The final set of reliable CRISPR cassettes contained 170 cassettes found by all three programs simultaneously plus 10 cassettes accepted on the basis of colocalization with *cas* genes and 12 more cassettes with repeats similar to repeats of already accepted cassettes. Only 3 (2%) cassettes included stretches of N′s, in contrast to 18,862 (65%) cassettes identified by at least one program and 551 (5%) cassettes obtained by excluding N-run-containing pseudocassettes from this set (see "Compilation of the initial set of CRISPR cassettes" above). Forty-three (23%) cassette-containing contigs included putative *cas* genes, and 10 (0.5% of 1,905) spacers from the reliable cassettes were similar to known phage sequences; this percentage would be equivalent to 138 (0.5%) of all 28,424 spacers.

All candidate cassettes identified in this study were collected in a database (http://iitp.bioinf.fbb.msu.ru/vsorokin/crispr). The database interface allows the user to browse and analyze precalculated CRISPR cassettes and their flanking sequences and, in particular, to compare the user's sequences with spacers, repeats, and metagenomic contigs containing at least one CRISPR cassette.

CRISPR cassette occurrence in the metagenome and sequenced genomes. To compare the numbers of CRISPR cassettes found in the GOS data set and in completely sequenced genomes, we used the CRISPRdb database (17), built with CRISPRFinder software (18). We divided the number of CRISPR cassettes in CRISPRdb (1,751 cassettes) by the total length of the corresponding sequenced genomes obtained from GenBank (2.43 Gbp). The average density of CRISPR cassettes in completely sequenced genomes is 0.72 per Mb.

For consistency, to estimate the relative CRISPR cassette occurrence in metagenomic contigs, we used the total number of CRISPR cassettes found by CRISPRFinder divided by the total length of the sequences studied. As already mentioned, CRISPRFinder has problems with N runs. To avoid the occurrence of false cassettes, contigs containing N runs were cut at these runs, and CRISPRFinder was applied to each part independently. We assumed that candidate cassettes found in adjacent fragments of such a contig were parts of the same cassette if their repeats were identical; otherwise, they were considered to form different cassettes.

Applying CRISPRFinder with the above correction, we obtained 8,153 candidate cassettes (3,233 in the original contigs plus 4,922 in contig fragments) for the total length of the metagenome, approximately 4.5 Gbp. Thus, the density of candidate CRISPR cassettes in the GOS data set is 1.8 per Mb. As already mentioned, high fragmentation of the metagenome data set yields a high rate of false-positive results, so this estimate could serve only as an upper boundary. As the lower boundary of the estimate, we can use the number of reliable cassettes, with a density of 0.042 per Mb. The average density of CRISPR cassettes in completely sequenced genomes (0.72 per Mb) is close to the middle of this interval.

The difference may possibly be explained by the bias in the taxonomical distribution of sequenced genomes compared to the relatively unbiased metagenomic sample. An alternative, more technical explanation is as follows. Since CRISPR cassettes are highly variable, homologous sequence loci containing CRISPR cassettes formed different contigs. On the other hand, other loci of very close strains in an environmental sample were assembled in single contigs, thus decreasing the denominator in the ratio.

All cassettes in CRISPRdb are classified as "confirmed" (872 cassettes) or "questionable" (879 cassettes). Applying the same classification scheme (18) to the metagenomic cassettes,

we classified 7,693 of 8,155 cassettes as questionable and only 462 cassettes as confirmed.

Notably, while in the complete genomes the ratio of confirmed to questionable cassettes is about 1:1, in the metagenomes the ratio is 1:17. This difference could arise mainly from the edge effects. Indeed, the metagenomic contigs are much shorter than genomic sequences, and their average length is comparable to the typical length of a CRISPR cassette. Thus, the probability of observing a short (and hence questionable) cassette adjacent to a contig end is high. To address this problem, we calculated that 3,372 (43.8%) questionable cassettes are adjacent to contig ends. Taking that into account brings the confirmed-to-questionable ratio down to 1:1.13, close to the ratio for complete genomes.

Similarities within the metagenome. Using spacers of the reliable cassettes found in the metagenome as queries in a similarity search against the whole GOS data set, we observed 1,658 similarities for 1,905 spacers. However, more than half of the similarities could be extended to the repeat area; hence, they were likely coming from homologous cassettes. We consider them in the next section. Excluding the cases where both repeats and spacers of a cassette were similar to the same contig, we obtained the final set of 818 similarities generated by 282 spacers from 87 cassettes, on the one hand, and 534 independent contigs, on the other hand. All these similarities likely come from metagenome contigs of viral/phage or plasmid origin representing the natural sources of the spacers (protospacer contigs). The phage origin of 147 out of these 534 contigs was confirmed by having at least one phage hit in a BLASTX search. Due to spacer duplications in the cassettes and/or repeated sequences in matching contigs, these 818 similarities represented only 765 unique spacer-protospacer pairs.

We studied the distribution of these similarities with respect to the geographical locations of corresponding reads, thus testing the natural assumption that a virus or phage and a host should coexist. We expected that the probability of observing a similarity between a spacer and a protospacer would be higher for pairs sampled from the same geographical location. To validate this expectation, we constructed sample lists of the spacers and protospacers, and we compared them.

Each (proto)spacer was compared with all GOS reads constituting the corresponding metagenomic contig as described in Materials and Methods. A sample list for each (proto)spacer was formed by the sample labels of corresponding reads. For each spacer-protospacer pair, we compared the sample lists. The number of cases where the sample lists shared at least one geographical location was 660. In order to estimate the significance of this finding, we shuffled the labels of all geographical locations, keeping the frequencies of different sample labels in the total of all sample lists fixed. Then we randomly selected the sample lists for the protospacers. To avoid statistical artifacts, the sizes of the sample lists were fixed to the observed sizes. Then we recalculated the number of locations shared between each spacer and the shuffled list of its protospacer. The shuffling procedure was repeated 10^4 times. The corresponding number of shared locations for each shuffled set was significantly lower than that observed for the nonshuffled sample. Hence, the statistical significance of the observation was at least 10^{-4}. The average number of shared locations for all 10,000 generations was 525.1, and the standard deviation was 11.0.

Thus, the spacer content of CRISPR cassettes reflects the diversity of phages from the same geographical location. It also provides indirect evidence for the considerable difference in the phage populations between ocean locations. This observation agrees with recently published results (23) and a recent hypothesis about the insular pattern of the biogeographical diversity of phages (reviewed in references 11 and 37).

We also analyzed the similarities of the ocean metagenomic cassette spacers to the marine viromes from the Sargasso Sea, the Arctic Ocean, the Bay of British Columbia, and the Gulf of Mexico (3). Using BLAST, we found 378 similarities resulting from 119 unique spacers with 181 matches in the Sargasso Sea samples, 92 matches in the Gulf of Mexico samples, 50 matches in the Bay of British Columbia samples, and 50 matches in the Arctic Ocean samples (the density ratios were 2.3, 1.1, 0.6, and 0.7 per Mbp, respectively). However, the total number of similarities did not differ significantly from the average number of similarities obtained for 10 randomly constructed sets of "pseudospacers," which was 289.3 (standard deviation, 99.5).

Repeat clusters. Since clustering was used for the construction of the reliable cassette set, initially the repeats of all CRISPR cassettes identified by at least one program were clustered by sequence similarity. The application of the BLASTCLUST procedure (see Materials and Methods) to the repeat consensuses resulted in 214 clusters, only 28 of which contained repeats from the reliable CRISPR cassettes.

While the main aim of this procedure was to identify cassettes likely to be evolutionarily related, for completeness we compared our clusters to clusters constructed in reference 27. For that purpose, we used HMMER (http://hmmer.janelia.org/) to construct 12 profiles of alignments provided in reference 27. In order to set the threshold e-value, we subjected the members of these 12 clusters to an HMMSEARCH against all calculated HMM profiles.

We found that under a strict e-value threshold (HMM SEARCH parameters: -E 1e−07 -Z 1), there were no false-positive results; that is, no repeat was recognized by a HMM profile of a different cluster. On the other hand, with this threshold, some repeats were not recognized by the profile of the cluster to which they belonged.

A tolerant threshold (HMMSEARCH parameters: -E 0.01 -Z 1) allowed for relatively effective matching: 8 of 12 clusters matched only themselves, while members of the remaining 4 clusters matched 1 or 2 additional clusters. Hence, we set the e-value threshold to 0.01 and then searched the metagenomic cassette repeats against the obtained set of HMM profiles. The results are shown in Table 1.

Only 15 of 28 metagenomic clusters matched HMM profiles (7 of 12). Fourteen of these 15 clusters contained only CRISPR cassettes found by all three programs simultaneously. One cluster (CLU015) matched three HMM profiles. Five HMM profiles matched more than one metagenomic cluster, demonstrating that our procedure was somewhat more sensitive than that in reference 27.

Taxonomy of CRISPR-containing metagenomic contigs. The distribution of reliable CRISPR cassettes is nonuniform with respect to geographical location (samples). The highest cassette density was observed in the most extreme samples in

TABLE 1. Relationships between repeat clusters

Matching metagenomic cluster[a]	e-value for match to the HMM profile for Kunin's cluster[b]:						
	CLU01	CLU02	CLU03	CLU04	CLU05	CLU09	CLU12
CLU115	0.00058						
CLU015	0.0036				0.0017	5.1e−07	
CLU035		5.3e−09					
CLU074		3.2e−07					
CLU037		3.4e−05					
CLU029			5.5e−12				
CLU027			8.6e−10				
CLU039			2.5e−09				
CLU072			0.00079				
CLU026			0.00086				
CLU040				4.4e−11			
CLU023				2.2e−09			
CLU022							1.2e−05
CLU038							3.7e−05
CLU075							0.0056

[a] Metagenomic clusters containing only reliable cassettes are shown in boldface.

[b] Seven (of 12) clusters from the report of Kunin et al. (27) that demonstrated similarity to the metagenomic clusters (see the text for details).

terms of salinity, that is, in the hypersaline sample from the Punta Cormorant Lagoon, Floreana, and the freshwater sample from Lake Gatun. The same two samples are outliers in terms of taxonomical composition (34). This might be explained by possible additional advantages of the CRISPR mechanism in extreme environments, or it might reflect the fact that both the lake and the lagoon are relatively closed locations, yielding more-competitive environments than usual.

We used a BLASTX-based procedure (see Materials and Methods) to analyze the taxonomical origins of the metagenomic contigs containing reliable CRISPR cassettes. Only 54 of 184 (29%) contigs could be assigned to a taxon. For two contigs, no strong evidence of any particular bacterial type was detected; all BLAST hits came from diverse bacteria. They were assigned the generic label "*Bacteria*." The largest group of annotated contigs ($n = 41$) belonged to the *Proteobacteria* and differed in the taxonomy level of the ascribed labels. Five contigs were assigned to *Cyanobacteria*, three to *Bacteroides*, one to *Chlamydiae*, and one to *Actinobacteria*. Only one archaeal contig was found. This distribution roughly coincides with the general metagenome composition estimated on the basis of 16S rRNA ribotyping (34). In the latter analysis, most contigs were assigned to various classes of the *Proteobacteria* (63%), followed by the *Bacteroidetes*, *Cyanobacteria*, *Firmicutes*, and *Actinobacteria*. The scarcity of the archaea was also observed (34).

For CRISPR cassettes whose spacers showed detectable similarity to phage sequences from GenBank, we compared the taxonomical label of the contig and the known specificity of the phage. Three such cases were detected, and the corresponding taxonomical assignments were in agreement. A spacer of the CRISPR cassette from JCVI_SCAF_1096628067806, assigned to an *Enterobacter* sp., was similar to two *Salmonella enterica* serovar Typhimurium phages, ST104 and ST64T (with two mismatches in both cases). A 32-nucleotide spacer from JCVI_SCAF_1101668599545, assigned to the *Proteobacteria*, was similar to the *Burkholderia* phage BcepNY3 (with six mismatches and a 1-nt indel). A spacer from the JCVI_SCAF_1096627143323 contig (*Cyanobacteria*) was similar to the *Synechococcus* phage S-PM2 with five mismatches.

Most contigs (71%) lacked obvious indicators of their taxonomical origins. These contigs could be divided into three categories: (i) contigs that contained no genes detectable by a BLASTX search, (ii) contigs that contained only universal *cas* genes (which cannot serve as a basis for taxonomy prediction, because they are subject to frequent lateral transfer and their phylogeny does not reflect taxonomy), and (iii) contigs that were (almost) completely covered by a CRISPR cassette(s). The only way to assign taxonomy in these cases is to detect spacer similarity to known phage sequences. The weak, but still detectable, similarity of one spacer of the CRISPR cassette shared by the JCVI_SCAF_1101668626807, JCVI_SCAF_1096627368521, and JCVI_SCAF_1096627147282 contigs to mycobacteriophage Pipefish allowed us to label these contigs as "*Bacteria*; *Actinobacteria*." One more contig, JCVI_SCAF_1096627829539, was assigned to "*Bacteria*; *Proteobacteria*" on the basis of spacer similarity to bacteriophage ϕ80. The JCVI_SCAF_1101668660722 contig contained a spacer almost identical to the *Pseudomonas aeruginosa* phage ϕCTX. Moreover, BLAST hits to the *cas2* gene from this contig came exclusively from *Gammaproteobacteria*; the best hit was from *Pseudomonas mendocina*. Thus, this contig was assigned to the *Pseudomonadaceae*.

Thus, at the current level of knowledge, taxonomy could be predicted only for a small fraction of metagenomic contigs containing CRISPR cassettes. This is caused by the short average length of metagenomic contigs and the limited knowledge of phages, precluding taxonomy prediction on the basis of spacer similarity to specific phages.

Compatibility of the taxonomical assignments of cassettes in clusters. We compared the taxonomical assignments of contigs whose cassettes belonged to the same clusters. Assignments in clusters CLU021 and CLU022 were generally compatible, so that different contigs were assigned labels of the same bacterial phylum and class. In CLU023, cassette c1330 originated from a contig assigned to the *Alphaproteobacteria*, while c1530 came from the *Gammaproteobacteria*.

In contrast, a striking difference could be observed in clusters CLU026 and CLU040. For example, cassette c1105, from *Burkholderia* (*Betaproteobacteria*), was in the same cluster, CLU026, as cassette c0309, from *Cyanobacteria*. Similarly, cluster CLU040 contained cassette c0368, from *Enterobacter* (*Gammaproteobacteria*), and cassette c1199, from *Chlamydia*. This taxonomic inconsistency may result from lateral transfer of CRISPRs, which has been described for complete genomes (16). Notably, in all cases, contigs with similar repeats and different taxonomical assignments shared no other similarities, either in the spacers or in the flanking sequences. This might indicate that not only the spacers, but also the regulatory region and the *cas* genes, are evolutionarily rather labile, in contrast to the repeats.

Evolution of CRISPR cassettes. Clusters of similar repeats seem to be a natural object for the study of CRISPR cassette evolution, since repeat similarity may indicate common ancestry of the cassettes.

We analyzed whether cassettes with very similar repeats can be transcribed in opposite directions. We assumed that the strand of the adjacent *cas* gene determines the direction of transcription. We found that only one cluster (CLU021) contains cassettes adjacent to *cas* genes whose repeats, while almost identical, have different orientations with respect to the

cas genes (c0289 and c1521). However, the distances between the cassettes and the adjacent *cas* genes in these cases are about 400 nt, longer than a standard leader region. Hence, it is possible that the *cas* operon and the cassette are transcribed independently.

In order to distinguish between cassette homology in close species and long-distance lateral transfer of cassettes, we compared flanking sequences for the clustered CRISPR cassettes. Each flanking sequence was used to perform the BLASTN search against the database of all flanking sequences, and the similarities observed were then analyzed manually.

We analyzed the spacers and flanking sequences of the cassettes in each cluster. Each pair of cassettes was described in terms of shared spacers and flanking sequences. The results are presented in files S1 and S2 (http://iitp.bioinf.fbb.msu.ru/vsorokin/crispr/files). We classified the observed patterns of positional relationships of shared spacers based on a limited set of local elementary events, listed below. We also described several complex patterns.

The elementary events (Fig. 3) are as follows.

(i) Simple (one-spacer) indel (class 1). Simple indels are the most frequent type of local difference between homologous cassettes. This event corresponds to a pair of locally identical cassettes, where the only difference in the homologous region is the presence of an additional unique spacer in one of the cassettes. In the absence of outgroups, it is impossible to distinguish between the gain and the loss of a spacer. Given our current knowledge about the mechanisms of CRISPR evolution, recombination between neighboring repeats with subsequent removal of one unit seems to be the most probable scenario.

(ii) Indel of two or more adjacent spacers (class 2). The same evolutionary scenarios described above apply here. This pattern was also observed as a nested event in a number of more complex cases (Fig. 3).

(iii) Simple (one-spacer) adjacent duplication (class 3). In this case, one of two otherwise locally identical cassettes contains a pair of adjacent exact copies of a spacer.

(iv) Duplicated spacer (class 4). This class represents a pair of identical spacers in the homologous regions of cassettes. This is a single-cassette event rather than a difference between two homologous cassettes. In contrast to the preceding class, no cassettes with nonduplicated spacers are known here. However, this class is likely a subset of the preceding class.

(v) Simple (one-spacer) nonadjacent duplication (class 5). In contrast to class 3, in class 5 the copy resulting from the duplication has been inserted at a distance of one or more units from the original spacer.

(vi) Adjacent duplication of several consecutive spacers (class 6). The only example of class 6 is described in the next section (case 3) and is diagramed in Fig. 3B. Six consecutive spacers were tandemly duplicated twice.

(vii) Nonadjacent duplication of several consecutive spacers (class 7). The only example of class 7, involving a two-spacer segment, is described in the next section (case 4) and is diagramed in Fig. 3B.

Reconstruction of evolutionary history in complex cases. The elementary events provide a convenient language for the description of CRISPR cassette evolution. Biological considerations are necessary to resolve specific situations. For example, we believe that deletions rather than insertions are responsible for the observed indels, since they have likely resulted from repeat-mediated recombination. Several likely examples of such recombination were recently demonstrated in experiments (12, 24). If a duplication of a fragment or a spacer is observed, one cannot formally distinguish between the following two possibilities. First, a cassette fragment might have been duplicated to an adjacent or nonadjacent region of the cassette. Second, the same outcome could result from simultaneous or independent acquisition of the same fragment of phage DNA. One should also keep in mind the possibilities that artifacts could arise during the assembly of a repeat-containing fragment and, in particular, that contigs containing CRISPR cassettes might be chimeric. To avoid assembly artifacts, all cassettes identified were analyzed at the level of individual sequencing reads.

Several examples of complicated evolutionary trajectories are listed below.

(i) Combination of a simple indel with a simple nonadjacent duplication (case 1). Case 1 is represented by c0356-c1602, a pair of cassettes from cluster CLU039. It can be considered a superposition of a simple indel (class 1) and a nonadjacent duplication (class 5). The copy resulting from the duplication was inserted at a distance of two units from the original spacer, and one of the units was a unique, newly acquired spacer.

(ii) Multiple nonadjacent duplications in a cassette (case 2). In cassettes c1525 and c1622 from cluster CLU023, a simple nonadjacent spacer duplication (class 5) created two identical spacers (spacers 30 and 02 and spacers 35 and 07) in the shared part of the cassettes. This shared part is extended in the contigs at the opposite sides. In the extension of cassette c1525, an additional copy (spacer 10) of the duplicated spacer was observed (Fig. 3B), likely resulting from one more round of duplication. Moreover, two spacers of these cassettes (spacers 23 of cassette c1525 and 11 of cassette 1622), located in the opposite extensions, are identical. Again, they likely resulted from a simple nonadjacent spacer duplication (class 5), although this duplication was not seen in the same cassette.

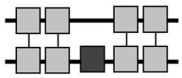
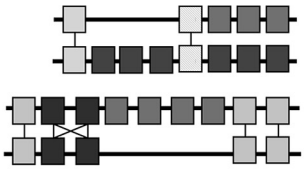
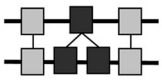
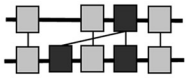
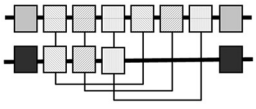
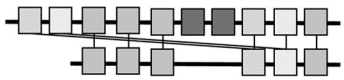
Similar events formed pairs c0291-c1525, c1304-c1340, and c1304-c1575, all from cluster CLU023.

(iii) Duplication of a fragment with several spacers (case 3). Cassette c1526 has two complete tandem copies and one incomplete tandem copy of a six-spacer fragment (class 6) that partly corresponds to cassette c1337.

(iv) Nonadjacent duplication of two spacers at a variable distance (case 4). Case 4 is represented by cassette pair c1575-c1340 from cluster CLU024. Copies of two adjacent duplicated spacers occurred at a distance of three or five units from the original spacers, respectively (class 7). This was accompanied by a two-spacer indel (class 2).

Conclusions. We have developed a procedure for the identification of candidate CRISPR cassettes in metagenomic data. This procedure was used to generate a set of 192 reliable CRISPR cassettes from the GOS data set. The comparison of spacers from these cassettes with the metagenome showed that CRISPR-containing reads tend to originate in the same local samples as protospacer-containing reads. This finding demonstrated that metagenomes are a good source of data for studying the dynamics of phage-host interactions reflected in CRISPR cassettes. Finally, comparison of homologous cassettes yielded a

A

Class	Class description	Examples	Scheme of similarity relationships[b]
1	simple indel	c0292 - c1525 (CLU023) c1330 - c1306 (CLU023) c0293 - c1340 (CLU024) c0293 - c1575 (CLU024)	
2	two or more adjacent spacers indel	c1327 - c1598 (CLU074) c1382 - c1609 (CLU039)	
3	simple (one-spacer) adjacent duplication	c1386 - c1602 (CLU039) c1507 - c1341 (CLU022)	
4	duplicated spacer	c0362 - c1382 (CLU039) c0362 - c1603 (CLU039) c0319 - c1581 (CLU30)	
5	simple (one-spacer) non-adjacent duplication	c1596 - c0289 (CLU021)	
6	adjacent duplication of several consequent spacers	c1337 - c1526 (CLU074)	
7	non-adjacent duplication of several consequent spacers	c1575 - c1340 (CLU024)	

B

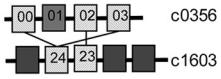
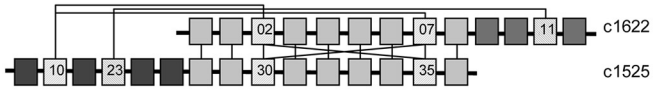
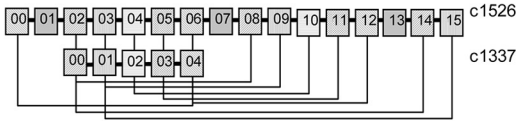
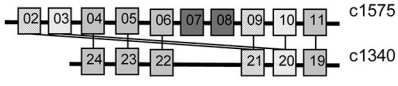
Case	Description	Examples	Scheme of similarity relationships[c]
1	combination of simple indel with simple non-adjacent duplication	c0356 - c1602 (CLU039)	
2	multiple non-adjacent duplications in a cassette	c1525 - c1622 (CLU023) c0291 - c1525 (CLU023) c1304 - c1340 (CLU023) c1304 - c1575 (CLU023)	
3	adjacent duplication of segment with multiple spacers	c1337 - c1526 (CLU074)	
4	non-adjacent duplication of segment with multiple spacers insertion	c1575 - c1340 (CLU024)	

FIG. 3. Evolutionary events forming CRISPR cassettes. Shown are classes of evolutionary events observed in reliable cassettes and examples of complex relationships. *b*, rectangles represent cassette spacers; highly similar spacers are indicated by coinciding tones and linked by thin lines. *c*, spacers are numbered separately in each cassette.

description of a set of elementary events that was used to reconstruct likely evolutionary histories in complicated cases.

ACKNOWLEDGMENTS

We thank Andrey A. Mironov and Konstantin V. Severinov for helpful discussions and the anonymous reviewers for critique and suggestions.

This work was partially supported by the Russian Fund of Basic Research (09-04-01098-a), the Russian Academy of Sciences ("Molecular and Cellular Biology," "Fundamental Problems of Oceanology," and "Biodiversity" programs), the Howard Hughes Medical Institute (55005610), and state contract P1737.

REFERENCES

1. **Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman.** 1990. Basic local alignment search tool. J. Mol. Biol. **215:**403–410.
2. **Andersson, A. F., and J. F. Banfield.** 2008. Virus population dynamics and acquired virus resistance in natural microbial communities. Science **320:** 1047–1050.
3. **Angly, F. E., B. Felts, M. Breitbart, P. Salamon, R. A. Edwards, C. Carlson, A. M. Chan, M. Haynes, S. Kelley, H. Liu, J. M. Mahaffy, J. E. Mueller, J. Nulton, R. Olson, R. Parsons, S. Rayhawk, C. A. Suttle, and F. Rohwer.** 2006. The marine viromes of four oceanic regions. PLoS Biol. **4:**e368.
4. **Arber, W.** 1979. Promotion and limitation of genetic exchange. Science **205:**361–365.
5. **Barrangou, R., C. Fremaux, H. Deveau, M. Richards, P. Boyaval, S. Moineau, D. A. Romero, and P. Horvath.** 2007. CRISPR provides acquired resistance against viruses in prokaryotes. Science **315:**1709–1712.
6. **Bickle, T. A., and D. H. Kruger.** 1993. Biology of DNA restriction. Microbiol. Rev. **57:**434–450.
7. **Bland, C., T. L. Ramsey, F. Sabree, M. Lowe, K. Brown, N. C. Kyrpides, and P. Hugenholtz.** 2007. CRISPR recognition tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. BMC Bioinformatics **8:**209.
8. **Bolotin, A., B. Quinquis, A. Sorokin, and S. D. Ehrlich.** 2005. Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. Microbiology **151:**2551–2561.
9. **Brouns, S. J., M. M. Jore, M. Lundgren, E. R. Westra, R. J. Slijkhuis, A. P. Snijders, M. J. Dickman, K. S. Makarova, E. V. Koonin, and J. van der Oost.** 2008. Small CRISPR RNAs guide antiviral defense in prokaryotes. Science **321:**960–964.
10. **DeLong, E. F., C. M. Preston, T. Mincer, V. Rich, S. J. Hallam, N. U. Frigaard, A. Martinez, M. B. Sullivan, R. Edwards, B. R. Brito, S. W. Chisholm, and D. M. Karl.** 2006. Community genomics among stratified microbial assemblages in the ocean's interior. Science **311:**496–503.
11. **Desnues, C., B. Rodriguez-Brito, S. Rayhawk, S. Kelley, T. Tran, M. Haynes, H. Liu, M. Furlan, L. Wegley, B. Chau, Y. Ruan, D. Hall, F. E. Angly, R. A. Edwards, L. Li, R. V. Thurber, R. P. Reid, J. Siefert, V. Souza, D. L. Valentine, B. K. Swan, M. Breitbart, and F. Rohwer.** 2008. Biodiversity and biogeography of phages in modern stromatolites and thrombolites. Nature **452:**340–343.
12. **Deveau, H., R. Barrangou, J. E. Garneau, J. Labonte, C. Fremaux, P. Boyaval, D. A. Romero, P. Horvath, and S. Moineau.** 2008. Phage response to CRISPR-encoded resistance in Streptococcus thermophilus. J. Bacteriol. **190:**1390–1400.
13. **Edgar, R. C.** 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. **32:**1792–1797.
14. **Edgar, R. C.** 2007. PILER-CR: fast and accurate identification of CRISPR repeats. BMC Bioinformatics **8:**18.
15. **Forde, S. E., J. N. Thompson, and B. J. Bohannan.** 2007. Gene flow reverses an adaptive cline in a coevolving host-parasitoid interaction. Am. Nat. **169:** 794–801.
16. **Godde, J. S., and A. Bickerton.** 2006. The repetitive DNA elements called CRISPRs and their associated genes: evidence of horizontal transfer among prokaryotes. J. Mol. Evol. **62:**718–729.
17. **Grissa, I., G. Vergnaud, and C. Pourcel.** 2007. The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats. BMC Bioinformatics **8:**172.
18. **Grissa, I., G. Vergnaud, and C. Pourcel.** 2007. CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. Nucleic Acids Res. **35:**W52–W57.
19. **Haft, D. H., J. Selengut, E. F. Mongodin, and K. E. Nelson.** 2005. A guild of 45 CRISPR-associated (Cas) protein families and multiple CRISPR/Cas subtypes exist in prokaryotic genomes. PLoS Comput. Biol. **1:**e60.
20. **Hale, C., K. Kleppe, R. M. Terns, and M. P. Terns.** 2008. Prokaryotic silencing (psi)RNAs in Pyrococcus furiosus. RNA **14:**2572–2579.
21. **Hayes, F.** 2003. Toxins-antitoxins: plasmid maintenance, programmed cell death, and cell cycle arrest. Science **301:**1496–1499.
22. **Heidelberg, J. F., W. C. Nelson, T. Schoenfeld, and D. Bhaya.** 2009. Germ warfare in a microbial mat community: CRISPRs provide insights into the co-evolution of host and viral genomes. PLoS One **4:**e4169.
23. **Held, N. L., and R. J. Whitaker.** 2009. Viral biogeography revealed by signatures in Sulfolobus islandicus genomes. Environ. Microbiol. **11:**457–466.
24. **Horvath, P., D. A. Romero, A. C. Coute-Monvoisin, M. Richards, H. Deveau, S. Moineau, P. Boyaval, C. Fremaux, and R. Barrangou.** 2008. Diversity, activity, and evolution of CRISPR loci in Streptococcus thermophilus. J. Bacteriol. **190:**1401–1412.
25. **Kobayashi, I.** 2001. Behavior of restriction-modification systems as selfish mobile elements and their impact on genome evolution. Nucleic Acids Res. **29:**3742–3756.
26. **Kristensen, D. M., A. R. Mushegian, V. V. Dolja, and E. V. Koonin.** 2010. New dimensions of the virus world discovered through metagenomics. Trends Microbiol. **18:**11–19.
27. **Kunin, V., R. Sorek, and P. Hugenholtz.** 2007. Evolutionary conservation of sequence and secondary structures in CRISPR repeats. Genome Biol. **8:**R61.
28. **Lillestøl, R. K., S. A. Shah, K. Brugger, P. Redder, H. Phan, J. Christiansen, and R. A. Garrett.** 2009. CRISPR families of the crenarchaeal genus Sulfolobus: bidirectional transcription and dynamic properties. Mol. Microbiol. **72:**259–272.
29. **Makarova, K. S., N. V. Grishin, S. A. Shabalina, 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 RNAi, and hypothetical mechanisms of action. Biol. Direct. **1:**7.
30. **Marraffini, L. A., and E. J. Sontheimer.** 2008. CRISPR interference limits horizontal gene transfer in staphylococci by targeting DNA. Science **322:** 1843–1845.
31. **Mojica, F. J., C. Diez-Villasenor, 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.
32. **Ptashne, M.** 2004. A genetic switch: phage lambda revisited, 3rd ed. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
33. **Riesenfeld, C. S., P. D. Schloss, and J. Handelsman.** 2004. Metagenomics: genomic analysis of microbial communities. Annu. Rev. Genet. **38:**525–552.
34. **Rusch, D. B., A. L. Halpern, G. Sutton, K. B. Heidelberg, S. Williamson, S. Yooseph, D. Wu, J. A. Eisen, J. M. Hoffman, K. Remington, K. Beeson, B. Tran, H. Smith, H. Baden-Tillson, C. Stewart, J. Thorpe, J. Freeman, C. Andrews-Pfannkoch, J. E. Venter, K. Li, S. Kravitz, J. F. Heidelberg, T. Utterback, Y. H. Rogers, L. I. Falcon, V. Souza, G. Bonilla-Rosso, L. E. Eguiarte, D. M. Karl, S. Sathyendranath, T. Platt, E. Bermingham, V. Gallardo, G. Tamayo-Castillo, M. R. Ferrari, R. L. Strausberg, K. Nealson, R. Friedman, M. Frazier, and J. C. Venter.** 2007. The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. PLoS Biol. **5:**e77.
35. **Sandaa, R. A., L. Gomez-Consarnau, J. Pinhassi, L. Riemann, A. Malits, M. G. Weinbauer, J. M. Gasol, and T. F. Thingstad.** 2009. Viral control of bacterial biodiversity—evidence from a nutrient-enriched marine mesocosm experiment. Environ. Microbiol. **11:**2585–2597.
36. **Sorek, R., V. Kunin, and P. Hugenholtz.** 2008. CRISPR—a widespread system that provides acquired resistance against phages in bacteria and archaea. Nat. Rev. Microbiol. **6:**181–186.
37. **Thurber, R. V.** 2009. Current insights into phage biodiversity and biogeography. Curr. Opin. Microbiol. **12:**582–587.
38. **Tijdens, M., H. L. Hoogveld, M. P. Kamst-van Agterveld, S. G. Simis, A. C. Baudoux, H. J. Laanbroek, and H. J. Gons.** 2008. Population dynamics and diversity of viruses, bacteria and phytoplankton in a shallow eutrophic lake. Microb. Ecol. **56:**29–42.
39. **Tyson, G. W., and J. F. Banfield.** 2008. Rapidly evolving CRISPRs implicated in acquired resistance of microorganisms to viruses. Environ. Microbiol. **10:**200–207.
40. **Williamson, S. J., S. C. Cary, K. E. Williamson, R. R. Helton, S. R. Bench, D. Winget, and K. E. Wommack.** 2008. Lysogenic virus-host interactions predominate at deep-sea diffuse-flow hydrothermal vents. ISME J. **2:**1112–1121.
41. **Williamson, S. J., D. B. Rusch, S. Yooseph, A. L. Halpern, K. B. Heidelberg, J. I. Glass, C. Andrews-Pfannkoch, D. Fadrosh, C. S. Miller, G. Sutton, M. Frazier, and J. C. Venter.** 2008. The Sorcerer II Global Ocean Sampling Expedition: metagenomic characterization of viruses within aquatic microbial samples. PLoS One **3:**e1456.
42. **Wommack, K. E., and R. R. Colwell.** 2000. Virioplankton: viruses in aquatic ecosystems. Microbiol. Mol. Biol. Rev. **64:**69–114.