The CRISPR Journal
Volume 2, Number 1, 2019
Mary Ann Liebert, Inc.
DOI: 10.1089/crispr.2018.0046

The CRISPR Journal

RESEARCH ARTICLE

CRISPRs for Strain Tracking and Their Application to Microbiota Transplantation Data Analysis

Tony J. Lam and Yuzhen Ye*

Abstract

CRISPR-Cas systems are adaptive immune systems naturally found in bacteria and archaea. Prokaryotes use these immune systems to defend against invaders, which include phages, plasmids, and other mobile genetic elements. Relying on the integration of spacers derived from invader sequences (protospacers) into CRISPR loci (forming spacers flanked by repeats), CRISPR-Cas systems are able to store the memory of past immunological encounters. While CRISPR-Cas systems have evolved in response to invading mobile genetic elements, invaders have also developed mechanisms to avoid detection. As a result of an arms race between CRISPR-Cas systems and their targets, CRISPR arrays typically undergo rapid turnover of spacers through the acquisition and loss events. Additionally, microbiomes of different individuals rarely share spacers. Here, we present a computational pipeline, CRISPRtrack, for strain tracking based on CRISPR spacer content, and we applied it to fecal transplantation microbiome data to study the retention of donor strains in recipients. Our results demonstrate the potential use of CRISPRs as a simple yet effective tool for donor-strain tracking in fecal transplantation and as a general purpose tool for quantifying microbiome similarity.

Introduction

The gut microbiome serves to provide a range of symbiotic functions, including metabolism, immune system development, and pathogen resistance.[1] While the gut microbiome plays an important role as a modulator of host health and disease, commensal colonizers are often susceptible to disruption, which has been shown to be associated with the development of disease states.[2–4] One such example is persistent and recurrent *Clostridium difficile* infection (CDI), which is often induced by the treatment of antibiotics.[5] In an attempt to increase intestinal microbial diversity and re-establish a stable microbiome, fecal microbiota transplantation (FMT) has often been prescribed as a form of treatment for patients with recurrent CDI and other gastrointestinal disorders.[6,7] The reported success rate of FMT based on thousands of patients with recurrent CDI is ~90% following one or more FMTs.[8,9] Although restoration of the gut microbiota appears to have a positive effect against gut dysbiosis that is thought to exist in FMT patients, the exact mechanisms of FMT have yet to be fully elucidated.[8] In addition to gastrointestinal disorders, recent studies have also shown promising applications of FMT to treat other types of diseases, including Parkinson's disease and autism.[10,11]

Prokaryotes constantly encounter mobile genetic elements (MGEs) such as phages and plasmids. While the exposure to MGEs may provide hosts with an adaptive advantage through the process of horizontal gene transfer events, these interactions also have wide ranging biological implications, which can disrupt a variety of the host's regulatory functions.[12–14] It is therefore unsurprising that prokaryotes have evolved various means of defending against invading MGEs.[15,16] One such defense mechanism is CRISPR and CRISPR-associated genes (Cas).[17,18] These CRISPR-Cas systems have been shown to provide prokaryotes with an adaptive immune response against the constant threat of MGEs while also providing a mechanism to acquire and retain the memory of past immunological engagements.[17–22] The actuation of CRISPR-Cas systems consists of three stages: (1) adaptation, in which new spacers are derived from target sequences of invading

School of Informatics, Computing, and Engineering, Indiana University, Bloomington, Indiana.

*Address correspondence to: *Yuzhen Ye, School of Informatics, Computing, and Engineering, Indiana University, Bloomington, IN 47408,* E-mail: yye@indiana.edu

mobile genetic elements known as protospacers and integrated into the CRISPR loci. forming the immunological memory of CRISPR-Cas systems; (2) expression and processing, where the CRISPR array is transcribed into mature CRISPR RNAs (crRNAs); and (3) interference, where Cas-crRNA complexes scan, target, and cleave foreign nucleic acids of matching complementary sequences.[22,23]

In addition to spacer acquisition, there is also spacer loss, which aids the turnover of spacers. While the mechanisms of spacer loss events have yet to be fully understood, aside from random loss events, several hypotheses have been proposed to suggest that spacer loss events may be related to functional mechanisms. One proposed hypothesis is based on the underlying immunity granted through the positionality of spacers within the spacer array, where it has been shown that positionality of spacers provides an optimization of immune response through differential expression of crRNAs against the most recent invader.[24] Alternatively, it has also been proposed that deletion events may be related to the maintenance of the CRISPR array through the removal of old, less utilized spacers, thus allowing room for the addition of more relevant spacers.[25–27] Homologous recombination events of CRISPRs have also been observed and proposed as a mechanism for the acquisition and deletion of spacers.[25,28,29]

While prokaryotes have evolved CRISPR-Cas systems to target foreign genetic elements, invaders have also evolved to evade CRISPR-Cas systems (e.g., through localized protospacer adjacent motif mutations and anti-CRISPR proteins).[30,31] The constant evolution of the CRISPR evasion tactics of MGEs have been proposed to be a significant contributor to the extreme diversification of Cas genes and the variety of CRISPR-Cas systems.[32–35] As such, the co-evolution of CRISPR-Cas systems and their targets illustrates the ever-perpetual arms race shaping the predator–prey dynamics of these systems.

Spacer acquisition, as well as spacer turnover, highlights the ever-evolving nature of CRISPR arrays. Similarly to bacteria found throughout the environment, microbial organisms found within the human microbiome (which mostly comprise of eubacteria) also carry CRISPR arrays that are dynamic in nature. With the constantly changing composition of the CRISPR loci, gut microbiota found within human individuals often bear CRISPR arrays with unique spacer sequences. By taking advantage of the properties of spacer acquisition and retention within CRISPRs, the CRISPR spacers can potentially be used as molecular markers for typing and strain-level species tracking purposes.[36]

While the underlying dynamics and mechanisms of FMT remain largely undiscovered, efforts have been made to unveil these details by first understanding the impact of FMT at a microbial ecology level. To understand the effects of FMT induced microbiome reconstruction, it becomes important to understand the success of bacterial engraftment following fecal transplantation. Several studies have shown the success of utilizing single-nucleotide variants (SNV)-based methods for tracking the dynamics between donor and recipient microbiomes following FMT.[37–39] The study by Li *et al.* on metabolic syndrome patients using SNV in metagenomes enabled the quantification of the extent of donor microbiota colonization after FMT, revealing extensive coexistence of donor and recipient strains, persisting 3 months after treatment. The authors also found that same-donor recipients displayed varying degrees of microbiota transfer, indicating individual patterns of microbiome resistance and donor–recipient compatibilities. Smillie *et al.* developed StrainFinder,[39] a tool to infer strain genotypes based on detected SNVs from FMT microbiomes and track strains over time. The successful usage of SNV-based methods highlights the importance of understanding the microbial ecology on a strain level. However, SNV calling in metagenomics can be complicated by the uneven abundance distribution of the bacterial species living in the same community and the coexistence of closely related species. Even worse, there is currently no strict definitions of what constitutes a bacterial or archaeal strain.[40]

Previous studies have explored the utilization of CRISPR arrays as a means to study community dynamics between microbiomes and their exposure to MGEs, including studies that utilized CRISPRs to track environmental and human microbiomes over time to uncover dynamics relating to the selective pressures of MGEs and CRISPR evolution.[41–46] Here, we propose the use of CRISPR arrays to study and track donor-strain retention in fecal transplantations and quantify microbiome similarity through spacer content. We developed a pipeline, CRISPRtrack, which takes advantage of the unique and subject specific spacers for the quantification of donor-strain retention in FMT recipient and leverages tools we have previously developed for the identification and characterization of CRISPR-Cas systems in metagenomes.[47,48] As compared to SNVs, spacers are relatively large entities of approximately 20–50 bps long. Thus, they are more straightforward to characterize—an apparent advantage of using CRISPR spacers for strain tracking. Although using CRISPR spacers has its caveats (for example, some genomes do not contain CRISPR-Cas systems), by applying our tool CRISPRtrack to two fecal microbiota transplantation data sets, we are able to demonstrate the potential use CRISPR spacers as a simple yet effective tool for donor-strain tracking in fecal transplantation.

Methods

Identification of CRISPR arrays in genomes and metagenomes

In this study, we utilized two approaches that we previously developed for the identification of CRISPR arrays from genomes and metagenome assemblies: a reference-based approach and a *de novo*–based approach. Contigs assembled from shotgun metagenome sequences were used as input for identification of CRISPR arrays in both approaches. Our previous study showed that CRISPR arrays, in some instances, have been difficult to assemble from shotgun metagenomics sequencing data, partly due to the presence of repetitive regions of CRISPR arrays.[47] However, we believe that these challenges are gradually resolving themselves through the constant improvements of sequencing quality and read length, as well as improvements in assembly techniques. We have since found that metagenomics specific assemblers such as MEGAHIT[49] have improved assembly capabilities for the identification of CRISPR-Cas systems comparatively to previous assemblers. Additionally, we have also demonstrated improvements of CRISPR array prediction using varying k-mer sizes during assembly.[50] As a result, this study utilized MEGAHIT to assemble metagenomes using k-mer size parameters (k-list = 21, 41, 61, 81, 99) and applied the assembled contigs to downstream analyses.

Specifically, for *de novo*–based prediction, CRISPR arrays were predicted from contigs utilizing our previously developed tool, CRISPRone.[48] CRISPRone utilizes metaCRT,[47] a modified variant of CRT,[51] for the initial identification of putative CRISPR arrays. The software metaCRT utilizes metagenomics assemblies and exploits the structure of CRISPR arrays to search for sequences containing repeat-spacer-like structures, and improves upon CRT by considering the inclusion of spacers flanked by an incomplete repeat. Following the prediction of putative CRISPR arrays, CRISPRone applies additional filters for the removal of suspicious CRISPR arrays that may have been erroneously identified by metaCRT through the identification of structures that constitute false-positive arrays (e.g., tandem repeats, STAR-like elements, and simple repeats).[52,53]

In our reference-based approach, we use a set of CRISPR repeat sequences associated with human gut microbiomes to be used as reference repeats. Utilizing this set of reference repeats with CRISPRAlign,[47] we are able to identify CRISPRs that share similar repeats as our reference repeats. Comparatively to *de novo* approaches, reference-based methods hold an advantage by using a set of carefully curated set of reference repeats, which reduces the chances of false-positive CRISPR arrays. Contrastingly, as reference-based approaches only search for CRISPRs sharing similar reference repeats, reference-based methods may result in an under-represented set of CRISPRs by missing any CRISPRs that may have dissimilar CRISPR repeats.

There are other approaches to reduce false CRISPR arrays. For example, Sorokin *et al.* revealed a large discrepancy in the prediction results when different approaches were used, and they only used the consistent predictions for downstream analyses. We also tested other approaches, including minCED and Crass,[54] and the consensus approach to test the impact of CRISPR array prediction approaches on the performance of our pipeline.

Quantification of the existence of donor species in recipient using CRISPR spacer contents

Using predicted CRISPR arrays, all corresponding putative CRISPR spacers were extracted. Utilizing the set of all extracted spacers, cd-hit-est[55] (-c 0.9) was used to cluster spacers into their representative groups to remove redundancy (the clusters of spacers are henceforth referred to as ''spacer cluster(s)''). Each spacer cluster represents a unique spacer, and all spacers grouped into the same cluster were considered the same in order to allow mismatches in the spacers and potential sequencing errors. The presence of donor species within recipient samples at a given time point, r_t, can be quantified by computing the sharing of spacers between the recipient microbiome and donor microbiome as following:

$$r_t = \frac{2S_c}{2S_r + S_d}$$

where S_c denotes the number of clusters containing spacers from both the recipient and the donor sample, S_r denotes the number of clusters containing only spacers from the recipient, and S_d denotes the number of clusters containing only spacers from the donor.

Fecal transplantation data sets and the HMP data sets

We utilized three sets of previously published data sets to test our new method. For clarity, we denote these data sets as FMT-Li,[38] FMT-Smillie,[39] and HMP.[56] FMT-Li data sets[38] were downloaded from the European Nucleotide Archive (ENA) under accession number PRJEB12357. The FMT-Li data sets include metagenomic sequencing data from five patients (FMT1–FMT5) receiving microbiota transplantation from three healthy donors (Don1–Don3), in which stool samples were collected from patients at multiple time points spanning between pretreatment to 84 days post treatment (the patients did not receive antibiotics or other medication before FMT). FMT-Smillie data sets[39] were downloaded from the ENA under accession number PRJEB23524. The FMT-Smillie data sets consist of microbiome sequencing data obtained from four donors, and 19 patients whose stool samples were collected at multiple time points spanning between pretreatment to 135 days post treatment (unlike the FMT-Li patients, the patients in this cohort did receive antibiotics treatments prior to FMT). We note this data set includes data from participants with only one or two microbiome samples post FMT. To exemplify the application of CRISPRtrack to microbiota transplantation data, we focus on the analysis of a subset of the individuals, each with at least four microbiome samples. HMP data sets[56] were obtained from the Human Microbiome Project Data Analysis and Coordination Center Web site. We used a total of 95 stool microbiome data sets from the HMP collection.

Availability of the software and results

We implemented a package called CRISPRtrack for identification of CRISPRs in metagenome assemblies and quantification of the retention of donor species in recipients. The package is available for download at SourceForge. The package contains a few tools that we previously developed and additional scripts that we developed for this study for computing the sharing of the CRISPR spacers. The package also includes tools for further analyses and visualizations of the results. CRISPRtrack supports both approaches for characterizing CRISPRs: the reference-based approach using gut microbiome-related reference repeats (called CRISPRtrack-ref) and *de novo* prediction by CRISPRone (CRISPRtrack-denovo). The package outputs spacer-subject tables, similarity scores between microbiomes based on spacer content sharing, and tracking plots of donor strains in recipients based on CRISPR spacer sharing.

We also made available the results from this work on our supplementary Web site* for this work, including the sequences of the reference CRISPR repeats and CRISPRtrack results for the FMT-Li and FMT-Smillie data sets.

Results

CRISPRs in common gut microbial species

To define a set of reference CRISPR repeats to be used in the reference-based identification of CRISPR arrays, we characterized a set of high confidence CRISPR-Cas systems associated with common gut microbiomes. We first checked 42 common strains found in the fecal microbiota samples in the FMT-Li data set. In our previous study,[50] we analyzed a different cohort of gut microbiome data sets,[57] from which we were able to identify 33 unique CRISPR repeats. Combining the two subsets of CRISPR repeats, we were able to compile a collection of 64 unique CRISPR repeats to be used as reference repeats. CRISPR arrays that were used to generate reference repeats were predicted by CRISPRone and curated manually through the identification of high-confidence CRISPR arrays, which appeared alongside a complete *cas* gene cassette. Figure 1 shows a subset of CRISPR-Cas systems that was utilized to compile the set of reference CRISPR repeats. Unsurprisingly, most of the CRISPR-Cas systems found in the gut-associated microbial genomes belonged to type I CRISPR-Cas systems. Additionally, we found type II, III, and V CRISPR-Cas systems in these genomes. The sequences of the reference repeats are included in the CRISPRtrack package and are available on the supplementary Web site.

Spacer sharing between human individuals (the baseline)

To determine the baseline of shared spacers among different individuals, we employed the HMP data set involving 95 microbiome samples derived from 79 human subjects (see a table on our supplementary Web site listing the metadata of the individuals and microbiome samples). Both *de novo* and reference-based approaches were used to identify CRISPR spacers in the HMP data sets. Spacers predicted from both methods were then clustered and used to estimate the baseline of spacers shared between individuals.

Using the referenced-based approach, a total of 26,074 spacer clusters were identified from the HMP data sets. Comparing the shared spacers between different individuals, we show that gut microbiomes from different individuals share significantly fewer spacers compared to gut microbiomes from the same individuals but at different time points. The median spacer content similarity between different individuals and the same individuals at different time points were 0.0049 and 0.66, respectively (Fig. 2A). To study spacer sharing further among unrelated individuals, we focused our analysis on individuals with more than one sample. This reduced the number of spacer clusters to 23,868, among which 19,634 ($\sim$82%) spacer clusters contained only a single spacer, indicating the spacer is unique to the individual. From the spacer clusters, only 63 (0.26%) clusters were found to be shared among ≥10 individuals. Through this analysis, we were able to show that while a small set of spacers are shared among different individuals, a large portion of spacers remain unique to individual subjects and are not shared with others. Spacers shared by many individuals are likely to originate from inactive orphaned CRISPR arrays, which has been shown not to exhibit active turnover of spacers.[29,58,59]

*http://omics.soic.indiana.edu/CRISPRtrack

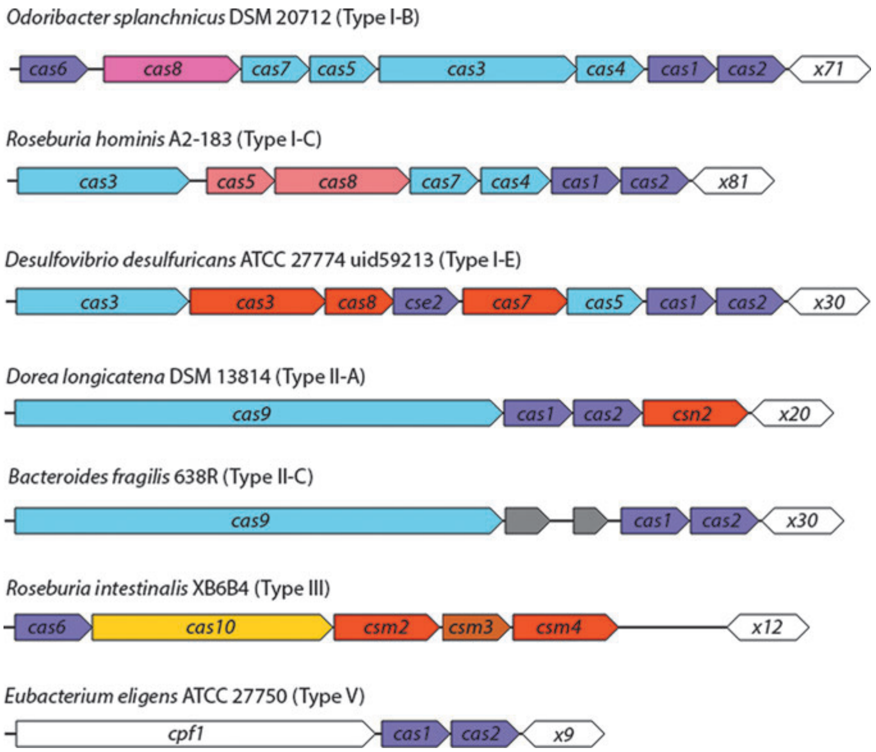
Odoribacter splanchnicus DSM 20712 (Type I-B)
cas6
cas8
cas7
cas5
cas3
cas4
cas1
cas2
x71
Roseburia hominis A2-183 (Type I-C)
cas3
cas5
cas8
cas7
cas4
cas1
cas2
x81
Desulfovibrio desulfuricans ATCC 27774 uid59213 (Type I-E)
cas3
cas3
cas8
cse2
cas7
cas5
cas1
cas2
x30
Dorea longicatena DSM 13814 (Type II-A)
cas9
cas1
cas2
csn2
x20
Bacteroides fragilis 638R (Type II-C)
cas9
cas1
cas2
x30
Roseburia intestinalis XB6B4 (Type III)
cas6
cas10
csm2
csm3
csm4
x12
Eubacterium eligens ATCC 27750 (Type V)
cpf1
cas1
cas2
x9

FIG. 1. Subset of CRISPR-Cas systems found in common human gut-associated bacterial species used to generate reference CRISPR repeats. The arrows in different colors represent the *cas* genes, and the open hexagons represent the CRISPR repeat-spacer arrays, with numbers following the letter *x* indicating the number of repeats found in each array.

Similarly, we analyzed the CRISPR spacer sharing between individuals using spacers predicted through the *de novo* approach (i.e., CRISPRone). The *de novo* approach predicted much more CRISPR spacers compared to the reference-based approach, resulting in a total of 48,275 spacer clusters. We note that CRISPRone applied a few steps to remove likely false CRISPR arrays predicted *de novo*, and Supplementary Table S1 shows the number of spacers predicted before and after applying the filtering steps. The difference seen between the *de novo* and reference-based methods is expected, as the reference-based approach only identified CRISPR arrays containing repeats similar to the reference repeats, whereas the CRISPR arrays identified through the *de novo* approach is not limited to CRISPR arrays sharing the same repeats as the reference repeats. Reassuringly, analysis of *de*

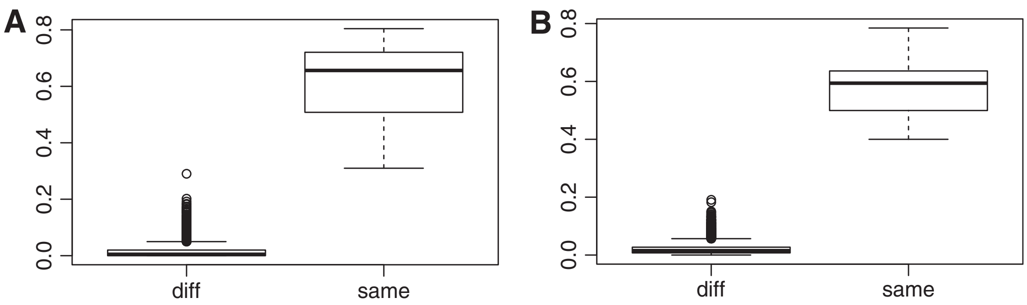

FIG. 2. Sharing of the CRISPR spacers among HMP individuals. These two figures show the boxplots of spacer content similarity between microbiomes from different individuals (diff) and microbiomes from the same human subject (same). **(A)** Based on spacers identified using the reference-based approach. **(B)** Based on *de novo* identification of CRISPR arrays.

novo predicted spacers revealed consistent results that microbiomes from different individuals shared few CRISPR spacers, whereas the microbiomes of the same individual shared substantially more spacers, as seen in Figure 2B.

Taking the pairwise comparison of spacer similarities across all spacers from each individual, we calculated the proportion of spacers shared between different individuals. Using the distribution of shared spacers, we calculated the 95th percentile for the proportion of spacers shared among different HMP individuals. The 95th percentiles were 0.062 (6.2%) and 0.056 (5.6%) for spacers derived by the reference-based approach and *de novo* approach, respectively, and represent the percentage of spacers shared >95% of the pairwise comparisons for HMP individuals. Our empirical estimates of spacer similarity within the HMP data set are consistent with other previous studies in that CRISPR repositories are observed to be mostly individual specific and sharing relatively few spacers between different individuals.[42,60,61] The calculated 95th percentiles were used as the baseline CRISPR spacer similarities for examining the sharing of spacers between FMT recipients and their donors in our analyses of the FMT-Li data set and the FMT-Smillie data set.

Application of CRISPRtrack to FMT-Li data set

We applied CRISPRtrack to characterize the CRISPR spacers from the donor and recipient microbiome data (FMT-Li), and quantified the retention of donor CRISPR spacers in recipients using the CRISPR spacer content. A total of 30,271 spacers and 16,091 spacer clusters were characterized in this data set using the *de novo* CRISPR array characterization approach (see Supplementary Tables S2 and S3 for details and comparison). The spacer similarity plots (Fig. 3; based on spacers predicted *de novo*) shows that the recipient microbiome contains similar CRISPR spacers as the donor microbiome, especially during the early time points post FMT, indicating that a significant amount of donor-sourced bacteria were transferred into the recipient and retained for that period of time. Our results also show that after significant reduction of the recipient's own bacteria (as indicated as the low spacer similarity between recipient and its pre-FMT microbiome), there is a resurgence of the recipient's original strains (Fig. 3B). We note using spacers predicted by the reference-based approach showed consistent results with the *de novo* approach.

Below, we highlight a few comparisons of our results with the results from the previously reported SNV-based analysis.[38]

- It was mentioned that marked differences in colonization success were observed between allogenic recipients whom shared a donor (FMT1, FMT2, and FMT3). Three months after treatment, FMT2 and FMT3 retained a higher amount of donor-specific SNVs compared to FMT1 (46.1%, 56.6%, and

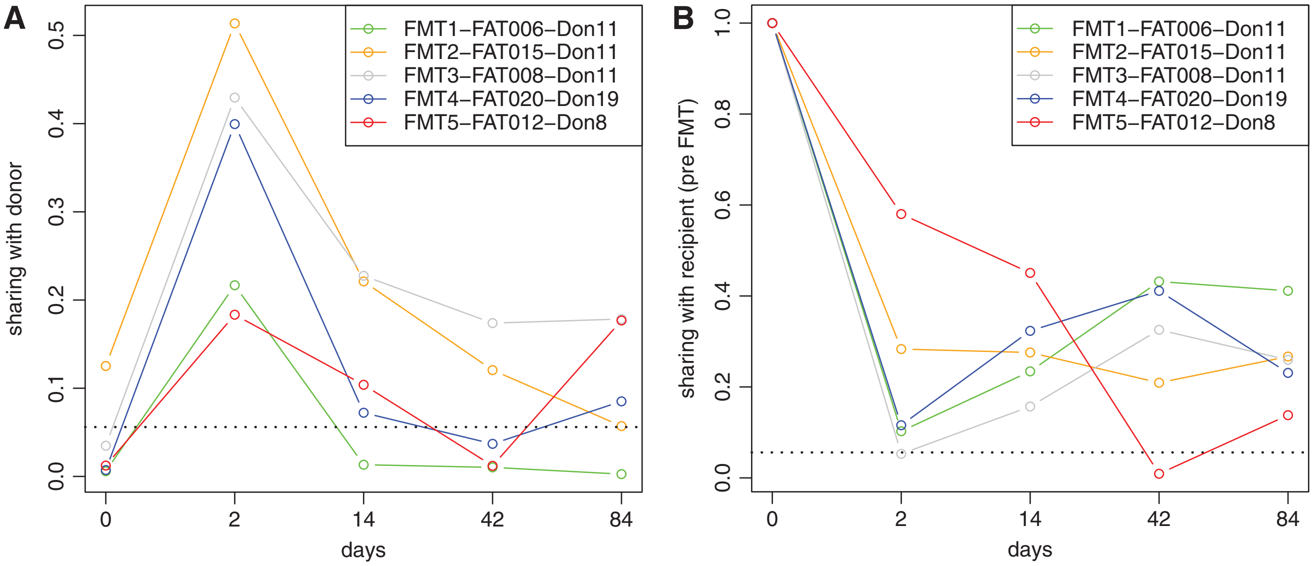

FIG. 3. Tracking of donor spacers and recipient's own spacers over time after fecal microbiota transplantation (FMT). **(A)** CRISPR spacer similarity between the recipients and their corresponding donors. **(B)** CRISPR spacer similarity between the recipients after FMT and their pre-FMT counterparts. Lines connect time-point samples from the same individual. The time axis (i.e., the *x*-axis) was not scaled for clarity. The dotted black lines in the plots indicate the 95th percentiles of the spacer similarities between different individuals, inferred from the HMP data sets.

12.0%, respectively). CRISPRtrack revealed a similar trend: FMT1 retained the fewest donor-specific spacers, whereas FMT2 and FMT3 retained more.

- The SNV-based analyses showed the highest retention of donor strains in FMT2 and FMT3; CRISPRtrack revealed the same trend.
- The SNV-based analyses showed the resurgence of donor-specific strains in FMT5 after day 14; our analysis showed a similar trend, although the resurgence peaked on day 84 instead of day 42.

In our analysis, we observed that FMT2 (prior to treatment, i.e., day 0) shares relatively more common CRISPR spacers with the donor microbiome compared to other recipients (see Fig. 3A). We also examined the spacers identified using the reference-based approach to study the spacers that are common to donor and recipient. For example, CRISPR arrays containing repeats that are almost identical (with 1 bp difference) to the repeat identified in the *Bacteroides fragilis* 638R genome (named BfragL47-II) are found to share five spacers. The sequence of BfragL47-II is GTTGTGATTTGCTTTCA AATTAGTATCTTTGAACCATTGGAAACAGC, and at position 33 the repeat identified in the donor and recipient CRISPRs have T instead of A. Another example is CRISPR arrays containing repeats that were identical to the CRISPR repeat found in reference genome *Alistipes shahii* WAL8301 (named AshahL36 whose repeat sequence is GTTGTGGTTTGATGTAGAATTTCGATAA GATACAAC). Interestingly, the CRISPR identified in *A. shahii* is an orphan CRISPR without *cas* genes. An array containing AshahL36 repeats was assembled from the recipient data set, which contains nine spacers, a subset of the 36 spacers in a much longer array assembled from the donor data set.

Application of CRISPRtrack to FMT-Smillie data set

We applied CRISPRtrack to a more recent FMT microbiome data set (FMT-Smillie).[39] Although this data set involved more human subjects than the earlier mentioned FMT data set (FMT-Li),[38] among the individuals, only five each had microbiome samples from at least four time points. We also note that the samples were collected at varying time intervals, unlike the FMT-Li data set, which sampled all recipients at regular time intervals. Another major difference between FMT-Smillie and FMT-Li data sets is that most recipients in the FMT-Smillie study received antibiotic treatments prior to FMT.

As an example, Figure 4A shows the tracking of *de novo* predicted CRISPR spacers in samples from recipient MGH02R. The patient's pre-FMT microbiome apparently contained fewer bacterial species compared to post-FMT microbiomes: the pre-FMT microbiome assembly contained approximately 85 Mbp, whereas the post FMT (day 2) microbiome assembly contained approximately 396 Mbp. As a result, isolates from the FMT-Smillie data set exhibit few commonly shared spacers between pre- and post-FMT samples (Fig. 4A). By contrast, the pre- and post-FMT microbiomes in the FMT-Li data set share significantly more spacers,

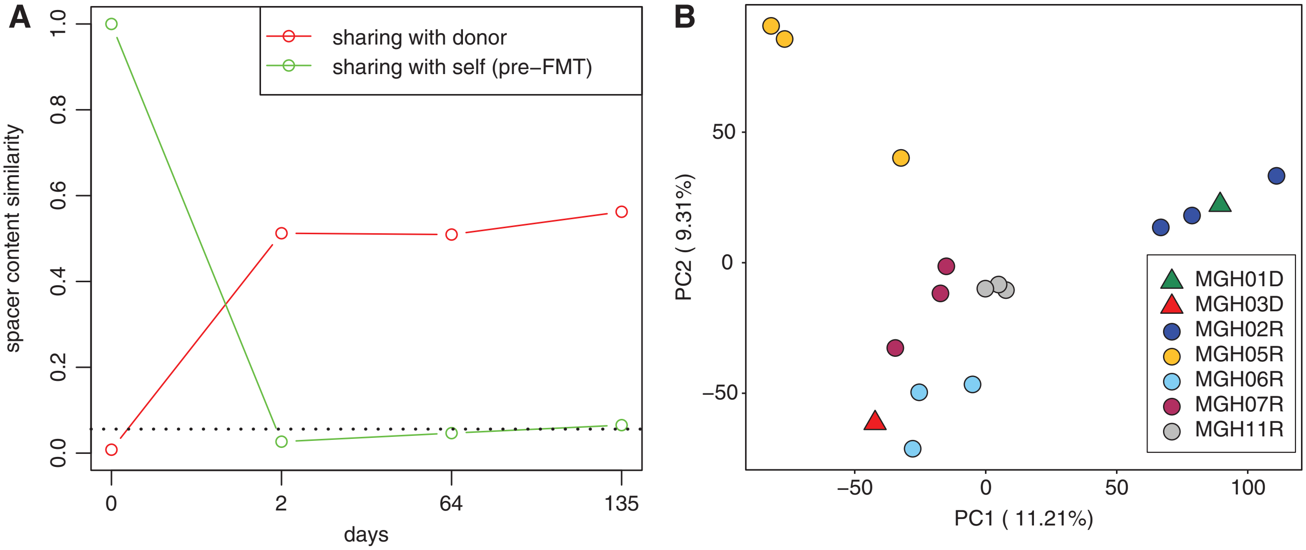

FIG. 4. CRISPRtrack results for the FMT-Smillie data set. **(A)** Tracking of the donor- and pre-FMT recipient-specific spacers in recipient MGH02R. **(B)** Principle component plot of principle component 1 and principle component 2. Donor samples are depicted as triangles, whereas recipients are depicted as circles. Samples are colored by subject.

which was likely due to FMT-Li patients not receiving antibiotic treatment prior to FMT (Fig. 3).

Focusing on the data involving two donors (MGH01D and MGH03D) from the FMT-Smillie data set, we performed a two-dimensional principal component analysis (PCA) on the spacer-subject table produced by CRISPRtrack. The two principle components that contributed the most to the variance explained, principle component 1 (PC1) and principle component 2 (PC2), were plotted in Figure 4B. PC1 separates the samples in relation to the donor, whereas PC2 seems to separate out samples further by subject. Unsurprisingly, the largest contributors to the variance explained are those related to the donor and subjects. We note for the same reason (antibiotic treatment prior to FMT) that pre-FMT microbiomes have very low bacterial diversity. We did not include the pre-FMT samples in this analysis. Recipient samples MGH06R (azure blue), MGH07R (brown), and MGH11R (gray) received FMT from MGH03D donor samples and cluster together. Similarly, MGH02R (blue), which received samples from donor MGH01D (green), clusters together. However, MGH05R (orange), which received FMT from MGH03D, shared the least CRISPR spacers compared to its donor.

Robustness of CRISPRtrack

Finally, we asked if CRISPRtrack would provide consistent results if different CRISPR identification tools were used. We compared CRISPRtrack against two other recently developed CRISPR array identification tools—minCED and Crass[54]—and compared the effectiveness of utilizing spacer predicted from each respective software for strain tracking.

Among the three approaches, minCED and our approach produced comparable numbers of spacers. Crass produced significantly fewer spacers for the FMT-Li data set. However, Crass produced much more spacers in the FMT-Smillie data set (see Supplementary Tables S2 and S3) compared to minCED and our approach. Supplementary Figure S1 shows the Venn diagram of spacer clusters shared among the three approaches in both FMT-Li and FMT-Smillie data sets. The differences of the spacer identification results are hardly surprising and have been reported before,[41] indicating that characterization of CRISPR arrays is still challenging, due to the difficulty of the problem (confusing the tandem repeats or other low-composition repeats as CRISPR arrays) and the limitation of the microbiome data (reads and assemblies are fragmented).

Reassuringly, using spacers derived by minCED and Crass and also the consensus spacers (predicted by all three methods) showed similar trends as compared to the results from CRISPRone, which is included in CRISPRtrack (see Supplementary Figs. S2, S3, and S4). This indicates that CRISPR spacers provide a robust way for tracking strains in experiments such as the FMT, which is to some extent tolerant to false-positives and false-negatives in spacer prediction.

Discussion

Our study shows the potential of using CRISPRs for tracking the engraftment of CRISPR containing donor strains in recipients following FMT. CRISPRtrack provides two approaches for spacer identification: reference based and *de novo*. We expect that the reference-based approach will be suitable for studying microbiota from well studied environments (i.e., gut microbiome), as this approach relies on the curation of reference CRISPR repeats. Alternatively, the *de novo* approach will be simpler to apply for less well studied microbiota data analysis. In terms of running time, the reference-based approach is slower than the *de novo* approach, since the reference-based approach involves an alignment step to find segments in metagenomic assemblies that are similar to reference CRISPR repeats. Still, both approaches are fast: using only a single process (Intel Xeon CPU E5-2623 v3 @ 3.00 GHz), the whole pipeline completed in 27 and 217 minutes for the *de novo* and reference-based approaches, respectively, on the FMT-Li collection. The run time for the *de novo* and reference-based approaches using the FMT-Smile collection were 32 and 342 minutes, respectively. We note that although various methods, including the method reported here, have been developed for tracking donor strains in FMT recipients, predictive models used for the prediction of successful FMT based on microbiome data still remain lacking.[38,39]

CRISPRs provide a unique advantage in that they can provide a unique subset of spacers (some of which might be found in inactive CRISPR arrays) that can be utilized as molecular markers, providing a high resolution approach to differentiate bacterial strains from separate individuals. While we show that CRISPR-based tracking methods hold the potential of revealing microbial community dynamics, we also acknowledge the limitations of such approaches. The fast-evolving nature of CRISPRs causes constant spacer acquisition and turnover, which limits its use to short-term tracking. As not all CRISPRs are active within a system, even with constant turnover of spacers within a system, existing spacers will have a greater contribution to the spacer diversity than newly acquired spacers. Additionally, not all prokaryotes contain CRISPR-Cas systems. Thus, CRISPR spacers cannot be used for the tracking of microbial strains that lack CRISPR-Cas systems. It is also important to note that even with the application of the aforementioned method of improving CRISPR assembly, not all CRISPRs within a sample will be assembled and thus under-represent the CRISPR count. Another caveat to consider, which applies broadly to sequencing data in general, is the potential for inclusion of genetic material from dead cells in which indistinguishable from the genetic material of living cells. However, this caveat is less problematic for time-series applications, as dead cells from donor strains that have failed to engraft to the host are unlikely to persist within subjects for an extended period of time. With these potential limitations of using CRISPR spacers in mind, we still believe that CRISPR spacers can be used to serve as sensitive molecular markers for tracking microbes, especially when we consider microbial communities as a whole.

Utilizing the methods we have developed for using CRISPRs to track donor strain retention, we can begin to consider exploring further questions about the dynamics of CRISPR spacers in FMT patients. Potential avenues of exploration may include the dynamics of CRISPR spacer turnover following FMT, as well as understanding if spacer acquisition in recipient CRISPRs can be correlated to donor microbiomes post FMT.

Acknowledgments

This work was supported by the National Institutes of Health grants 1R01AI108888 and 1R01AI143254 to Y.Y., and partially supported by the Indiana University Precision Health Initiative.

Supplementary Materials

Supplementary Figure S1
Supplementary Figure S2
Supplementary Figure S3
Supplementary Figure S4
Supplementary Table S1
Supplementary Table S2
Supplementary Table S3

References

1. Koskella B, Hall LJ, Metcalf CJE. The microbiome beyond the horizon of ecological and evolutionary theory. *Nat Ecol Evol* 2017;1:1606–1615. DOI: 10.1038/s41559-017-0340-2.
2. Wang J, Qin J, Li Y, et al. A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* 2012;490:55–60. DOI: 10.1038/nature11450.
3. Lane ER, Zisman TL, Suskind DL. The microbiota in inflammatory bowel disease: current and therapeutic insights. *J Inflamm Res* 2017;10:63–73. DOI: 10.2147/JIR.S116088.
4. Shreiner AB, Kao JY, Young VB. The gut microbiome in health and in disease. *Curr Opin Gastroenterol* 2015;31:69–75. DOI: 10.1097/MOG.0000000000000139.
5. Borody TJ, Khoruts A. Fecal microbiota transplantation and emerging applications. *Nat Rev Gastroenterol Hepatol* 2011;9:88–96. DOI: 10.1038/nrgastro.2011.244.
6. Vrieze A, Van Nood E, Holleman F, et al. Transfer of intestinal microbiota from lean donors increases insulin sensitivity in individuals with metabolic syndrome. *Gastroenterology* 2012;143:913–916. DOI: doi.org/10.1053/j.gastro.2012.06.031.
7. Brandt LJ, Aroniadis OC. An overview of fecal microbiota transplantation: techniques, indications, and outcomes. *Gastrointest Endosc* 2013;78:240–249. DOI: 10.1016/j.gie.2013.03.1329.
8. Goldenberg SD, Batra R, Beales I, et al. Comparison of different strategies for providing fecal microbiota transplantation to treat patients with recurrent *Clostridium difficile* infection in two english hospitals: a review. *Infect Dis Ther* 2018;7:71–86. DOI: 10.1007/s40121-018-0189-y.
9. Youngster I, Sauk J, Pindar C, et al. Fecal microbiota transplant for relapsing clostridium difficile infection using a frozen inoculum from unrelated donors: a randomized, open-label, controlled pilot study. *Clin Infect Dis* 2014;58:1515–1522. DOI: 10.1093/cid/ciu135.
10. Tian Z, Liu J, Liao M, et al. Beneficial effects of fecal microbiota transplantation on ulcerative colitis in mice. *Dig Dis Sci* 2016;61:2262–2271. DOI: 10.1007/s10620-016-4060-2.
11. Hsiao EY, McBride SW, Hsien S, et al. Microbiota modulate behavioral and physiological abnormalities associated with neurodevelopmental disorders. *Cell* 2013;155:1451–1463. DOI: 10.1016/j.cell.2013.11.024.
12. Jore MM, Brouns SJJ, van der Oost J. RNA in defense: CRISPRs protect prokaryotes against mobile genetic elements. *Cold Spring Harb Perspect Biol* 2012;4:1–12. DOI: 10.1101/cshperspect.a003657.
13. Heuer H, Smalla K. Plasmids foster diversification and adaptation of bacterial populations in soil. *FEMS Microbiol Rev* 2012;36:1083–1104. DOI: 10.1111/j.1574-6976.2012.00337.x.
14. Obeng N, Pratama AA, Elsas JD van. The significance of mutualistic phages for bacterial ecology and evolution. *Trends Microbiol* 2016;24:440–449. DOI: 10.1016/j.tim.2015.12.009.
15. Labrie SJ, Samson JE, Moineau S. Bacteriophage resistance mechanisms. *Nat Rev Microbiol* 2010;8:317–327. DOI: 10.1038/nrmicro2315.
16. Koonin E V, Makarova KS, Wolf YI. Evolutionary genomics of defense systems in archaea and bacteria. *Annu Rev Microbiol* 2017;71:233–261. DOI: 10.1146/annurev-micro-090816-093830.
17. Bolotin A, Quinquis B, Sorokin A, et al. Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. *Microbiology* 2005;151:2551–2561. DOI: 10.1099/mic.0.28048-0.
18. Barrangou R, Fremaux C, Deveau H, et al. CRISPR provides acquired resistance against viruses in prokaryotes. *Science* 2007;315:1709–1712. DOI: 10.1126/science.1138140.
19. Mojica FJM, Díez-Villaseñor C, García-Martínez J, et al. Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. *J Mol Evol* 2005;60:174–182. DOI: 10.1007/s00239-004-0046-3.
20. Garneau JE, Dupuis MÈ, Villion M, et al. The CRISPR/cas bacterial immune system cleaves bacteriophage and plasmid DNA. *Nature* 2010;468:67–71. DOI: 10.1038/nature09523.
21. Levy A, Goren MG, Yosef I, et al. CRISPR adaptation biases explain preference for acquisition of foreign DNA. *Nature* 2015;520:505–510. DOI: 10.1038/nature14302.
22. Amitai G, Sorek R. CRISPR-Cas adaptation: insights into the mechanism of action. *Nat Rev Microbiol* 2016;14:67–76. DOI: 10.1038/nrmicro.2015.14.
23. Marraffini LA. CRISPR-Cas immunity in prokaryotes. *Nature* 2015;526:55–61. DOI: 10.1038/nature15386.
24. McGinn J, Marraffini LA. CRISPR-Cas systems optimize their immune response by specifying the site of spacer integration. *Mol Cell* 2016;64:616–623. DOI: 10.1016/j.molcel.2016.08.038.
25. Horvath P, Romero DA, Coûté-Monvoisin AC, et al. Diversity, activity, and evolution of CRISPR loci in *Streptococcus thermophilus*. *J Bacteriol* 2008;190:1401–1412. DOI: 10.1128/JB.01415-07.
26. Lopez-Sanchez MJ, Sauvage E, Da Cunha V, et al. The highly dynamic CRISPR1 system of *Streptococcus agalactiae* controls the diversity of its mobilome. *Mol Microbiol* 2012;85:1057–1071. DOI: 10.1111/j.1365-2958.2012.08172.x.
27. Han P, Niestemski LR, Barrick JE, et al. Physical model of the immune response of bacteria against bacteriophage through the adaptive CRISPR-Cas immune system. *Phys Biol* 2013;10:025004. DOI: 10.1088/1478-3975/10/2/025004.
28. Kupczok A, Landan G, Dagan T. The contribution of genetic recombination to CRISPR array evolution. *Genome Biol Evol* 2015;7:1925–1939. DOI: 10.1093/gbe/evv113.
29. Lillestøl RK, Redder P, Garrett RA, et al. A putative viral defence mechanism in archaeal cells. *Archaea* 2006;2:59–72.

30. Deveau H, Barrangou R, Garneau JE, et al. Phage response to CRISPR-encoded resistance in *Streptococcus thermophilus*. *J Bacteriol* 2008;190:1390–1400. DOI: 10.1128/JB.01412-07.
31. Bondy-Denomy J, Garcia B, Strum S, et al. Multiple mechanisms for CRISPR-Cas inhibition by anti-CRISPR proteins. *Nature* 2015;526:136–139. DOI: 10.1038/nature15254.
32. Takeuchi N, Wolf YI, Makarova KS, et al. Nature and intensity of selection pressure on crispr-associated genes. *J Bacteriol* 2012;194:1216–1225. DOI: 10.1128/JB.06521-11.
33. Koonin EV, Wolf YI. Evolution of the CRISPR-Cas adaptive immunity systems in prokaryotes: models and observations on virus–host coevolution. *Mol Biosyst* 2015;11:20–27. DOI: 10.1039/c4mb00438h.
34. 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.
35. Van Houte S, Ekroth AKE, Broniewski JM, et al. The diversity-generating benefits of a prokaryotic adaptive immune system. *Nature* 2016;532:385–388. DOI: 10.1038/nature17436.
36. Barrangou R, Dudley EG. CRISPR-based typing and next-generation tracking technologies. *Annu Rev Food Sci Technol* 2016;7:395–411. DOI: 10.1146/annurev-food-022814-015729.
37. Kumar R, Yi N, Zhi D, et al. Identification of donor microbe species that colonize and persist long term in the recipient after fecal transplant for recurrent *Clostridium difficile*. *NPJ Biofilms Microbiomes* 2017;3:12. DOI: 10.1038/s41522-017-0020-7.
38. Li SS, Zhu A, Benes V, et al. Durable coexistence of donor and recipient strains after fecal microbiota transplantation. *Science* 2016;352:586–589. DOI: 10.1126/science.aad8852.
39. Smillie CS, Sauk J, Gevers D, et al. Strain tracking reveals the determinants of bacterial engraftment in the human gut following fecal microbiota transplantation. *Cell Host Microbe* 2018;23:229–240.e5. DOI: 10.1016/j.chom.2018.01.003.
40. Segata N. On the road to strain-resolved comparative metagenomics. *mSystems* 2018;3:e00190-17. DOI: 10.1128/mSystems.00190-17.
41. Sorokin VA, Gelfand MS, Artamonova II. Evolutionary dynamics of clustered irregularly interspaced short palindromic repeat systems in the ocean metagenome. *Appl Environ Microbiol* 2010;76:2136–2144. DOI: 10.1128/AEM.01985-09.
42. Pride DT, Sun CL, Salzman J, et al. Analysis of streptococcal CRISPRs from human saliva reveals substantial sequence diversity within and between subjects over time. *Genome Res* 2011;21:126–136. DOI: 10.1101/gr.111732.110.
43. Pride DT, Salzman J, Relman DA. Comparisons of clustered regularly interspaced short palindromic repeats and viromes in human saliva reveal bacterial adaptations to salivary viruses. *Environ Microbiol* 2012;14:2564–2576. DOI: 10.1111/j.1462-2920.2012.02775.x.
44. Gogleva AA, Gelfand MS, Artamonova II. Comparative analysis of CRISPR cassettes from the human gut metagenomic contigs. *BMC Genomics* 2014;15. DOI: 10.1186/1471-2164-15-202.
45. Davison M, Treangen TJ, Koren S, et al. Diversity in a polymicrobial community revealed by analysis of viromes, endolysins and CRISPR spacers. *PLoS One* 2016;11:1–23. DOI: 10.1371/journal.pone.0160574.
46. Hidalgo-Cantabrana C, Sanozky-Dawes R, Barrangou R. Insights into the human virome using CRISPR spacers from microbiomes. *Viruses* 2018;10. DOI: 10.3390/v10090479.
47. Rho M, Wu Y-W, Tang H, et al. Diverse CRISPRs evolving in human microbiomes. *PLoS Genet* 2012;8:e1002441. DOI: 10.1371/journal.pgen.1002441.
48. Zhang Q, Ye Y. Not all predicted CRISPR-Cas systems are equal: Isolated cas genes and classes of CRISPR like elements. *BMC Bioinformatics* 2017;18:92. DOI: 10.1186/s12859-017-1512-4.
49. Li D, Liu CM, Luo R, et al. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* 2015;31:1674–1676. DOI: 10.1093/bioinformatics/btv033.
50. Ye Y, Zhang Q. Characterization of CRISPR RNA transcription by exploiting stranded metatranscriptomic data. RNA 2016;22:945–956. DOI: 10.1261/rna.055988.116.
51. 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.
52. Cramton SE, Schnell NF, Götz F, Brückner R. Identification of a new repetitive element in *Staphylococcus aureus*. *Infect Immun* 2000;68:2344–2348. DOI: 10.1128/IAI.68.4.2344-2348.2000.
53. Geissmann T, Chevalier C, Cros MJ, et al. A search for small noncoding RNAs in *Staphylococcus aureus* reveals a conserved sequence motif for regulation. *Nucleic Acids Res* 2009;37:7239–7257. DOI: 10.1093/nar/gkp668.
54. Skennerton CT, Imelfort M, Tyson GW. Crass: identification and reconstruction of CRISPR from unassembled metagenomic data. *Nucleic Acids Res* 2013;41:e105. DOI: 10.1093/nar/gkt183.
55. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 2006;22:1658–1659. DOI: 10.1093/bioinformatics/btl158.
56. Huttenhower C, Gevers D, Knight R, et al. Structure, function and diversity of the healthy human microbiome. *Nature* 2012;486:207–214. DOI: 10.1038/nature11234.
57. Franzosa EA, Morgan XC, Segata N, et al. Relating the metatranscriptome and metagenome of the human gut. *Proc Natl Acad Sci U S A* 2014;111:E2329–2338. DOI: 10.1073/pnas.1319284111.
58. Barrangou R, Fremaux C, Deveau H, et al. CRISPR provides acquired resistance against viruses in prokaryotes. *Science* 2007;315:1709–1712. DOI: 10.1126/science.1138140.
59. Lillestøl RK, Shah SA, Brügger K, et al. CRISPR families of the crenarchaeal genus Sulfolobus: bidirectional transcription and dynamic properties. *Mol Microbiol* 2009;72:259–272. DOI: 10.1111/j.1365-2958.2009.06641.x.
60. Robles-Sikisaka R, Ly M, Boehm T, et al. Association between living environment and human oral viral ecology. *ISME J* 2013;7:1710–1724. DOI: 10.1038/ismej.2013.63.
61. Robles-Sikisaka R, Naidu M, Ly M, et al. Conservation of streptococcal CRISPRs on human skin and saliva. *BMC Microbiol* 2014;14:1–15. DOI: 10.1186/1471-2180-14-146.

Received for publication October 11, 2018;
Revised November 29, 2018;
Accepted January 9, 2019.