The CRISPR Journal
Volume 1, Number 2, 2018
Mary Ann Liebert, Inc.
DOI: 10.1089/crispr.2017.0022

The CRISPR Journal

RESEARCH ARTICLE

CRISPRdisco: An Automated Pipeline for the Discovery and Analysis of CRISPR-Cas Systems

Alexandra B. Crawley[1], James R. Henriksen[2], and Rodolphe Barrangou[1]

Abstract

CRISPR-Cas adaptive immune systems of bacteria and archaea have catapulted into the scientific spotlight as genome editing tools. To aid researchers in the field, we have developed an automated pipeline, named CRISPRdisco (CRISPR discovery), to identify CRISPR repeats and *cas* genes in genome assemblies, determine type and subtype, and describe system completeness. All six major types and 23 currently recognized subtypes and novel putative V-U types are detected. Here, we use the pipeline to identify and classify putative CRISPR-Cas systems in 2,777 complete genomes from the NCBI RefSeq database. This allows comparison to previous publications and investigation of the occurrence and size of CRISPR-Cas systems. Software available at http://github.com/crisprlab/CRISPRdisco provides reproducible, standardized, accessible, transparent, and high-throughput analysis methods available to all researchers in and beyond the CRISPR-Cas research community. This tool opens new avenues to enable classification within a complex nomenclature and provides analytical methods in a field that has evolved rapidly.

Introduction

CRISPR-Cas* bacterial and archaeal immune systems remain of high interest across many domains of the life sciences, including food science, molecular biology, prokaryotic evolution, and as a technology from pharma to next-generation crops.[1–4] The unifying interest in CRISPR is the tremendous wealth of applications this technology affords. While application and tool development using a handful of characterized CRISPR-Cas systems has exploded, the annotation and discovery of systems remains an ongoing challenge for microbiologists and bioinformaticians to solve. The ability to identify CRISPR-Cas systems can benefit the greater scientific community, from microbiologists attempting to learn about adaptive immunity in prokaryotes, to molecular biologists interested in harnessing the nucleic acid-targeting functions of various Cas proteins. In recent years, our understanding of the vast diversity in CRISPR-Cas has led to the identification of two classes of systems,[4–8] six types and 23 subtypes.[8–10] Beyond that, an additional uncharacterized type with five subtypes has been recently proposed.[5] The CRISPR nomenclature is updated over time by the field, which can be difficult for novice CRISPR users to learn and keep up with, let alone use in additional gene discovery. There is a need to create a pipeline that can accurately, easily, and reproducibly identify and annotate CRISPR-Cas systems in DNA sequences.

There has been a great development in the tools instrumental to detect CRISPR repeats, notably CRISPRFinder and CRT,[11–17] and a great effort to group and establish canonical Cas proteins,[18–20] but there is a need to meld these worlds and develop an integrated usable tool. Currently, investigators interested in CRISPR must utilize separate tools for repeat identification and *cas* annotation. Furthermore, there is a lack of automated software that performs both of these tasks and can apply logic to determine basic CRISPR and *cas* identity. Here, we have developed a bioinformatic pipeline that builds on current tools and nomenclature to allow automated detection of entire CRISPR-Cas systems. It applies homology-based propagation of annotation, repeat detection, and nomenclatural logic to determine type, subtype, and the completeness

*Clustered Regularly Interspaced Short Palindromic Repeats.

[1]Department of Food, Bioprocessing, and Nutrition Sciences, North Carolina State University, Raleigh, North Carolina; [2]AgBiome, Durham, North Carolina.

Address correspondence to: Rodolphe Barrangou, rbarran@ncsu.edu

of a system. An important function of our pipeline is to determine repeat orientation—a feature that is extremely vital in determining crRNAs and investigating the function and activity of CRISPR-Cas systems. The pipeline also has optional capabilities to determine the novelty of systems identified compared to user-defined databases.

There is a critical need for software used in research to be reproducible, accessible, and transparent. Release of code and data under open source licenses in public repositories facilitates transparency, reuse, and improvement,[21] and is essential for replication and reproducibility of published results.[22] Release of bioinformatic analysis in interactive notebooks improves transparency and reuse.[23] Containerization technology can be used to address aspects of reproducibility and ease dependencies requirements accessibility.[23–26] All software and data used in this analysis are available under a GPL2 license and are provided in a reusable Docker container in order to facilitate widespread use. Further development by the wider CRISPR and bioinformatic community to a public code repository at http://github.com/crisprlab/CRISPRdisco is welcomed.

Materials and Methods

CRISPRdisco bioinformatic pipeline

The pipeline, CRISPRdisco, is written in python 2.7 and is distributed in a Docker container with all dependencies (Fig. 1). Homology between reference sets of proteins and the detection of CRISPR repeats, along with typing logic, are used to categorize systems. The included reference sets were generated by manual review from Makarova *et al.*,[10] Shmakov *et al.*,[5] and Burstein *et al.*[9] for all Cas proteins identified to date. From nucleotide and protein files for each genome, annotation propagation by homology was by BLAST of reference sets against a protein database of coding sequence on a DNA sequence with an

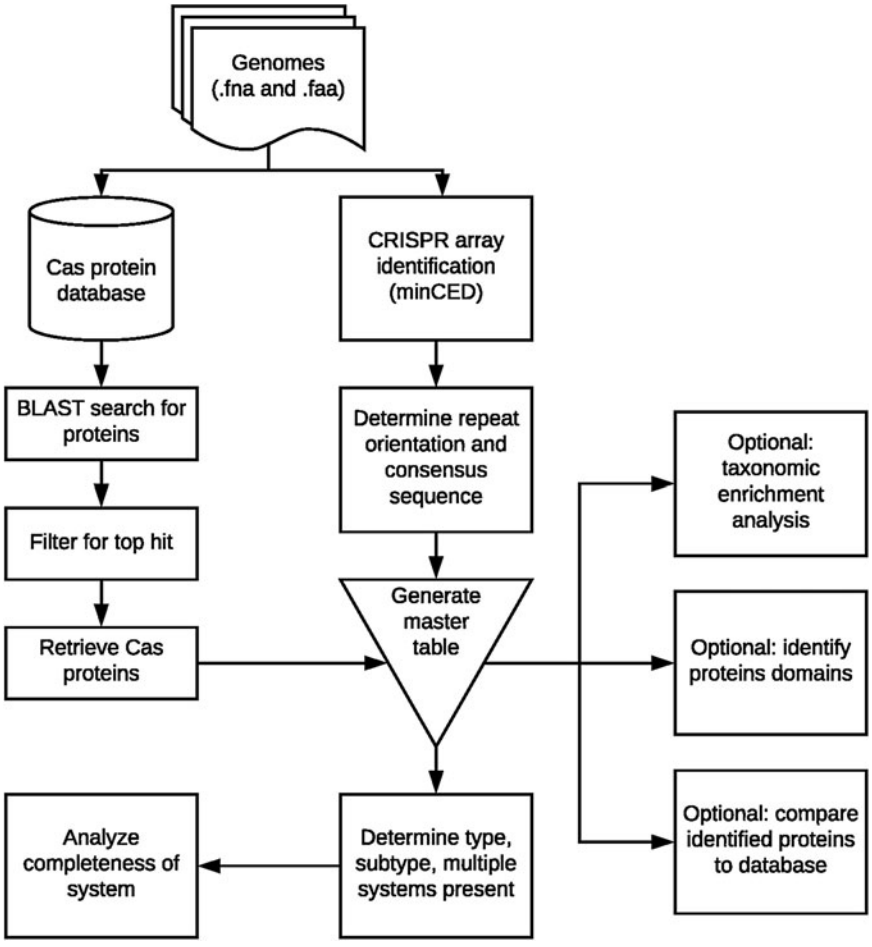

FIG. 1. Overview of bioinformatic pipeline. The bioinformatic pipeline uses nucleotide and coding sequence information to annotate both CRISPR repeats and Cas proteins. This information is combined to determine type, subtype and completeness of CRISPR-Cas systems in archaeal and bacterial genomes.

evalue cutoff of 1e-6, requiring 40% identity to reference protein sequence >50% of length of reference, and using bitscore to determine the top match.

The CRISPR arrays are identified using minCED (mining CRISPRs in environmental data sets; http://github.com/ctSkennerton/minced), a derivative of CRISPR Recognition Tool[11] that is more conservative in repeat calling and allows more flexible user outputs. Custom code determines the orientation of the repeats, generates the consensus repeat sequences, and returns the number of repeats, indicating the size of the array. Once the CRISPR loci have been identified, the presence and absence of genes are used to assign type and subtype, detect multiple systems in a genome, and determine the completeness of the system through identification of missing repeats and Cas proteins.

In addition, the user has the option to compare the proteins identified in their data set to local databases and look for protein domains using hmmscan.[27] The default output for the pipeline includes a master summary table that lists the number of repeat loci, all Cas proteins, and the putative assigned type and subtype. Additionally, this table contains the annotations for completeness of the system and any missing elements. There are additional supplemental tables that can be saved, including the consensus repeat information and sequence for each locus, the number of individual proteins detected in a genome and the blast results for each protein detected. For more information on how to use the pipeline, please visit the Github repository.

Occurrence of CRISPR-Cas systems

A collection of genomes from Bacterial and Archaeal taxa categorized as full and complete from the RefSeq database released before December 16, 2016, containing 5,201 replicons (2,777 genomes, 2,424 extrachromosomal sequences including plasmids) was used to build and fine-tune the pipeline. We compared the output with the currently used gold standard tools and publications to determine the accuracy of our pipeline. The accuracy of our repeat detection and Cas identification were compared to CRISPRdb (September 2016, hypothetical annotations not used) containing 2,052 genomes and Makarova *et al.*[10] containing 1,263 genomes, respectively.

Only loci where *cas* (and the Cas proteins they encode) and CRISPR co-occur were considered for the distribution of loci and size of array analysis. Systems were said to ''co-occur'' when CRISPR repeats were within 20 kb of the signature protein in a genome. Signature proteins are the primary nuclease in each type of CRISPR-Cas system. Specifically, Cas3 is a single-strand DNA exonuclease that is the signature protein for type I[28–31]; Cas9 is a blunt-cutting double-strand DNA endonuclease that is the signature protein for type II[32–34]; Cas10 is an RNA or DNA nickase that is the signature protein for type III[35–37]; Csf4 is the signature nuclease for type IV[5]; Cas12 (formerly Cpf1) is a staggered-cutting double-strand DNA endonuclease signature protein for type V[5,8,38–41]; and finally, Cas13 is an RNA-shredding nuclease that is the signature protein for type VI.[42–44] If the repeats from these loci were also detected elsewhere in the genome, custom code was used to calculate the true size of split arrays when there were multiple repeat-spacer arrays with the same repeat sequence located throughout the genome. Notched box plots display the distribution of these sizes, with the notch indicating the confidence interval around the mean. The missing components were annotated using the count for number of CRISPR loci co-localized with Cas. Means comparisons were performed in JMP Pro 12 using Tukey's HSD test with a 0.05 alpha level.

Phylogenetic distribution and enrichment

Only loci where Cas and repeats co-occurred were used for the phylogenetic distribution and enrichment analysis. Custom code (http://github.com/crisprlab/test) was used to determine the rate of occurrence of CRISPR-Cas systems in bacterial and archaeal genomes by different levels of taxonomic classification. Phylogenetic enrichment analyses were performed using the NCBI taxonomic information. The occurrence of CRISPR-Cas systems by taxonomic level was normalized to the number of sequences in each clade (Fig. 2 and Supplementary Table S1; Supplementary Data are available online at www.liebertpub.com/crispr).

Results

Pipeline performance and comparison

CRISPR arrays and Cas sequences were annotated using the CRISPRdisco pipeline in 5,201 genomes and plasmids from the RefSeq repository at NCBI (Supplementary Table S1). The runtime for the entire pipeline on a single replicon was between 2 and 5 min, depending on the amount of sequence, when parallelizing with 10 CPUs per replicon on a system with 20 virtual 2.13 GHz Intel® Xeon® CPUs and 128 GB RAM total. The majority of the runtime is spent running BLAST. Total runtime scales nearly linearly with number of sequences.

The pipeline presented here showed agreement with CRISPR repeat detection in 94% of genomes analyzed with the CRISPRdb (Supplementary Table S2)[12,45] and ~99% agreement in Cas detection with supplemental table nrmicro3569-s7 from Makarova *et al.*[10] (Table 1). When our pipeline disagreed with the presence or absence of CRISPR components in genomes, we used the entire

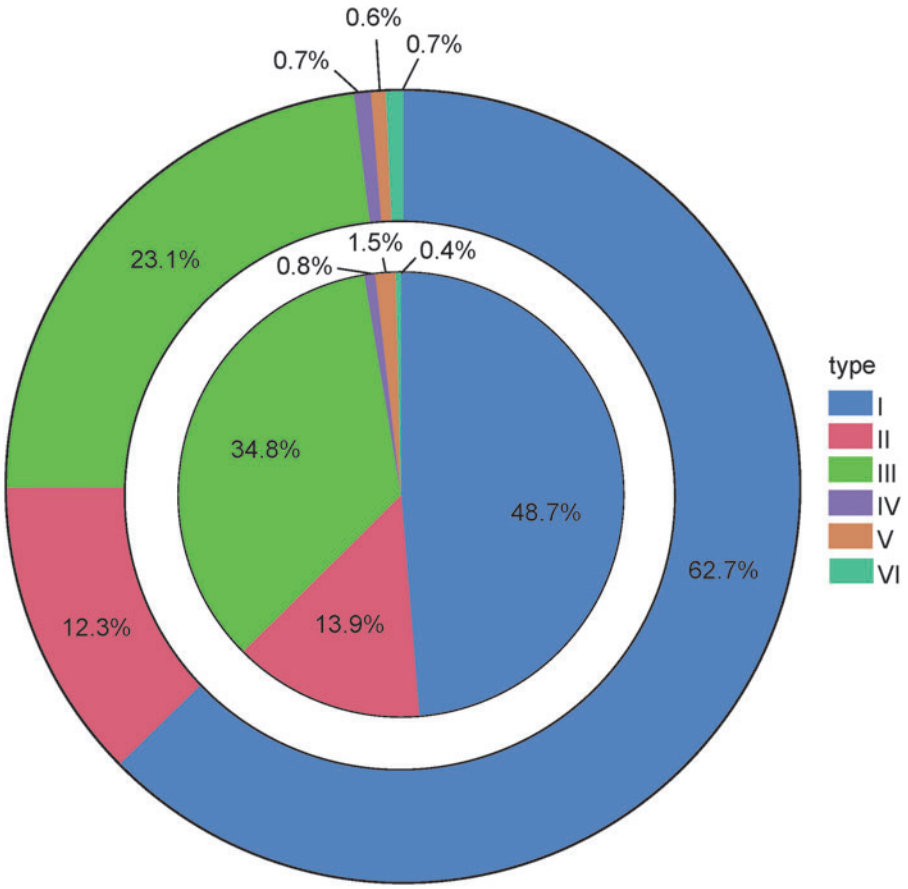

FIG. 2. Distribution of type of CRISPR-Cas system. The outer ring demonstrates the total distribution of CRISPR-Cas systems detected in archaea and bacteria. The inner ring shows the same distribution with a correction for sampling bias in the data set used. The inner ring is normalized to the number of replicons in each taxonomic class used for this analysis.

CRISPR-Cas locus to determine which annotation software was more likely to be accurate. When comparing CRISPR occurrence, if CRISPR repeats were detected in one software and not the other, we used the presence of *cas* genes to determine whether the repeats were likely a true CRISPR-Cas system and the absence of *cas* genes to determine whether the repeats were likely non-CRISPR repetitive elements in a genome. The inverse of this logic was used when CRISPR repeats were not detected. When comparing Cas occurrence, if Cas coding sequences were detected using one software and not the other, we used the presence of CRISPR repeats to determine if the locus was likely a true CRISPR-Cas system and vice versa. When using the whole locus to determine the accuracy of the pipeline, we are >98% accurate in CRISPR repeat calling and 99% accurate in Cas detection relative to these other sources.

Overall rate of occurrence

CRISPR or Cas elements were detected in 1,963/5,201 genomes and plasmids from the RefSeq database (Supplementary Table S1). Of the 1,963 CRISPR elements containing replicons, only 1,065 complete CRISPR-Cas systems were identified where *cas* genes co-localized

Table 1. Pipeline agreement of CRISPR repeat and Cas identification

CRISPR identification	*Number (%)*	*Cas identification*	*Number (%)*
Agree (presence/absence of CRISPR arrays)	1,931 (94%)	Agree (presence/absence of Cas)	1,249 (98.9%)
Cas presence favors CRISPRdisco pipeline	77 (3.7%)	CRISPR presence favors CRISPRdisco pipeline	8 (0.6%)
Cas presence favors CRISPRfinder	44 (2.1%)	CRISPR presence favors orig. publication	6 (0.5%)
Total	2,052	*Total*	1,263

Table 2. Number of systems where CRISPR repeats and Cas co-occur

Class	*Type*	*Signature Cas protein*	*Total number of proteins identified*	*Number of systems that co-occur with repeats*	*Percent of Cas that co-occur with repeats*
Class 1	Type I	Cas3	1257	686	54.6%
	Type III	Cas10	507	234	46.1%
	Type IV	Csf4	14	0	0%
		DinG	609	7	1.1%
Class 2	Type II	Cas9	245	125	51.0%
	Type V	Cas12a	8	6	75.0%
		Cas12b	2	0	0%
		Cas12c	0	N/A	N/A
		Cas12d	0	N/A	N/A
		Cas12e	0	N/A	N/A
	Type VI	Cas13a	4	3	75%
		Cas13b	71	4	5.6%
		Cas13c	0	N/A	N/A

with CRISPR repeats (Table 2). Only about half of all signature proteins occurred with repeats and a complete *cas* locus. The majority of complete systems identified were type I systems (686/1,065 systems; 64%), followed by type III (234 systems; 22%), then type II (125 systems; 12%), with IV, V, and VI accounting for a combined 20 systems (2%; Fig. 2 and Table 2). In many systems where the *cas* co-occur with repeats, we were able to detect additional repeats elsewhere in the genome; these are called ''split arrays'' (Table 3). Split arrays were extremely common in type III systems ($\sim$73%), frequent in type I systems ($\sim$50%), and rare in type I systems ($\sim$5%). Split arrays are proposed to utilize the same set of Cas proteins, though they only occur in a single locus. Additionally, we identified 459 genomes that contained partial or complete Cas proteins but were missing repeats entirely (Table 4). Genomes with complete Cas systems but lacking repeats may be lacking repeat-spacer arrays due to poor assemblies. Newer repeat identification tools such as CRISPRdetect may perform better in detecting unusual or degenerate CRISPR arrays.[17]

Table 3. Number of systems missing repeats

Type of system	*Completeness of system*	*Number of potential loci*
Type I	Partial	31
	Complete	14
Type II	Partial	7
	Complete	16
Type III	Partial	24
	Complete	6
Type IV	Partial	271
	Complete	16
Type V	Partial	1
	Complete	0
Type VI	Partial	29
	Complete	0

When looking at the size of repeat-spacer arrays, we noticed a small effect size but statistically significant differences in the number of repeats based on the type of CRISPR-Cas system (Fig. 3). Only type II systems were statistically smaller than types I and III. Type II systems contained on average 25 repeats, while type I and type III systems both contain on average 41 and 47 repeats, respectively. The larger array size in type I and type III systems may be due to the higher rate of split loci.

Another metric to characterize the canonical subtypes is to look at the size of the CRISPR repeat. Each subtype has its own unique repeat length that is often different from subtypes within a CRISPR type. In type I systems this is extremely evident (Fig. 4). The type I-A subtype has the shortest repeats, 24–25 nucleotides, followed by type I-E and I-F with 28–29 nucleotide repeats, I-B with 30 nucleotide repeats, I-C with 32 nucleotide repeats, and finally I-U with 36–37 nucleotide repeats (Fig. 4). As CRISPR repeat length is conserved within a subtype, this information can be used to confirm subtypes and ensure the accuracy of CRISPR repeat detection software.

Taxonomic distribution

When looking at the rate of occurrence of CRISPR-Cas systems at different taxonomic levels, we see divergences from the canonical rate of occurrence of these systems

Table 4. Number of split loci

Type	*Number single loci*	*Number split loci*	*Percent split*
I	230	226	49.6%
II	115	6	5.0%
III	34	95	73.6%
IV	2	2	50%
V	4	1	20%
VI	4	1	20%
Total	389	331	45.9%

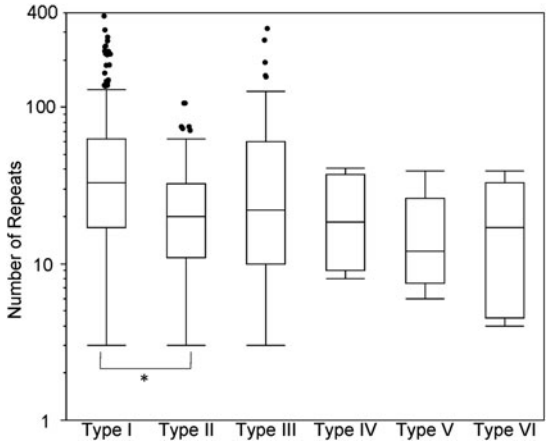

FIG. 3. Size of CRISPR loci. The total number of repeats from genomes where repeats co-occur with *cas* genes broken down by type. *Average size is statistically different based on Tukey's HSD (alpha 0.05).

(Fig. 1, Table 3, and Supplementary Table S3). The current published rate of occurrence for CRISPR-Cas systems in archaea is around 90% of genomes and in bacteria around 45%. When looking across the entire kingdom, we see a rate of occurrence of any type of CRISPR feature in chromosomes around 90% in archaea and 66% in bacteria. When considering only complete systems that contain both a signature protein and repeat-spacer array that co-occur in the genome, the rates are much lower—around 20% of bacteria and 60% of archaea. There are a number of taxonomic classes of bacteria that hover around the 30–50% rate of occurrence of CRISPR-Cas systems, but there are many classes that fall at the extremes of CRISPR occurrence. In our analyses, the Chlamydiia group (173 sequences surveyed) was the only class that did not contain a single CRISPR-Cas feature. Several classes of bacteria had very low rates of occurrence, including Alphaproteobacteria, Mollicutes, and Spirochaetia. Conversely there were several bacterial classes that had high rates of occurrence of CRISPR-Cas systems, including Actinobacteria, Clostridia, Deinococci, Thermotogae, Aquificae, and Chlorobia (Fig. 5). Within the archaeal classes, CRISPR features were almost ubiquitously identified in all classes except the Halobacteria class. Overall, there does appear to be a taxonomic dependence when investigating the rate of occurrence of CRISPR-Cas systems in isolates in this data set. The inherent biases in which strains are isolated, sequenced, and deposited preclude extension of these rates to prokaryotic populations in nature.

CRISPR-Cas features were also identified on plasmids (Supplementary Table S4). Often, putative CRISPR-Cas systems on plasmids were missing key features, such as signature proteins or co-occurring repeats. In total, 64 bacterial and 8 archaeal genera harbor plasmids with CRISPR features. In bacteria, we could detect complete systems for types I, II, III, and IV systems on plasmids, while only complete type I systems were detected on archaeal plasmids. The occurrence of these systems, and system parts, on plasmids suggests horizontal gene transfer aids in disseminating these genetic features through environments.

Looking at the diversity of CRISPR-Cas systems in different taxonomic classes, there is no class that carries all six types of systems. In archaea, types I and III are the predominant types; components of these types were detected in 198/202 genomes in that domain. Type I systems were detected in 121 archaeal genomes, while type III systems were detected in 78 genomes. Type V systems have not previously been reported in archaeal genomes. This analysis detected two complete type IV systems detected in the Methanomicrobia class and one type V-A system detected in an unclassified archaeal (NC_020913 archaeon Mx1201). In bacteria, type I systems are the dominant type; they were detected in 16% of genomes (787/4,949). Types II, III, and IV systems occurred in roughly the same number of genomes (II in 4.6%, III in 6.1%, and IV in 5.1%). Type V systems were by far the rarest, as only seven were detected in all bacterial genomes, primarily in the Gammaproteobacteria and Clostridia classes. Type VI systems were detected in $\sim$1% of bacterial genomes.

Discussion

The pipeline presented here relies on homology-based propagation of annotation and canonical *cas* gene presence and absence to assign type and subtype of systems. Solely using this approach may lead to the inability to assign subtypes properly. For example, in a genome that is misassembled or incomplete, a system that more closely functions as a type II-A system may be called a II-C due to a single gene, *csn2*, missing. The genomes used in this analysis are all complete, so this issue is hopefully minimized. The pipeline represents a preliminary step in the identification of potentially functional CRISPR-Cas systems. Developments of the pipeline and CRISPR knowledge base will be needed to connect the coding potential of CRISPR-Cas systems further to system functionality based solely on *in silico* sequence.

When comparing the functionality of the pipeline described here, we were only able to compare positive results to published Cas data by Makarova *et al.*,[10] Shmakov *et al.*,[8] and Burstein *et al.*[9] The publications that have defined our CRISPR-Cas nomenclature only

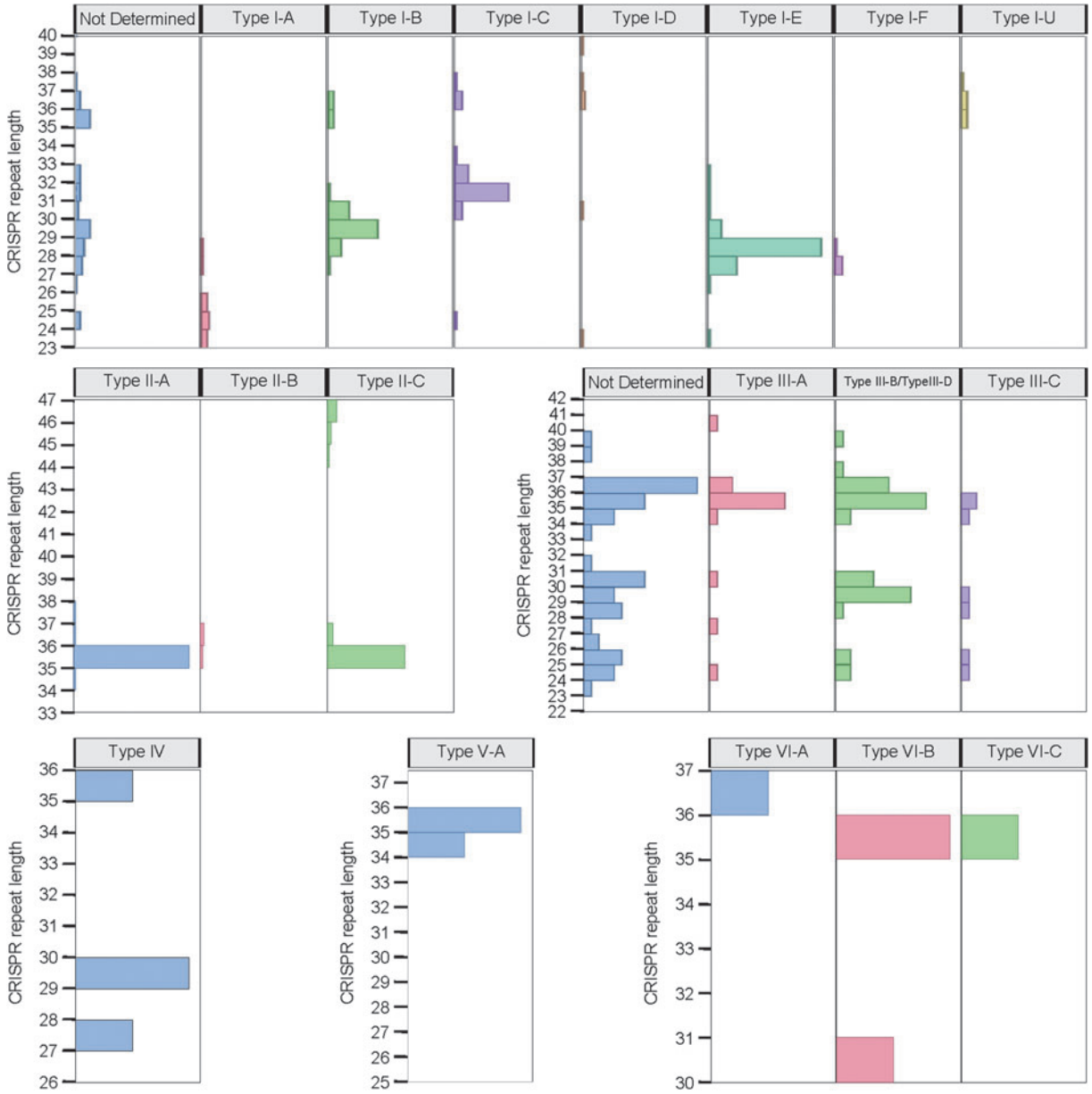

FIG. 4. Length of CRISPR repeats by CRISPR subtype. The distribution of CRISPR repeat length by subtype is shown. The bar height is scaled by CRISPR-Cas type. Relative abundance of each subtype within a type can be determined by bar height on the *x*-axis.

encompass the occurrence of these systems and not the absence of these systems in genomes. We cannot comment on the false-positive rate of our Cas protein detection, as there is no current public repository of genomes predicted to be lacking CRISPR-Cas systems. Publishing only the Cas-positive genomes may obfuscate the true rate of occurrence of CRISPR-Cas systems. One benefit of using a whole locus approach to annotate CRISPR-Cas systems is the ability to use information from both repeat and Cas detection to inform us whether potential proteins identified by the pipeline are true Cas proteins or merely homologs. The high rate of incomplete systems suggests the majority of these systems may not be active in all three stages of CRISPR immunity. Functional characterization of the stages of immunity in partial and complete systems will aid in translating the *in silico* prediction of a system to its biological activity.

No type II systems were detected in archaeal genomes included in this study, though archaeal Cas9s have been reported in the literature.[9] The first detection of Cas9s

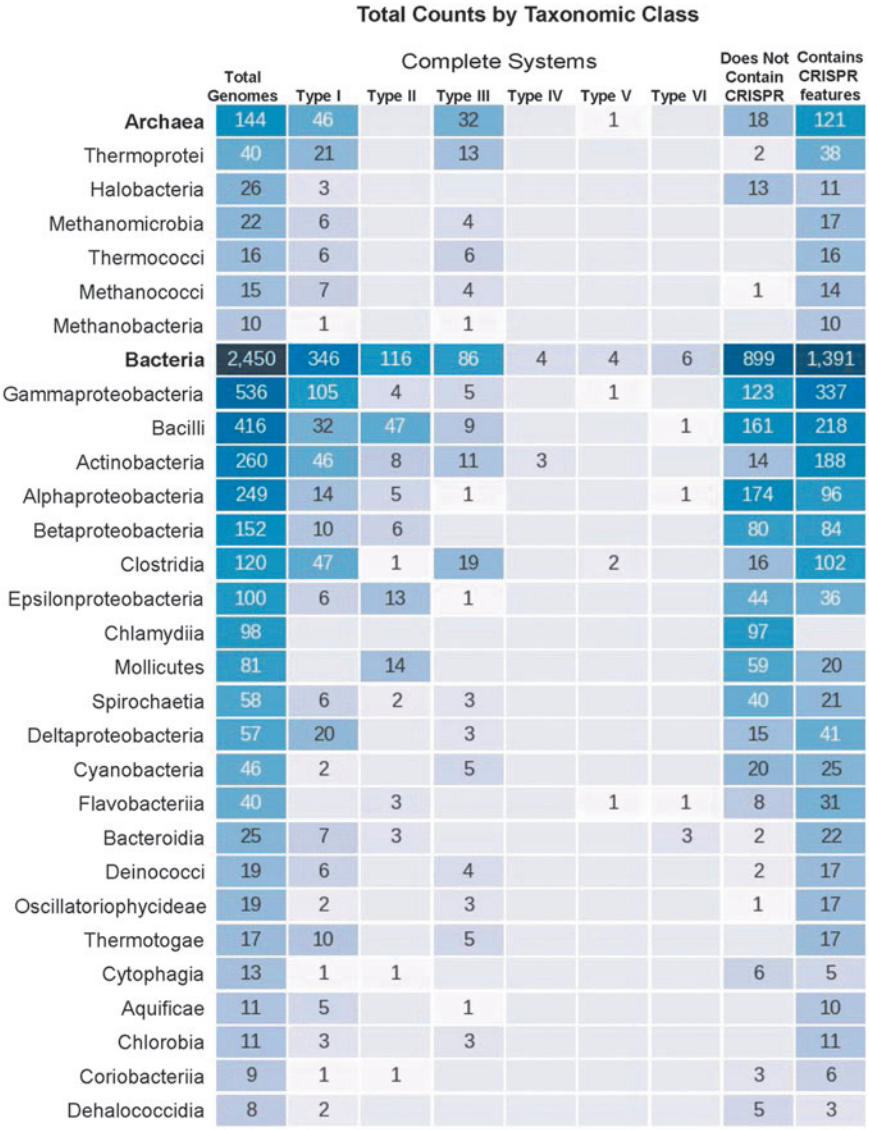

FIG. 5. Occurrence of CRISPR-Cas systems in archaea and bacteria. The total number of complete CRISPR-Cas systems where proteins co-occur with repeats in chromosomes is depicted in the left heat map. The taxonomic class level was used to group the genomes. Total counts for archaea and bacteria appear at the top of each kingdom. The classes are ranked by total number of genomes included in the search. The column titled "Does Not Contain CRISPR" includes all genomes without a single *cas* gene or CRISPR repeat detected, while the "Contains CRISPR features" column denotes that a protein coding sequence with homology to known Cas proteins or CRISPR repeats were detected in the chromosome. The right heat map displays the percent of genomes containing CRISPR-Cas systems when normalized for total number of genomes included in this analysis.

in archaeal genomes was from metagenomic assemblies. While the pipeline was not run on metagenomes in this analysis, the tool we present here can be used to annotate Cas proteins and CRISPR repeats in binned assemblies and in more fragmented shotgun genome assemblies of isolated strains. Additionally, as sequencing of novel microbiomes continues to expand our knowledge of the known microbial world, we expect to detect novel CRISPR-Cas systems more frequently. This pipeline should be updated to include all novel reference proteins.

Determining type and subtype of some of the newer CRISPR-Cas systems will require some additional functional analyses to determine which protein profiles actually constitute an active CRISPR-Cas system. The current

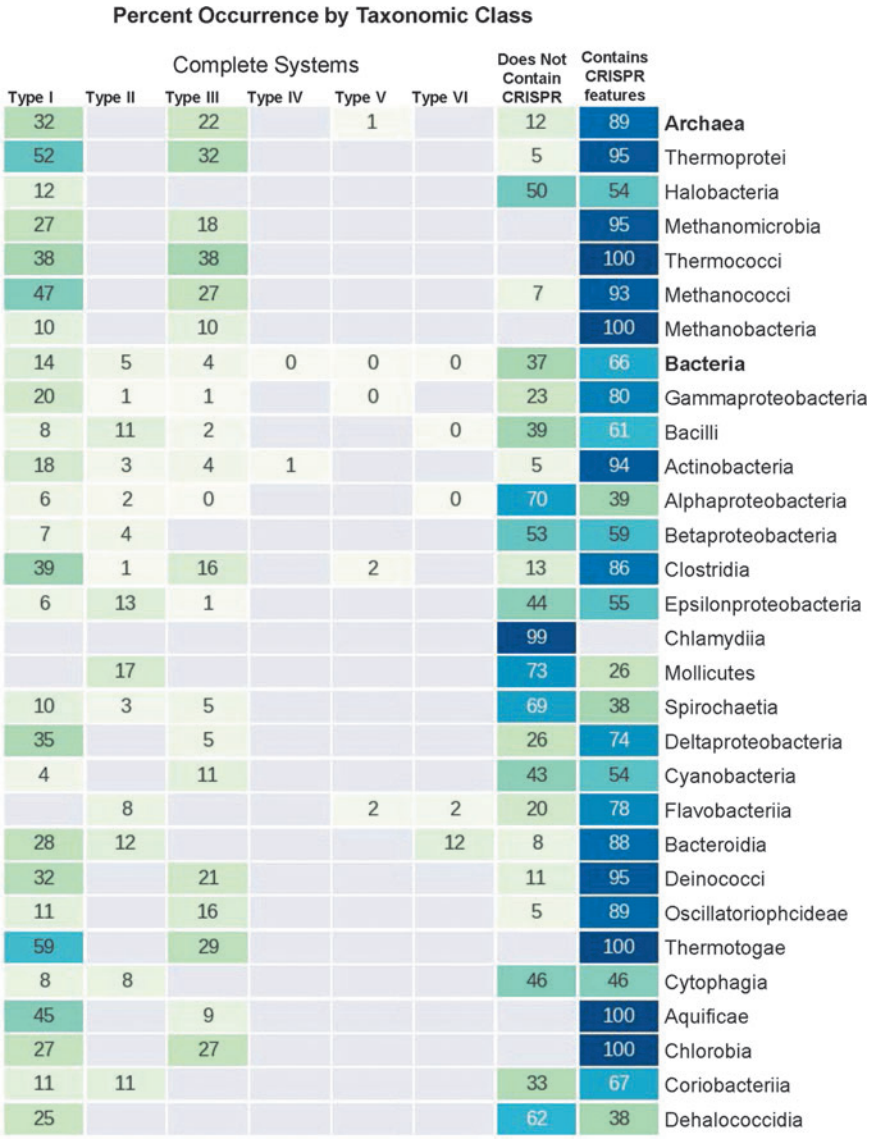

FIG. 5. (Continued).

signature proteins for type IV systems are Csf4 and DinG. Using DinG to determine type IV CRISPR-Cas systems drastically changes the rate of occurrence of these systems. DinG occasionally appears in bacterial genomes without CRISPR arrays and without other Cas proteins. This occurrence will cause us to overestimate the number of type IV CRISPR-Cas systems if the whole locus information is not considered when determining type. The same issue arises with the V-U proteins. Proteins with similar sequences and domain structures to the c2c-proteins previously identified as V-U potential class 2 systems are found throughout bacterial genomes and many are currently annotated as putative transposases.

The current canonical rate of occurrence of CRISPR-Cas systems in bacterial and archaeal genomes is skewed by the genomes deposited in public repositories, with notable bias toward pathogenic bacteria and model organisms. As novel microbes from diverse environments are sequenced, we will likely see the rate of CRISPR-Cas occurrence change. Recent reports have suggested that the true rate of CRISPR occurrence is much lower than the canonical 40% of bacteria and 90% of archaea.[46,47] We propose that the environment of the native host as well as the taxonomic classification is greatly important when discussing the rate of occurrence of CRISPR, as some taxonomic groups are virtually devoid of CRISPR while others are extremely rich.

Conclusion

There is a tremendous opportunity to determine the link between CRISPR functionality *in vivo* and CRISPR-Cas diversity *in silico*, as the current link remains unknown. This pipeline can help provide the bioinformatic resources required to identify, categorize, and characterize putative CRISPR-Cas systems in bacterial and archaeal genomes. We hope by having a single tool to detect and assign class, type, and subtypes accurately and reproducibly to potential CRISPR-Cas systems, users will be able to generate a greater knowledge base on how truly diverse these systems are and understand how they occur and are distributed in nature. The rate of occurrence of CRISPR-Cas systems changes with the taxonomic granularity used to assess this rate, and thus the rate should change across taxonomic groups of interest. As the majority of CRISPR-Cas systems identified *in silico* are incomplete, it is likely very few of these systems are truly active functionally as adaptive immune systems.

Acknowledgments

We would like to thank the CRISPR lab for additional support and input on this manuscript. This research was funded by AgBiome/LifeEDIT and NCSU start-up funds. The ''AgBiome Team'' includes Charles Pepe-Ranney, Kira Roberts, Rebecca Thayer, Matt Biggs, Mauricio Borgen, Jessica Parks, Mark Moore, Tyson Bowen, Lynn Dickey, Vinh Pham, Tedd Elich, Kelly Williamson, W. Murray Spruill, Eric Ward, Scott Uknes, and Dan Tomso.

Author Disclosure Statement

A.B.C. and R.B. are inventors on several patents related to CRISPR technology and derived applications. R.B. is a shareholder of Caribou Biosciences, a co-founder and SAB member of Intellia Therapeutics, and co-founder and chairman of the SAB for Locus Biosciences. R.B. is the Editor-in-Chief of *The CRISPR Journal* and has been excluded from all editorial proceedings with this manuscript. This work was funded by AgBiome/LifeEDIT.

References

1. Jackson SA, McKenzie RE, Fagerlund RD, et al. CRISPR-Cas: adapting to change. *Science* 2017;356. DOI: 10.1126/science.aal5055.
2. Knight S, Tjian R, Doudna J. Genomes in focus: development and applications of Crispr-Cas9 imaging technologies. *Angew Chem Int Ed Engl* 2017 Oct 27 [Epub ahead of print]; DOI: 10.1002/anie.201709201.
3. Barrangou R, Horvath P. A decade of discovery: CRISPR functions and applications. *Nat Microbiol* 2017;2:17092. DOI: 10.1038/nmicrobiol.2017.92
4. Koonin EV, Makarova KS, Zhang F. Diversity, classification and evolution of CRISPR-Cas systems. *Curr Opin Microbiol* 2017;37:67–78. DOI: 10.1016/j.mib.2017.05.008.
5. Shmakov S, Smargon A, Scott D, et al. Diversity and evolution of class 2 CRISPR-Cas systems. *Nat Rev Microbiol* 2017;15:169–182. DOI: 10.1038/nrmicro.2016.184.
6. Makarova KS, Zhang F, Koonin EV. SnapShot: class 1 CRISPR-Cas systems. *Cell* 2017;168:946–946.e1. DOI: 10.1016/j.cell.2017.02.018.
7. Makarova KS, Zhang F, Koonin EV. SnapShot: class 2 CRISPR-Cas systems. *Cell* 2017;168:328–328.e1. DOI: 10.1016/j.cell.2016.12.038.
8. Shmakov S, Abudayyeh OO, Makarova KS, et al. Discovery and functional characterization of diverse class 2 CRISPR-Cas systems. *Mol Cell* 2015;60:385–397. DOI: 10.1016/j.molcel.2015.10.008.
9. Burstein D, Harrington LB, Strutt SC, et al. New CRISPR-Cas systems from uncultivated microbes. *Nature* 2017;542:237–241. DOI: 10.1038/nature21059.
10. Makarova KS, Wolf YI, Alkhnbashi OS, et al. An updated evolutionary classification of CRISPR-Cas systems. *Nat Rev Microbiol* 2015;13:722–736. DOI: 10.1038/nrmicro3569.
11. Bland C, Ramsey TL, Sabree F, et al. CRISPR recognition tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. *BMC Bioinformatics* 2007;8:209. DOI: 10.1186/1471-2105-8-209.
12. Grissa I, Vergnaud G, Pourcel C. CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. *Nucleic Acids Res* 2007;35:W52–57. DOI: 10.1093/nar/gkm360.
13. Alkhnbashi OS, Costa F, Shah SA, et al. CRISPRstrand: predicting repeat orientations to determine the crRNA-encoding strand at CRISPR loci. *Bioinformatics* 2014;30:i489–496. DOI: 10.1093/bioinformatics/btu459.
14. Alkhnbashi OS, Shah SA, Garrett RA, et al. Characterizing leader sequences of CRISPR loci. *Bioinformatics* 2016;32:i576–i585. DOI: 10.1093/bioinformatics/btw454.
15. Wei Y, Chesne MT, Terns RM, et al. Sequences spanning the leader-repeat junction mediate CRISPR adaptation to phage in *Streptococcus thermophilus*. *Nucleic Acids Res* 2015;43:1749–1758. DOI: 10.1093/nar/gku1407.
16. Toms A, Barrangou R. On the global CRISPR array behavior in class I systems. *Biol Direct* 2017;12:20. DOI: 10.1186/s13062-017-0193-2.
17. Biswas A, Staals RH, Morales SE, et al. CRISPRDetect: a flexible algorithm to define CRISPR arrays. *BMC Genomics* 2016;17:356. DOI: 10.1186/s12864-016-2627-0.
18. Fonfara I, Le Rhun A, Chylinski K, et al. Phylogeny of Cas9 determines functional exchangeability of dual-RNA and Cas9 among orthologous type II CRISPR-Cas systems. *Nucleic Acids Res* 2014;42:2577–2590. DOI: 10.1093/nar/gkt1074.
19. Chylinski K, Le Rhun A, Charpentier E. The tracrRNA and Cas9 families of type II CRISPR-Cas immunity systems. *RNA Biol* 2013;10:726–737. DOI: 10.4161/rna.24321.
20. Carte J, Christopher RT, Smith JT, et al. The three major types of CRISPR-Cas systems function independently in CRISPR RNA biogenesis in *Streptococcus thermophilus*. *Mol Microbiol* 2014;93:98–112. DOI: 10.1111/mmi.12644.
21. Prins P, de Ligt J, Tarasov A, et al. Toward effective software solutions for big biology. *Nat Biotechnol* 2015;33:686–687. DOI: 10.1038/nbt.3240.
22. Peng RD. Reproducible research in computational science. *Science* 2011;334:1226–1227. DOI: 10.1126/science.1213847.
23. Gruning BA, Rasche E, Rebolledo-Jaramillo B, et al. Jupyter and Galaxy: easing entry barriers into complex data analyses for biomedical researchers. *PLoS Comput Biol* 2017;13:e1005425. DOI: 10.1371/journal.pcbi.1005425.
24. Shen H. Interactive notebooks: sharing the code. *Nature* 2014;515:151–152. DOI: 10.1038/515151a.
25. Silver A. Software simplified. *Nature* 2017;546:173–174. DOI: 10.1038/546173a.
26. Boettiger C. An introduction to Docker for reproducible research. *SIGOPS Oper Syst Rev* 2015;49:71–79. DOI: 10.1145/2723872.2723882.
27. HHMI. hmmscan: search sequence(s) against a profile database. http://hmmer.org/ (last accessed December 2017).
28. Beloglazova N, Petit P, Flick R, et al. Structure and activity of the Cas3 HD nuclease MJ0384, an effector enzyme of the CRISPR interference. *EMBO J* 2011;30:4616–4627. DOI: 10.1038/emboj.2011.377.
29. Mulepati S, Bailey S. Structural and biochemical analysis of nuclease domain of clustered regularly interspaced short palindromic repeat (CRISPR)-associated protein 3 (Cas3). *J Biol Chem* 2011;286:31896–31903. DOI: 10.1074/jbc.M111.270017.
30. Cady KC, O'Toole GA. Non-identity-mediated CRISPR-bacteriophage interaction mediated via the Csy and Cas3 proteins. *J Bacteriol* 2011;193:3433–3445. DOI: 10.1128/JB.01411-10.

31. Sinkunas T, Gasiunas G, Fremaux C, et al. Cas3 is a single-stranded DNA nuclease and ATP-dependent helicase in the CRISPR/Cas immune system. *EMBO J* 2011;30:1335–1342. DOI: 10.1038/emboj.2011.41.
32. Deltcheva E, Chylinski K, Sharma CM, et al. CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. *Nature* 2011;471:602–607. DOI: 10.1038/nature09886.
33. Gasiunas G, Barrangou R, Horvath P, et al. Cas9-crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria. *Proc Natl Acad Sci U S A* 2012;109:E2579–2586. DOI: 10.1073/pnas.1208507109.
34. Sapranauskas R, Gasiunas G, Fremaux C, et al. The *Streptococcus thermophilus* CRISPR/Cas system provides immunity in *Escherichia coli*. *Nucleic Acids Res* 2011;39:9275–9282. DOI: 10.1093/nar/gkr606.
35. Walker FC, Chou-Zheng L, Dunkle JA, et al. Molecular determinants for CRISPR RNA maturation in the Cas10–Csm complex and roles for non-Cas nucleases. *Nucleic Acids Res* 2017;45:2112–2123. DOI: 10.1093/nar/gkw891.
36. Hale CR, Cocozaki A, Li H, et al. Target RNA capture and cleavage by the Cmr type III-B CRISPR-Cas effector complex. *Genes Dev* 2014;28:2432–2443. DOI: 10.1101/gad.250712.114.
37. Ichikawa HT, Cooper JC, Lo L, et al. Programmable type III-A CRISPR-Cas DNA targeting modules. *PLoS One* 2017;12:e0176221. DOI: 10.1371/journal.pone.0176221.
38. Zetsche B, Gootenberg JS, Abudayyeh OO, et al. Cpf1 is a single RNA-guided endonuclease of a class 2 CRISPR-Cas system. *Cell* 2015;163:759–771. DOI: 10.1016/j.cell.2015.09.038.
39. Wu D, Guan X, Zhu Y, et al. Structural basis of stringent PAM recognition by CRISPR-C2c1 in complex with sgRNA. *Cell Res* 2017;27:705–708. DOI: 10.1038/cr.2017.46.
40. Liu L, Chen P, Wang M, et al. C2c1–sgRNA complex structure reveals RNA-guided DNA cleavage mechanism. *Mol Cell* 2017;65:310–322. DOI: 10.1016/j.molcel.2016.11.040.
41. Yang H, Gao P, Rajashankar KR, et al. PAM-dependent target DNA recognition and cleavage by C2c1 CRISPR-Cas endonuclease. *Cell* 2016;167:1814–1828.e12. DOI: 10.1016/j.cell.2016.11.053.
42. Cox DBT, Gootenberg JS, Abudayyeh OO, et al. RNA editing with CRISPR-Cas13. *Science* 2017;358:1019–1027. DOI: 10.1126/science.aaq0180.
43. Abudayyeh OO, Gootenberg JS, Essletzbichler P, et al. RNA targeting with CRISPR-Cas13. *Nature* 2017;550:280–284. DOI: 10.1038/nature24049.
44. Abudayyeh OO, Gootenberg JS, Konermann S, et al. C2c2 is a single-component programmable RNA-guided RNA-targeting CRISPR effector. *Science* 2016;353:aaf5573. DOI: 10.1126/science.aaf5573.
45. Grissa I, Vergnaud G, Pourcel C. The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats. *BMC Bioinformatics* 2007;8:172. DOI: 10.1186/1471-2105-8-172.
46. Burstein D, Sun CL, Brown CT, et al. Major bacterial lineages are essentially devoid of CRISPR-Cas viral defence systems. *Nat Commun* 2016;7:10613. DOI: 10.1038/ncomms10613.
47. Sun CL, Thomas BC, Barrangou R, et al. Metagenomic reconstructions of bacterial CRISPR loci constrain population histories. *ISME J* 2016;10:858–870. DOI: 10.1038/ismej.2015.162.

Received for publication December 20, 2017;
Revised January 23, 2018;
Accepted February 18, 2018.