BMC Genomics

RESEARCH Open Access

Diversity and dynamics of the CRISPR-Cas systems associated with *Bacteroides fragilis* in human population

Tony J. Lam, Kate Mortensen and Yuzhen Ye*

Abstract

Background: CRISPR-Cas (clustered regularly interspaced short palindromic repeats—CRISPR-associated proteins) systems are adaptive immune systems commonly found in prokaryotes that provide sequence-specific defense against invading mobile genetic elements (MGEs). The memory of these immunological encounters are stored in CRISPR arrays, where spacer sequences record the identity and history of past invaders. Analyzing such CRISPR arrays provide insights into the dynamics of CRISPR-Cas systems and the adaptation of their host bacteria to rapidly changing environments such as the human gut.

Results: In this study, we utilized 601 publicly available *Bacteroides fragilis* genome isolates from 12 healthy individuals, 6 of which include longitudinal observations, and 222 available *B. fragilis* reference genomes to update the understanding of *B. fragilis* CRISPR-Cas dynamics and their differential activities. Analysis of longitudinal genomic data showed that some CRISPR array structures remained relatively stable over time whereas others involved radical spacer acquisition during some periods, and diverse CRISPR arrays (associated with multiple isolates) co-existed in the same individuals with some persisted over time. Furthermore, features of CRISPR adaptation, evolution, and microdynamics were highlighted through an analysis of host-MGE network, such as modules of multiple MGEs and hosts, reflecting complex interactions between *B. fragilis* and its invaders mediated through the CRISPR-Cas systems.

Conclusions: We made available of all annotated CRISPR-Cas systems and their target MGEs, and their interaction network as a web resource at https://omics.informatics.indiana.edu/CRISPRone/Bfragilis. We anticipate it will become an important resource for studying of *B. fragilis*, its CRISPR-Cas systems, and its interaction with mobile genetic elements providing insights into evolutionary dynamics that may shape the species virulence and lead to its pathogenicity.

Keywords: CRISPR-Cas systems, Spacer heterogeneity, Host-MGE interaction network

Background

Microorganisms play a crucial role in human health by forming endosymbiotic relationships with their hosts and other microorganisms. These complex networks of microbial communities found throughout various environments, particularly in the human gut, are referred to as microbiomes [1–3]. Aside from bacteria-host interactions, bacteria are constantly engaged in an evolutionary arms-race with mobile genetic elements (MGEs), such as phage and plasmids. To defend against antagonistic actors, prokaryotes have developed a variety of mechanisms to alleviate such threats, one of which are CRISPR-Cas systems, an adaptive immune system that provides sequence-specific defense against invading MGEs [4–6].

CRISPR-Cas systems are highly prevalent, existing in approximately half of bacterial and most of the archaeal

*Correspondence: yye@indiana.edu
School of Informatics, Computing and Engineering, Indiana University, Bloomington, IN, USA

genera [7–9]. The extreme diversity of CRISPR-Cas systems is reflected by their ever-changing classification scheme, owing to the constant discovery of new CRISPR-Cas system types and subtypes [5, 10, 11]. CRISPR-Cas systems can be grouped into two main classes: Class I and Class II CRISPR-Cas Systems. Class I CRISPR-Cas Systems includes Types I, III and IV and use a complex of Cas proteins to degrade foreign nucleic acids. Class II CRISPR-Cas Systems include Types II, V, and VI and use a single, large Cas protein for the same purpose (Type II, V and VI use Cas9, Cas12 and Cas13, respectively) [12]. The diversity of CRISPR-Cas systems provides a fitness edge against invaders and is suggested to be a product of advantageous evolution [13–15]. Similarly, evolution of invaders have been observed to occur in tandem with host adaptive immunity as to evade host defense mechanisms, such as anti-CRISPR genes [4, 5, 16–19].

CRISPR arrays are comprised of short DNA segments, known as spacers, and these provide a cornerstone to CRISPR-Cas derived adaptive immunity. Spacers retain the memory of past immunological encounters, and are primarily acquired as a result of Cas protein complex mediated acquisition [5]. Newly acquired spacers are typically integrated towards the leader ends of arrays [20, 21]. Additionally, leader sequences usually found upstream of CRISPR arrays are attributed to the efficiency of CRISPR-Cas derived immune response [22]. Several studies have also suggested that spacer acquisition remains possible through several alternative means such as homologous recombination [20, 23, 24], and ectopic spacer integration where spacers are inserted into the middle of arrays as a result of leader sequence mutations [22, 25]. While CRISPRs hold immunological memory of past encounters with some arrays spanning several hundred spacers long [26], CRISPR arrays are typically found to be on average less than 30 spacers long suggesting that some spacers are purged over time [27]. While a specific underlying mechanism of CRISPR array maintenance has not yet been elucidated, various studies have suggested several mechanisms of spacer loss, such as spontaneous deletions, recombination, and DNA polymerase slippage during replication [23, 28–30].

In recent years, much effort has been placed into expanding our understanding of the interactions between microbiomes and their host, as well as, the potential modulation of the human microbiome to improve human health. One particular member of the microbiome, *B. fragilis*, has been proposed as a potential probiotic due to its ability to facilitate the alleviation of certain disease conditions [31]. In contrast, Bacteroidetes is one of the most common genera of bacteria in the lower intestinal tract, and while this member of the microbiome only accounts for a small fraction ($\sim$ 2%) of the total Bacteroides found in the gut microbiome, this species contributes to over 70% of Bacteroides infections [32–34]. This is due to *B. fragilis*' extensive pan-genome and susceptibility to horizontal gene transfer events. As a result, certain strains of *B. fragilis* have become known pathobionts and opportunistic pathogens [35–37]. The perplexing interplay between the pathogenic and probiotic nature of *B. fragilis* strains highlights the importance of understanding pathobiont evolutionary dynamics, elements that contribute to a species' pathogenicity, and CRISPR-Cas dynamics.

Many studies of the adaptation process in CRISPR-Cas systems involve an individual bacterial species challenged with invaders in controlled assays. Taking advantage of the increasing number *B. fragilis* reference genomes, and more importantly, the large number of *B. fragilis* isolates from 12 individuals [38], we re-investigated the CRISPR-Cas systems in *B. fragilis* in its natural living environment. The availability of hundreds of time-resolved genomes from *B. fragilis* isolates from 7 individuals (some involving multiple time points) allowed us to investigate both the intra- and inter-personal dynamics of interactions between *B. fragilis* and their invaders, and expand upon previous surveys of *B. fragilis* CRISPR-Cas systems [37]. Insights into how *B. fragilis* interacts with its invaders, as well as how its CRISPR-Cas systems confer immunity help improve our understanding of the factors that contribute to *B. fragilis* virulence, horizontal gene transfer, and evolution.

Results

CRISPR-Cas systems in *B. fragilis*

To better understand the dynamics of CRISPR-Cas systems within *B. fragilis*, we analyzed a total of 823 *B. fragilis* genomes, which included 222 NCBI (National Center for Biotechnology Information) reference genomes and 601 isolates from 12 healthy individuals, 7 of which include longitudinal observations (referred as the Zhao2019 dataset; see Methods) [38].

Our analysis shows that among all *B. fragilis* genomes, three types of CRISPR-Cas systems were identified, Type I-B (class 1), Type II-C (class 2), and Type III-B (class 1; see a review article [39] for the classification of CRISPR-Cas systems). The three types of CRISPR-Cas systems contain universal *cas* genes including *cas1* and *cas2*, and other type specific genes. For example, the Type II-C CRISPR-Cas systems contain the Type II signature *cas9* gene; the Type III-B CRISPR-Cas systems contain the signature *cas10* gene; and the Type I-B CRISPR-Cas systems contain the subtype I-B specific *cas5* gene, although *cas5* is universally found in all class I CRISPR-Cas systems [39]. Example illustrations of these CRISPR-Cas systems are depicted in Fig. 1A, and Table 1 shows the signature *cas* genes and the repeat sequence of the CRISPR arrays in these CRISPR-Cas systems. We note that besides the three

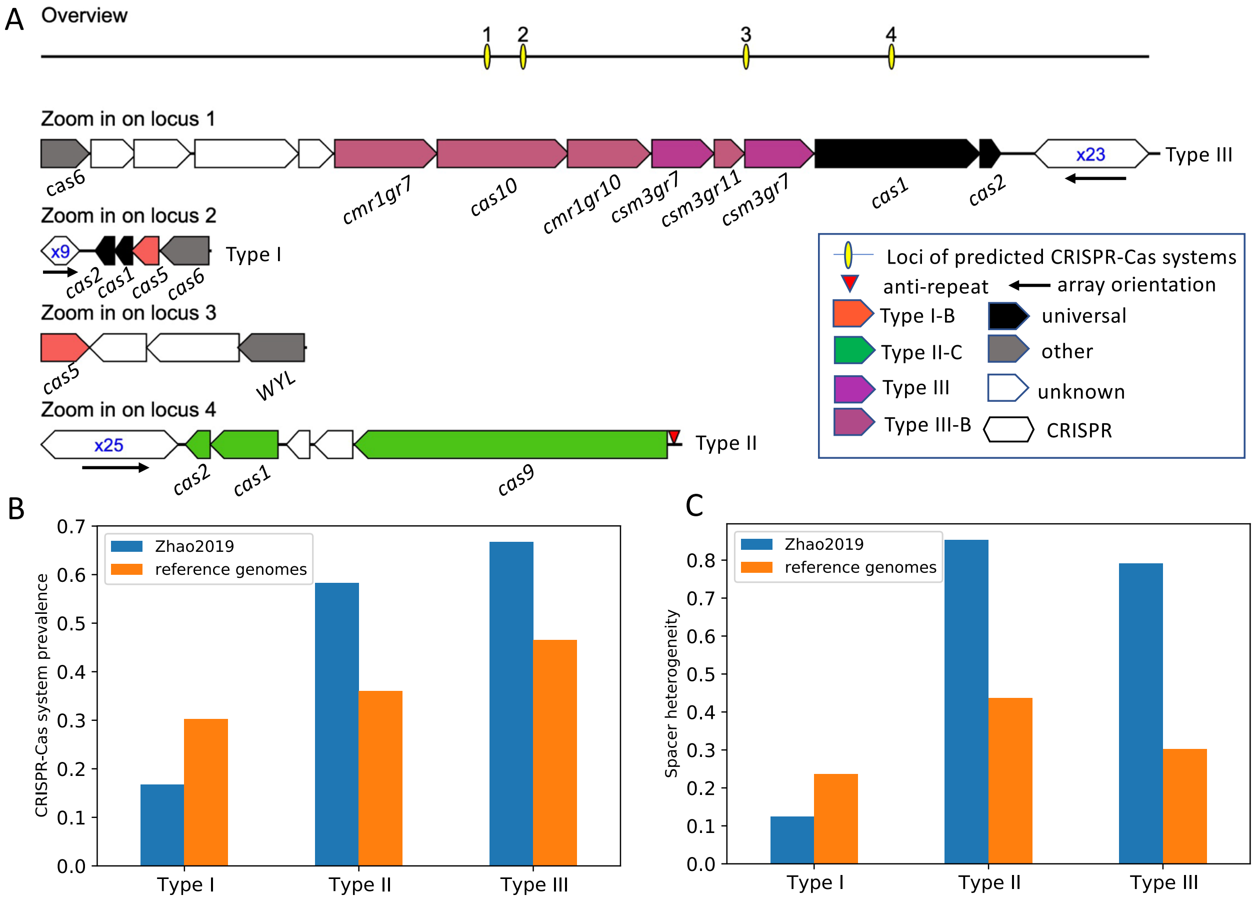

Fig. 1 The three types of CRISPR-Cas systems found in *B. fragilis* genomes. (**A**) Representatives of *B. fragilis* CRISPR-Cas systems found in *B. fragilis* strain S14 (accession number: GCA_001682215.1_ASM168221v1). The figure legend indicates the CRISPR-Cas types by color where each arrow represents a gene and its orientation. While some *cas* genes are labeled in italicized text below the arrows, further details can be found on the companion website for this accession number. CRISPR arrays are labeled by their number of repeats in blue text. The predicted anti-repeat that is partially complementary to the CRISPR repeat in the Type II system is shown as a red triangle in the plot. (**B**) Prevalence of CRISPR-Cas systems among the NCBI reference genomes and Zhao2019 isolates. (**C**) Spacer content heterogeneity of the CRISPR arrays

types of predicted CRISPR-Cas systems, additional putative CRISPR arrays were predicted in a *de novo* fashion. However, they were deemed to be false CRISPR arrays due to their lack of spacer content heterogeneity, despite the fact that they superficially contain the repeat-spacer structures (see details of these CRISPR artifacts and reasons why there were discarded at the companion website). Among the discarded CRISPR groups include a putative fourth CRISPR-Cas system that was previously reported in *B. fragilis genomes* [37]. This CRISPR-like artifact was found in 137 out of the 222 (61.7%) reference *B. fragilis* genomes, and isolates of *B. fragilis* in 10 out of 12 individuals in the Zhao2019 dataset (other CRISPR artifacts were rare, found in one or very few genomes). Additionally, this CRISPR artifact was predicted to contain protein-coding genes encoding for transcriptional regulators in some genomes (e.g., CP036550.1 and CP0811922.1), further suggesting that it is unlikely a genuine CRISPR.

Table 1 Signature *cas* genes and CRISPR repeat sequences for the three types of CRISPR-Cas systems found in *B. fragilis*

Type	CRISPR repeat seq	repeat ID [a]	*cas* genes
Type I-B	ATTTCAATTCCATAAGGTACAATTAATAC	BfragL29	*cas5* [b]
Type II-C	GCTGTTTCCAATGGTTCAAAGATACTAATTTGAAAGCAAATCACAAC	BfragL47	*cas9*
Type III-B	GTCTTAATCCTTATTATACTGGAATACATCTACAT	BfragL35	*cas10*

[a]: repeat IDs are BfragL followed by the length of the corresponding repeat sequence. [b]: *cas5* subtype I-B

Table 2 Presence of the different types of CRISPR-Cas systems in the 12 individuals

Type	S01 (t10)	S02 (t2)	S03 (t2)	S04 (t2)	S05 (t3)	S06 (t2)	S07 (t3)	S08	S09	S10	S11	S12
Type I-B							yes	yes				
Type II-C		yes	yes			yes	yes			yes		
Type III-B	yes	yes				yes	yes		yes	yes		

Inter-subject spacer diversity of *B. fragilis* cRISPR-Cas systems

An evaluation of the CRISPR-Cas system distribution among all isolates of the Zhao2019 dataset showed that CRISPR-Cas system types were unevenly and dis-similarly distributed between individuals (Fig. 1B and Table 2). Type I-B CRISPR-Cas systems were among the least prevalent with only isolates from two individuals (S07 and S08) containing this type of CRISPR-Cas system. Type II-C and Type III-B CRISPR-Cas systems are similarly prevalent with isolates from five and six individuals, respectively (see Table 2 for the lists of individuals that contain these systems). No CRISPR-Cas systems were found to be present within Zhao2019 subjects S04, S05, S11 and S12. The lack of uniformity of CRISPR-Cas system presence, or lack thereof, suggests that lineages of *B. fragilis* between individuals are, for the most part, unique from each other, reaffirming the findings of Zhao et al. [38]. Similarly, there was an observed lack of shared inter-individual spacer content, with majority of the spacers observed being individual specific (Fig. 3), with the most common shared spacer being the anchor spacer on the trailer end of observed CRISPR arrays.

Spacer content heterogeneity score (Fig. 1C) and compressed spacer graph of Type I-B CRISPR-Cas systems (Fig. 2A) found within Zhao2019 isolates show that the CRISPR arrays of Type I-B systems have low heterogeneity and are less active in terms of spacer turnover compared to other CRISPR-Cas systems in *B. fragilis*. In comparison, Type III-B CRISPR-Cas Systems contained mostly individual specific spacers and shared very few spacers between individuals. This pattern of individual-specific spacers is reflected in the branching structures observed in the compressed spacer graphs (Fig. 2B). Each branch within the spacer graph represents an unique CRISPR array structure; bottle-neck nodes (e.g. 79, 53, and 48) represent uniformly shared spacer(s) in spacer sharing CRISPR arrays. The observed branching structure in the compressed spacer graph indicates a diverse CRISPR array structure between individuals, indicative of the activity and heterogeneity of Type III-B CRISPR-Cas systems within *B. fragilis*. For comparison, the spacer content heterogeneity score shows similar trends among the *B. fragilis* reference genomes (see Fig. 1C).

The Type II-C CRISPR-Cas systems found within the Zhao2019 isolates were among the most diverse between

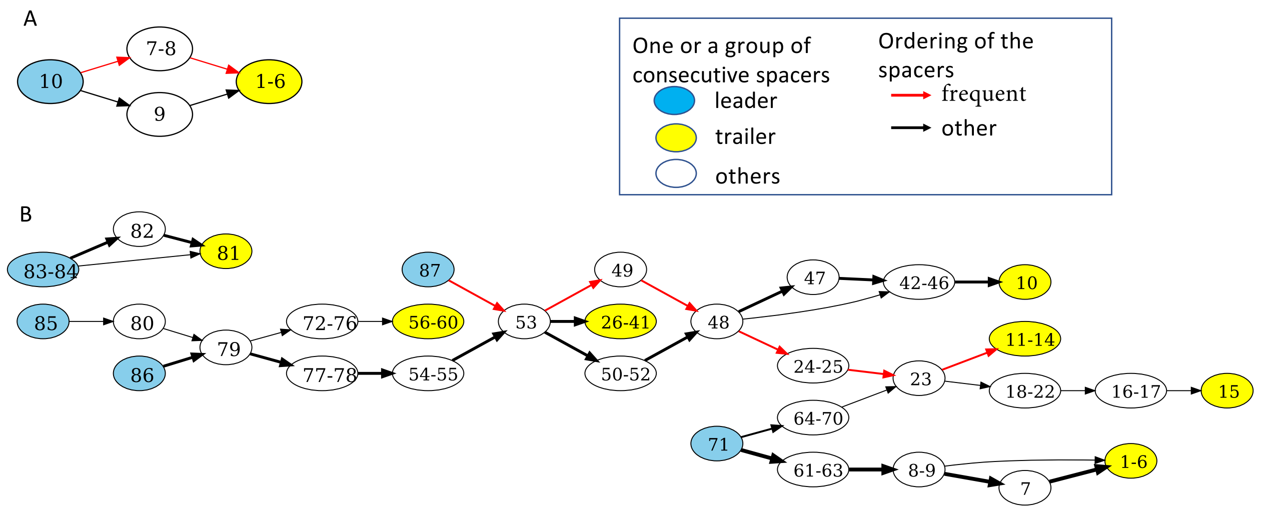

Fig. 2 Compressed spacer graphs of Type I-B (**A**) and Type III-B (**B**) CRISPR-Cas systems in *B. fragilis*. Numbered nodes represent a single, unique spacer whereas nodes labeled with numerical intervals represent several consecutive spacers, and directed edges represent the ordering of the spacers in arrays. For the spacers in the same array, spacers closer to the trailer end are numbered with smaller numbers whereas the spacers closer to the leader end get bigger numbers. Nodes containing leader end spacers are highlighted in blue, and nodes containing trailer end spacers are shown in yellow. The compressed spacer graphs provide visual summary of 80 Type I-B CRISPR arrays identified from 80 Zhao2019 isolates (**A**), and 465 Type III-B CRISPR arrays identified from 410 isolates (**B**), respectively

the three observed types of CRISPR-Cas systems found within *B. fragilis*. Among the seven subjects that contained Type II-C CRISPR-Cas systems, many of the identified Type II-C CRISPRs did not share inter-subject spacers, except for trailer spacers (i.e., end spacers such as nodes 1 and 19 shown in Fig. 3) which have been previously hypothesized as ancient spacers or anchor spacers [28, 40, 41]. The spacer sequence diversity can be seen in Fig. 3, where each branch path represents a unique CRISPR array observed. The diversity of the CRISPR arrays observed in Type II-C CRISPR-Cas systems suggests that Type II-C systems have greater spacer activity (e.g. spacer acquisition and loss), and also highlight the evolutionary pressures that MGEs exert on *B. fragilis*.

Intra-individual cRISPR-Cas dynamics of *B. fragilis*

Taking advantage of the temporal intra-individual isolates of *B. fragilis*, we were able to study micro-dynamics of *B. fragilis* community dynamics and the adaptation of its CRISPR arrays over time per individual. Overall CRISPR array structures remain relatively stable across samples from the same individual, with some slight variations between observed CRISPR arrays (e.g., Fig. 3 S01, S02, S03, S06, S07). A notable example is the arrays of type III-B CRISPR-Cas systems found in S01 isolates (Fig. 4A). Isolates were derived from this individual at 10 different time points spanning more than 2 years, and we only observed a small variation of the arrays in those isolates, resulting in a simple spacer graph with only one branching structure involving the loss (or gain) of a spacer in the middle of the arrays.

However, in some instances, periods of diverse spacer acquisition were observed from samples from the same individual (e.g., S08). As shown in Fig. 3, various strains of *B. fragilis* with varying CRISPR array structures were observed from isolates obtained in the same individual (S08) at a single time point. The spacer graph shows a funneling pattern, where multiple nodes on the leader end of the CRISPR array converge into a single neighboring node (see blue nodes on the left of Fig. 3; highlighted in a red triangle). This observed pattern in the spacer graph suggests that multiple 'lineages' have gained alternative leader end spacers in comparison to each other, specifically when the bacteria are exposed to different MGEs and are evolving according to the observed threat.

More examples of divergent lineages can be found in Fig. 4B–D, where instances of time-point-specific non-spacer sharing CRISPR arrays were present, as well as cross-time-point shared CRISPR arrays were present within *B. fragilis* Type II-C CRISPR-Cas systems. There are multiple lineages of *B. fragilis* containing diverse Type II-C CRISPR arrays in S02 (at time point S02-0001) and S06 (at time point S06-0001). Figure 4D shows a few representative CRISPR arrays found in individual S06, in which the three representative spacer-sharing CRISPR arrays in two time points, S06-0001 and S06-0122, were mostly similar except for differences likely a result of loss from two consecutive spacers in their corresponding isolates. We note that the time-point specific spacers/arrays were rare and found in a small number of isolates as compared to the arrays that share many spacers (see the numbers of isolates in Fig. 4). While previous studies have shown that intra-individual populations of *B. fragilis* are dominated by a single strain [42–44], our findings here show that in some cases many lineages, or strains of *B. fragilis*, remain present within the same individual at any given

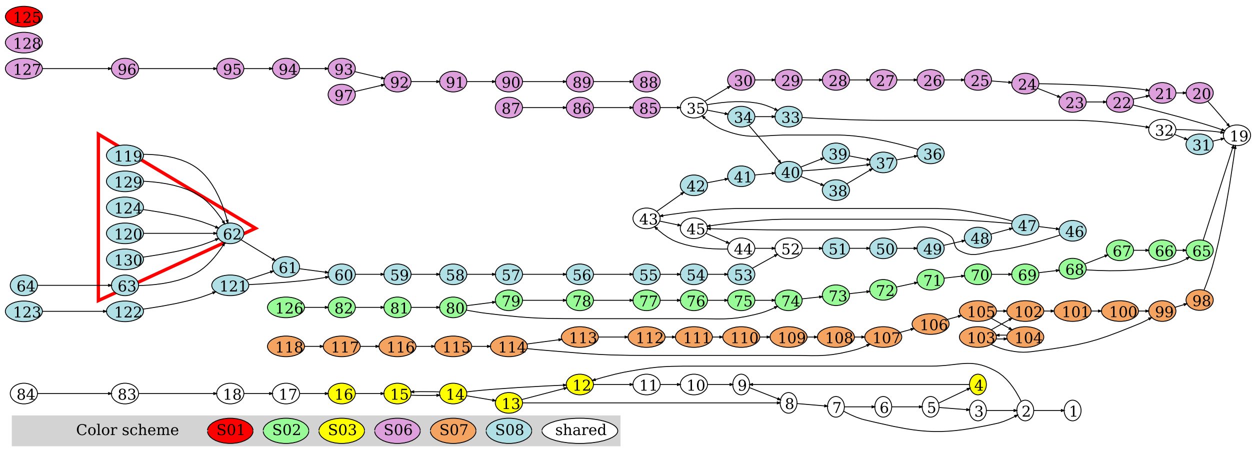

Fig. 3 Spacer graph of the Type II-C CRISPR-Cas systems in *B. fragilis*. Each numbered node represents a unique spacer. Spacers that are unique to individuals are highlighted in different colors (see color scheme in the plot; spacers shared by isolates from at least two individuals are shown in white). This spacer graph summarizes the spacer organization in 250 CRISPR arrays of Type II-C identified from 187 Zhao2019 isolates. Of particular interest, the funneling shape highlighted in a red triangle is a result of a total of 28 CRISPR arrays involving insertion of six different new spacers (nodes numbered 119, 129, 124, 120, 130 and 63) into the CRISPR arrays after the same spacer (numbered 62)

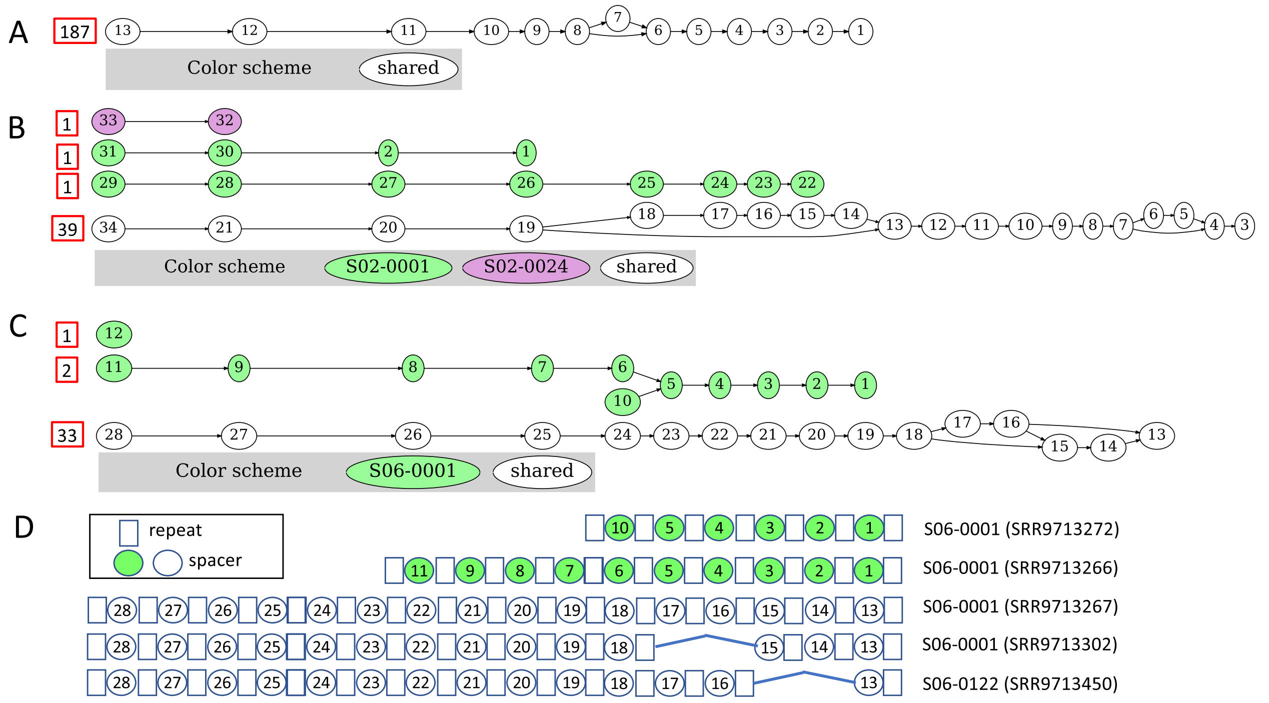

Fig. 4 Intra-individual CRISPR array diversity. Each numbered node represents a single, unique spacer, and directed edges represent the ordering of spacers in the CRISPR arrays. (**A**) Spacer graph of Type III-B CRISPR arrays from individual S01 isolates, consisting of 10 time points, showing low spacer diversity. (**B**) Spacer graph of Type II-C CRISPR arrays for individual S02. Green nodes are unique spacers identified from time point 1 (S02-0001), purple nodes are from time point 24 (S02-0024), and white nodes represent spacers found in multiple time points. (**C**) and (**D**) are the spacer graph, and the pileup view of five representative CRISPR arrays found in five isolates from individual S06, respectively. The five isolates were from two time points, including four from time point S06-0001 and one from time point S06-0122. The arrays involving spacers shown in green are unique to S06-0001 time point isolates, whereas the arrays involving spacers shown in open ovals are common to isolates from both time points. The numbers included in red boxes in figures A-C represent the number of isolates contributing to each major form of CRISPR arrays; for example, 1, 2, and 33 isolates contributed to the three forms of Type II-C CRISPR arrays in individual S06 shown in figure C, respectively

time point. The observation of various intra-individual *B. fragilis* strains is yet another example of the evolutionary arms race between host and the invading MGE.

Interaction network of *B. fragilis* and its invaders

A total of 1531 unique spacers were identified from *B. fragilis* genomes (including the 823 Zhao2019 isolates and 222 reference genomes). Among these spacers, 136 were shared by the two collections, 1290 were found in reference genomes only, and 104 were unique to the Zhao2019 isolates. We note that although Zhao2019 isolates outnumbered the collection of reference genomes we analyzed, due to the redundant nature of the Zhao2019 isolates (from 12 individuals), fewer unique spacers were identified in the Zhao2019 collection. All the spacers were used to identify potential MGEs that had left their traces in the *B. fragilis* genomes.

Among the 1531 unique spacers identified from *B. fragilis* isolates, 522 found matches (protospacers) in 161 MGEs (153 phages and 8 plasmids). 108 out of the 153 phages could be assigned to a family by PhaGCN [45] with a majority of them being Siphoviridae (93, 86%). Using these spacers, interaction networks between *B. fragilis* and its invaders were inferred. Analysis of the networks (Fig. 5A and B) showed varying levels of micro-dynamics within *B. fragilis* CRISPR-Cas systems. The spacer-MGE network (Fig. 5A) contains a few modules each containing a large number of MGEs and spacers (e.g., modules a, b, c and d highlighted in the Figure), likely a result of the arms-race between *B. fragilis* and MGEs (*B. fragilis* acquired new spacers to maintain immunity and invaders mutated to evade immunity). The spacer-MGE network shows that *B. fragilis* used its Type I-B and II-C CRISPR-Cas systems extensively to defend against MGEs that were mostly phages (the network contains 353, 163, and 3 spacers that were exclusively caught in Type II-C, Type I-B, and Type III-B CRISPR-Cas systems, respectively). It also suggests differential defense activities of the Type I-B and II-C CRISPR-Cas systems against some invaders (e.g., those included in modules a and b were preferentially targeted by Type II-C CRISPR-Cas systems; by contrast, invaders included in modules c and d don't show such preference).

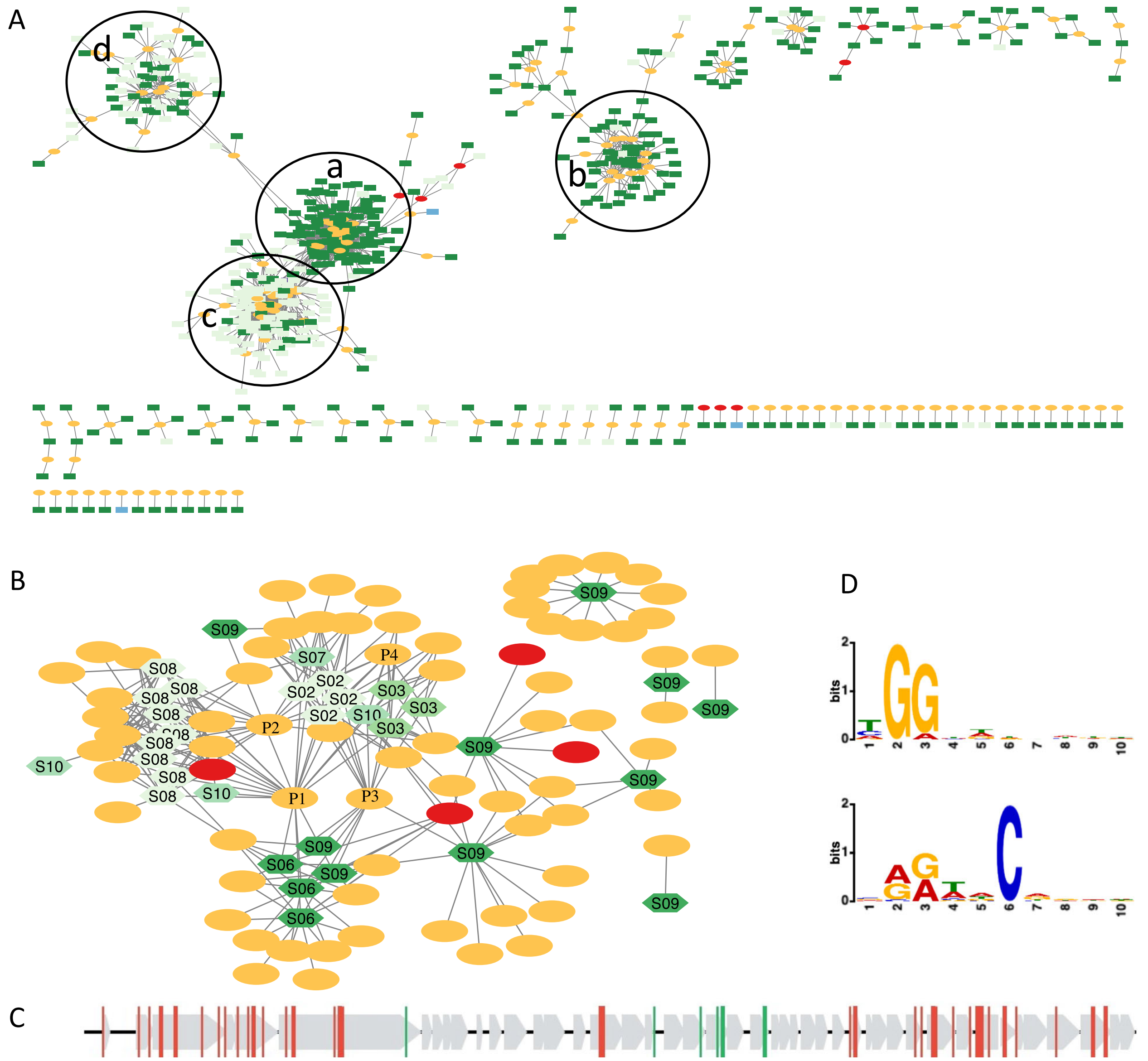

Fig. 5 Interactions between *B. fragilis* and MGEs, as well as, the distribution of protospacers in NC_011222. MGEs containing protospacers are shown in ovals with plasmids highlighted in red and phages in orange. (**A**) The spacer-MGE network includes spacers identified from Zhao2019 and the genomes of their MGE targets (if protospacer were identified in any known MGE reference genomes). Spacers identified in different types of CRISPR-Cas systems are shown in rectangles of different colors (Type I-B in light green, Type II-C in dark green, and Type III-B in blue). (**B**) The host-MGE network includes spacer-contributing *B. fragilis* isolates from Zhao2019 only. Hosts are shown in hexagons with *B. fragilis* isolates from different individuals shown in different shades of green; four phage MGEs are highlighted with labels: P1 (uvig_425872, Siphoriridae), P2 (uvig_422350, Siphoviridae), P3 (k141_68_round8-12 1718861, Microviridae), and P4 (NC_011222, Siphoriridae). (**C**) The distribution of protospacers in NC_011222, with green and red vertical lines representing protospacers matching spacers found in Type I-B and Type II-C CRISPR-Cas systems, respectively (gray arrows represent the genes). (**D**) PAMs detected in the downstream regions of the protospacers associated with Type I-B and Type II-C CRISPR-Cas systems, respectively

Figure 5B (focusing on *B. fragilis* isolates from several individuals) showed that some invaders (such as P1, P2, P3 and P4 located at the center of the network) have their traces found in *B. fragilis* in many different individuals, likely the result of ubiquitous presence of these MGEs in human gut. Despite of these central MGEs that make the whole network highly connected, we observed groupings of *B. fragilis* isolates from one or two subjects with some more localized MGEs (e.g., the MGEs that were targeted by the *B. fragilis* CRISPR-Cas systems in individual S09). Similar trends can be observed in the heatmap visualizations of the spacer-MGE and host-MGE interaction networks, as shown in Supplementary Fig. 1 and Fig. 2, respectively.

We analyzed the protospacers and their downstream regions in the identified MGEs. Figure 5C shows the distribution of protospacers in NC_011222 (Bacteroides phage B40-8, labelled as P4 in Fig. 5B) that was targeted extensively by both Type I-B and Type II-C CRISPR-Cas systems. We were able to infer the protospacer adjacent motif (PAM) that follow the protospacers. We extracted 10 bases of the downstream regions of all the protospacers and applied MEME [46] to detect and visualize the motifs among these sequences. Figure 5D shows the PAMs found in the downstream regions of the protospacers that matched spacers found in the Type I-B and Type II-C CRISPR-Cas systems, respectively. The logos show that the Type I-B and Type II-C systems target segments with distinct PAMs: Type I-B CRISPR-Cas system tends to target segments with base G at their 2 and 3 downstream positions, whereas Type II-C CRISPR-Cas system tends to target segments with conserved base C at downstream position 6.

Discussion

In this paper, we expanded upon previous works [37] and explored the CRISPR-Cas dynamics within *B. fragilis* genomes, while focusing on dynamics pertaining to a time-resolved study of *B. fragilis* within and between individuals. We analyzed a total of 823 genomes, a 7.5 fold difference in number of genomes analyzed in previous *B. fragilis* CRISPR-Cas papers [37]. While *B. fragilis* is a common commensal bacterium of the human gut microbiome, sometimes a probiotic candidate and sometimes pathogen, its role as one of the most virulent members of the *Bacteroides* genus should not be overlooked [47]. Part of *B. fragilis* virulence is due to its potent virulence factors, and as such, a thorough understanding of the mechanisms and factors that contribute to its virulence, horizontal gene transfer, and evolution are important. By utilizing CRISPR-Cas systems and focusing on time series isolates, we were able to reveal micro-dynamics found in *B. fragilis* isolates within and between individuals.

The analysis of NCBI's reference genomes and genomes from the Zhao2019 dataset enabled us to update the evaluation of known CRISPR-Cas systems found within *B. fragilis*. Particularly, we found three types of CRISPR-Cas systems (Type I-B, Type II-C, and Type III-B) with varying distributions among the genomes. Our analysis also shows that a fourth previously reported CRISPR-Cas system in *B. fragilis* was a false CRISPR-like artifact. This CRISPR-like artifact was previously characterized as an orphaned CRISPR array [37], but due to its structure containing only two spacers, three repeats, as well as non-uniform repeat sequences, we believe this is not an orphaned CRISPR array.

While differentiating between active, in-active, and false-positive CRISPRs remains a challenging and active research area, we employed various methods to mitigate the potential of including false-positive CRISPR arrays in our analysis. Identification of CRISPR arrays can be challenged by repetitive sequences that mimic CRISPR array structures. Here we employed the use of CRISPRone (which employs an ensemble method to remove potential false-positives) [44], and additionally introduce the filtering of putative CRISPR arrays through the use of spacer content heterogeneity. Our analysis shows that while all *B. fragilis* CRISPR-Cas system types had some level of plasticity, where CRISPR arrays across different time points and individuals were heterogeneous, the level of heterogeneity varied between CRISPR-types and even time-points. Intra-individual variations of CRISPR arrays, such as those found in Individual S08 (Fig. 3), showed periods of rapid expansion and diversification of CRISPR spacers between strains of observed isolates; these periods of diversification can be observed in the branching structures of the spacer graph. In comparison, periods of contraction where little to no CRISPR spacer content heterogeneity was observed were similarly present in intra-individual CRISPR-Cas systems, such as those found in Individual S01 (Fig. 4A). Unsurprisingly, most inter-individual CRISPR-Cas systems did not share many spacers between individuals. This could be explained that individuals picked up different isolates of *B. fragilis*. Here we also show that CRISPRs can go through periods of expansion, while others go through periods of stability, suggesting that CRISPR evolution is not a constant process but occurs in modes. Uncovering these CRISPR-Cas dynamics would not be possible without time series analysis of the same bacterial lineage. We found that *B. fragilis* CRISPR-Cas systems seemed to prefer targeting phage genomes over plasmid genomes while exploring the interplay/dynamics of *B. fragilis* and its MGEs. This is a contrast to some studies which found CRISPRs favoring the targeting of plasmids over phages [48, 49]. CRISPR spacer-MGE networks also revealed micro-dynamics of *B. fragilis* CRISPR targets, where we observed several notable network structures. Hairball-like structures, where a single spacer targeted many unique MGE targets, and exemplified that in some cases CRISPR spacers were likely able to target multiple MGEs through the same CRISPR spacer. This suggests that the protospacer is conserved across many targets. In addition to hairball like structures, it was also observed that several spacer nodes and MGE nodes formed cliques/modules, where nodes clustered together more closely to each other than other members of the network. Within these modules, MGE nodes shared an edge with many spacer nodes, suggesting that these MGEs contained many protospacers. This observation of many spacers targeting the same MGE may be suggestive of a process known as 'primed CRISPR adaptation'. In primed CRISPR adaptation, the presence of an existing spacer is used to enhance the acquisition of new spacers on the same MGE target [50, 51]. Alternatively, it may be possible that these instances of multiple targeting are a result of naive adaptation where spacers were independently acquired.

Not all spacers identified in *B. fragilis* had a matching MGE protospacer target, which might have biased our analysis to spacer targets based on available MGE database genomes. However, it has been suggested that most unidentified spacers relate to host-specific mobile elements [52, 53] and thus without adequate sequencing and annotation of the hosts' microbiome, many of the spacer targets will remain unresolved. Another hypothesis to the limited spacer-MGE associated matches, especially in trailer end (older) spacers, is that protospacer sites of targeted MGEs have since mutated to evade detection by the CRISPR spacer and the MGE target pre-dates sequencing technology; thus, spacers are unable to match to any known protospacer targets within the available MGE databases.

Additionally, in compressed spacer graphs, we observed periods of expansion and contraction of CRISPR arrays. Funneling patterns are of particular interest and were mostly observed at the leader end of spacer graphs. The lack of these funnel shaped patterns in the middle or trailer end of compressed spacer graphs suggests that certain spacers may provide an evolutionary advantage compared to other spacers, and establish itself as the dominant strain, out competing strains containing less fit CRISPR arrays; thus we do not see this branching structure in 'older' segments of the CRISPR array.

Although CRISPR-Cas systems are commonly found in prokaryotes, only about half of the bacterial species contain them [9, 44]. We recently showed that human related bacterial species have a broad spectrum of the prevalence of the CRISPR-Cas systems; for example, *Staphylococcus aureus* has the least tendency of obtaining the CRISPR–Cas systems with only 0.55% of its isolates containing CRISPR–Cas systems, whereas most isolates of *Clostridioides difficile* analyzed have CRISPR–Cas systems each having multiple CRISPRs [54]. It is reflected in the Zhao2019 collection—isolates from 8 out of 12 individuals contain one or more of the CRISPR-Cas systems found in *B. fragilis* (see Table 2). This poses a limitation of using the evolution of CRISPR arrays to study the adaption of bacterial species to the changing environments. On the other hand, due to the hypervariable nature of the CRISPR arrays, they provide a sensitive approach for studying the microevolution of bacterial species, as shown in [55, 56].

While our work improves the understanding of *B. fragilis* adaptation to MGE exposure by using inferred host-MGE networks, more work is still needed to understand how CRISPR adaptation plays a role in *B. fragilis* acquisition of virulence factors, evolution, and horizontal gene transfer. In particular, one main challenge to Host-Invader analysis is the limitation of available MGE databases. Future efforts and resources to maintain databases of MGEs and other elements of the microbiome (e.g. fungome) remain invaluable for further understanding of the microbiome, and not just prokaryotic members. A better understanding of how *B. fragilis* and other pathobionts interact with their invading mobile elements will enable a better understanding of their evolution and the elements responsible for their pathogenicity.

Conclusions

By exploring CRISPR-Cas systems present in *B. fragilis* and the dynamics of its host-MGE networks, we uncovered micro-dynamics of *B. fragilis* adaptation against invaders. We made available of all annotated CRISPR-Cas systems and their target MGEs, and their interaction network as a web resource at https://omics.informatics.indiana.edu/CRISPRone/Bfragilis. We anticipate it will become an important resource for studying of *B. fragilis*, its CRISPR-Cas systems, and its interaction with mobile genetic elements providing insights into evolutionary dynamics that may shape the species virulence and lead to its pathogenicity.

Methods

Genomic data processing and assembly

Reads from 601 *B. fragilis* isolates from the Zhao et al. study [38] were downloaded from the NCBI BioProject Accession PRJNA524913, henceforth referred to as the 'Zhao2019 dataset'. All isolates from the Zhao2019 dataset were obtained from the OpenBiome stool bank whose donors abstained from antibiotics for a minimum of 3 months prior to donation [38]. Raw shotgun sequencing reads were trimmed using Trimmomatic v0.39 [57] (parameters used: LEADING:5 TRAILING:5 SLIDINGWINDOW:4:10 MINLEN:60). Trimmed reads were then assembled using SPAdes v3.12 [58] with default settings. FragGeneScan [59] was then used to predict protein coding genes of metagenome assemblies.

A total of 222 *B. fragilis* reference genomes, 16 complete and 202 draft genomes, were downloaded from the NCBI ftp website as of Jan 18, 2021. A list of genomes included in this analysis can be found at the companion web resource.

Characterization of cRISPR-Cas systems

To identify CRISPR-Cas systems in *B. fragilis* genomes, we utilized CRISPRone [44] which predicts both CRISPR arrays and *cas* genes within a given input genome sequence. Predicted CRISPR-Cas systems were then further refined through a reference based approach. Repeat sequences of CRISPRone predicted spacers were extracted and clustered to obtain consensus reference repeats using CD-HIT-EST [60] with 85% sequence identity. Consensus reference repeats were then used as input for CRISPRAlign [61], a reference based approach to identify CRISPR arrays. As the exact boundaries of CRISPR arrays predicted by *de novo* approaches may sometimes be blurred due to small CRISPR arrays, sequencing errors, and mutations in repeat sequences, we utilize a reference based approach to redefine the repeat-spacer boundaries of CRISPR arrays predicted by CRISPRone.

To compare spacer sequences across different arrays, reduce spacer redundancy, and the eventual computation of spacer content heterogeneity, spacers were clustered with CD-HIT-EST [60] at 85% sequence identity. An 85% sequence identity was used to provide greater flexibility in spacer sequences, and allow for a small amount of sequence variation either due to sequencing error or real mutations found between individual spacers. Spacer sequences that clustered together were considered identical spacer sequences. Spacer clusters were reserved for downstream computation of spacer content heterogeneity (Fig. 6A) and construction of compressed spacer graphs (Fig. 6B).

In some cases, it may be difficult to differentiate between true CRISPR-Cas systems and false positive CRISPR-Cas systems (e.g., false CRISPR-arrays, inactive CRISPR-Cas systems). While manual curation can help filter out some of these issues, it becomes difficult to screen out hundreds to thousands of genomes. CRISPRone utilizes a set of heuristics to identify and filter out potential false-positive CRISPR arrays, including STAR-like element [44]. To additionally help filter out potential false positive arrays and inactive CRISPR-Cas systems, we propose a metric of heterogeneity to measure the rate of change (i.e., growth and turnover of spacers) in CRISPR arrays with the assumption that CRISPR arrays of active CRISPR-Cas systems undergo active expansion and turnover of spacers. In instances where spacer content heterogeneity was zero, but arrays had adjacent cas genes, these arrays were considered to be true CRISPRs. Here we define spacer content heterogeneity score as:

$$\text{Spacer Heterogeneity} = \frac{m - max(c_i)}{\sum_i^n c_i - max(c_i)} \qquad (1)$$

Where n is defined as the number of CRISPR arrays, with each CRISPR array containing c_1, c_2,..., c_n unique spacers (in some rare cases, CRISPR arrays may contain multiple copies of the same spacer, which will be considered as one spacer) and m denotes the number of unique spacers found from all n arrays combined. Spacer heterogeneity scores range from 0 to 1, where 0 indicates no spacer heterogeneity (i.e., two CRISPR arrays share all spacers), and 1 indicates the greatest possible extent of spacer content heterogeneity (i.e. two CRISPR arrays share no spacers).

Because spacer content heterogeneity alone is not enough to rule out false positive or inactive CRISPR-Cas systems, predicted CRISPR-Cas systems were further filtered out by coupling spacer content heterogeneity with gene content information. CRISPR groups that lack spacer content heterogeneity and had no adjacent *cas* genes were considered inactive or false positive, and thus discarded from further analysis; all filtered arrays were also manually inspected prior to their removal.

Compressed spacer graph for summarizing the sharing of spacers among a group of CRISPR arrays

Compressed spacer graphs [41] were constructed for each CRISPR-Cas type to summarize and illustrate CRISPR array dynamics. For every spacer in a given array, where each spacer was represented by a node of its representative spacer cluster, a directed edge was built between nodes of neighboring spacers in sequential order. Once all CRISPR arrays were represented in the graph struc-

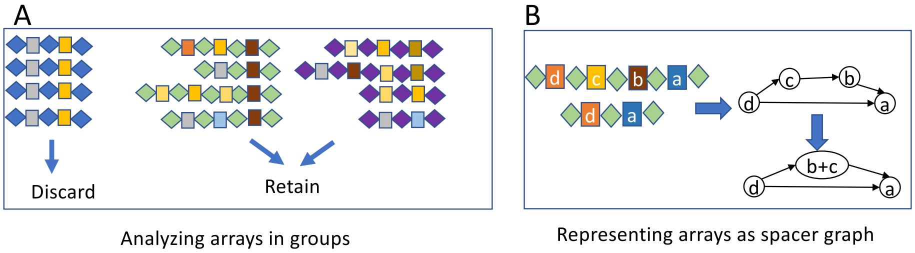
A
Discard
Retain
Analyzing arrays in groups
B
d
c
b
a
d
a
c
b
d
a
b+c
d
a
Representing arrays as spacer graph

Fig. 6 Approaches used for the identification and refinement of the CRISPR arrays and construction of spacer graphs. (**A**) CRISPR arrays are analyzed in groups such that each group shares identical or very similar repeats (repeats are shown as diamonds and spacers are shown as boxes). CRISPR arrays that lack spacer content heterogeneity and have no adjacent cas genes were considered to be false-positive and discarded. (**B**) Example of spacer sharing CRISPR arrays can be represented as a simplified graphical structure (spacer graph), in which the edges record the ordering of the spacers in arrays

ture, the spacer graph was then simplified by collapsing neighboring nodes if two neighboring nodes shared an "in-degree" and "out-degree" equal to or less than one (Fig. 6B). Compressed spacer graphs highlight CRISPR array structure and dynamics (e.g. branching structures representing spacer gain and loss). Arrays that share no spacers result in disconnected components in the compressed spacer graph.

Mobile genetic element databases

A collection of mobile genetic element (MGE) databases were gathered, including phage and plasmid databases. The phage databases included the Gut Phage Database [62] (GPD), MicrobeVersusPhage [63] (MVP) database, and the reference viral database [64] (RVDB). The plasmid databases included the Comprehensive and Complete Plasmid Database [65] (COMPASS), and PLSDB [66]. The phage and plasmid databases included sequences from the NCBI reference database, NCBI nucleotide database, MGEs identified from metagenomic assemblies, and prophages identified in prokaryotic genomes. We collectively refer to these databases as the 'MGE database' for simplicity.

Identification of CRISPR targets

All unique spacer sequences extracted from *B. fragilis*' CRISPR arrays were queried against the MGE database using BLASTN [67] to search for putative invaders that were targeted by *B. fragilis*. For this analysis, we used all unique spacers instead of 85%-similarity nonredundant set to increase the search sensitivity. Results were filtered to retain hits with a greater than 90% sequence identity, query coverage per hsp greater than 80%, and an e-value of less than 0.001. We noticed that even after dereplication by dRep [68] (with default parameters), there was still a large redundancy in the identified MGEs. Instead, we devised a greedy algorithm to select the minimum number of MGEs that collectively contain all protospacers matching the spacers. Similarly, we selected the minimum number of *B. fragilis* isolates that contained all identified spacers and only included them in the network. Selected MGEs and isolates are then used for building spacer-MGE and host-MGE networks. In the spacer-MGE network, spacer sequence clusters (called spacers for simplicity) and MGEs are represented as nodes and an edge is added between a spacer node and MGE node if the MGE contains a segment that matches the spacer (i.e., protospacer). In the host-MGE network, an edge is added to a host and a MGE if the host and MGE pair contain at least one matching protospacer and spacer. For MGEs that are phages (or prophages), we applied PhaGCN [45] to assign their taxonomic groups (ICTV [69] families). All visualizations and manual inspection of the networks were performed using Cytoscape [70].

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-022-08770-8.

Additional file 1: Additional file 1: Supplementary information. Supplementary Figure 1 shows the heatmap visualization of the spacer-MGE network and Supplementary Figure 2 shows the heatmap visualization of the host-MGE network.

Authors' contributions

TL carried out the implementation, analysis, and drafted the manuscript. KM participated in the analysis, and drafted the manuscript. YY conceived the study, participated in its design and implementation, participated in the analysis, and helped to draft the manuscript. All authors have read and approved the final manuscript.

Funding

The NIH grant 5R01AI143254 and NSF grant EF-2025451.

Availability of data and materials

We made the CRISPR-Cas annotations of all the genomes available for download along with visualization as a web resource at https://omics.informatics.indiana.edu/CRISPRone/Bfragilis. Phages and plasmids that were predicted to be targeted by the CRISPR-Cas systems, and their interaction networks are also available at the web resource.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Received: 15 February 2022 Accepted: 15 July 2022
Published online: 11 August 2022

References

1. O'Hara AM, Shanahan F. The gut flora as a forgotten organ. EMBO Rep. 2006;7(7):688–93.
2. West CE, Renz H, Jenmalm MC, Kozyrskyj AL, Allen KJ, Vuillermin P, et al. The gut microbiota and inflammatory noncommunicable diseases: associations and potentials for gut microbiota therapies. J Allergy Clin Immunol. 2015;135(1):3–13.
3. Quigley EM. Gut bacteria in health and disease. Gastroenterol Hepatol. 2013;9(9):560.
4. Labrie SJ, Samson JE, Moineau S. Bacteriophage resistance mechanisms. Nat Rev Microbiol. 2010;8(5):317-27. https://doi.org/10.1038/nrmicro2315.
5. Koonin EV, Makarova KS, Wolf YI. Evolutionary Genomics of Defense Systems in Archaea and Bacteria. Annu Rev Microbiol. 2017;71(1):233-61. https://doi.org/10.1146/annurev-micro-090816-093830.
6. Doron S, Melamed S, Ofir G, Leavitt A, Lopatina A, Keren M, et al. Systematic discovery of antiphage defense systems in the microbial pangenome. Science. 2018;359(6379):03.
7. Barrangou R, Fremaux C, Deveau H, Richards M, Boyaval P, Moineau S, et al. CRISPR Provides Acquired Resistance Against Viruses in Prokaryotes. Science. 2007;315(5819):1709-12. https://doi.org/10.1126/science.1138140.
8. Garneau JE, Dupuis MÈ, Villion M, Romero DA, Barrangou R, Boyaval P, et al. The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. Nature. 2010;468(7320):67-71. https://doi.org/10.1038/nature09523.
9. Levy A, Goren MG, Yosef I, Auster O, Manor M, Amitai G, et al. CRISPR adaptation biases explain preference for acquisition of foreign DNA. Nature. 2015;520(7548):505-10. https://doi.org/10.1038/nature14302.

10. Shmakov SA, Makarova KS, Wolf YI, Severinov KV, Koonin EV. Systematic prediction of genes functionally linked to CRISPR-Cas systems by gene neighborhood analysis. Proc Natl Acad Sci USA. 2018;115(23):E5307-16.
11. Shmakov S, Smargon A, Scott D, Cox D, Pyzocha N, Yan W, et al. Diversity and evolution of class 2 CRISPR-Cas systems. Nat Rev Microbiol. 2017;15(3):169-82.
12. Makarova KS, Zhang F, Koonin EV. SnapShot: class 2 CRISPR-Cas systems. Cell. 2017;168(1):328–8.
13. Stern A, Sorek R. The phage-host arms race: Shaping the evolution of microbes. BioEssays. 2010;33(1):43-51. https://doi.org/10.1002/bies.201000071.
14. Takeuchi N, Wolf YI, Makarova KS, Koonin EV. Nature and Intensity of Selection Pressure on CRISPR-Associated Genes. J Bacteriol. 2011;194(5):1216-25. https://doi.org/10.1128/jb.06521-11.
15. 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(1):20-7. https://doi.org/10.1039/c4mb00438h.
16. Deveau H, Barrangou R, Garneau JE, Labonte J, Fremaux C, Boyaval P, et al. Phage Response to CRISPR-Encoded Resistance in Streptococcus thermophilus. J Bacteriol. 2007;190(4):1390-400. https://doi.org/10.1128/jb.01412-07.
17. Hynes AP, Rousseau GM, Agudelo D, Goulet A, Amigues B, Loehr J, et al. Widespread anti-CRISPR proteins in virulent bacteriophages inhibit a range of Cas9 proteins. Nat Commun. 2018;9(1):. https://doi.org/10.1038/s41467-018-05092-w.
18. Künne T, Zhu Y, da Silva F, Konstantinides N, McKenzie RE, Jackson RN, et al. Role of nucleotide identity in effective CRISPR target escape mutations. Nucleic Acids Res. 2018;46(19):10395-404. https://doi.org/10.1093/nar/gky687.
19. Bondy-Denomy J, Garcia B, Strum S, Du M, Rollins MF, Hidalgo-Reyes Y, et al. Multiple mechanisms for CRISPR–Cas inhibition by anti-CRISPR proteins. Nature. 2015;526(7571):136-9. https://doi.org/10.1038/nature15254.
20. Weinberger AD, Sun CL, Pluci?ski MM, Denef VJ, Thomas BC, Horvath P, et al. Persisting viral sequences shape microbial CRISPR-based immunity. PLoS Comput Biol. 2012;8(4):e1002475.
21. McGinn J, Marraffini LA. Molecular mechanisms of CRISPR–Cas spacer acquisition. Nat Rev Microbiol. 2019;17(1):7–12.
22. McGinn J, Marraffini LA. CRISPR-Cas Systems Optimize Their Immune Response by Specifying the Site of Spacer Integration. Mol Cell. 2016;64(3):616-23. https://doi.org/10.1016/j.molcel.2016.08.038.
23. Gudbergsdottir S, Deng L, Chen Z, Jensen JVK, Jensen LR, She Q, et al. Dynamic properties of the Sulfolobus CRISPR/Cas and CRISPR/Cmr systems when challenged with vector-borne viral and plasmid genes and protospacers. Mol Microbiol. 2010;79(1):35-49. https://doi.org/10.1111/j.1365-2958.2010.07452.x.
24. Garrett RA, Shah SA, Vestergaard G, Deng L, Gudbergsdottir S, Kenchappa CS, et al. CRISPR-based immune systems of the Sulfolobales: complexity and diversity. Biochem Soc Trans. 2011;39(1):51-7. https://doi.org/10.1042/bst0390051.
25. Achigar R, Magadán AH, Tremblay DM, Pianzzola MJ, Moineau S. Phage-host interactions in Streptococcus thermophilus: Genome analysis of phages isolated in Uruguay and ectopic spacer acquisition in CRISPR array. Sci Rep. 2017;7(1):. https://doi.org/10.1038/srep43438.
26. Pourcel C, Touchon M, Villeriot N, Vernadet JP, Couvin D, Toffano-Nioche C, et al. CRISPRCasdb a successor of CRISPRdb containing CRISPR arrays and cas genes from complete genome sequences, and tools to download and query lists of repeats and spacers. Nucleic Acids Res. 2020;48(D1):D535–44.
27. Garrett SC. Pruning and tending immune memories: spacer dynamics in the CRISPR array. Front Microbiol. 2021;12:739.
28. Lopez-Sanchez MJ, Sauvage E, Da Cunha V, Clermont D, Ratsima Hariniaina E, Gonzalez-Zorn B, et al. The highly dynamic CRISPR1 system of Streptococcus agalactiae controls the diversity of its mobilome. Mol Microbiol. 2012;85(6):1057–71.
29. Jansen R, van Embden JDA, Gaastra W, Schouls LM. Identification of genes that are associated with DNA repeats in prokaryotes. Mol Microbiol. 2002;43(6):1565-75. https://doi.org/10.1046/j.1365-2958.2002.02839.x.
30. Yosef I, Goren MG, Qimron U. Proteins and DNA elements essential for the CRISPR adaptation process in Escherichia coli. Nucleic Acids Res. 2012;40(12):5569-76. https://doi.org/10.1093/nar/gks216.
31. Zhang W, Zhu B, Xu J, Liu Y, Qiu E, Li Z, et al. Bacteroides fragilis protects against antibiotic-associated diarrhea in rats by modulating intestinal defenses. Front Immunol. 2018;9:1040.
32. Yekani M, Baghi HB, Naghili B, Vahed SZ, Sóki J, Memar MY. To resist and persist: Important factors in the pathogenesis of Bacteroides fragilis. Microb Pathog. 2020;149:104506.
33. Solomkin JS, Mazuski JE, Bradley JS, Rodvold KA, Goldstein EJ, Baron EJ, et al. Diagnosis and management of complicated intra-abdominal infection in adults and children: guidelines by the Surgical Infection Society and the Infectious Diseases Society of America. Surg Infect. 2010;11(1):79–109.
34. Wexler HM. The genus bacteroides. In: The Prokaryotes. Berlin, Heidelberg: Springer; 2014. p. 459–484.
35. Casterline BW, Hecht AL, Choi VM, Bubeck Wardenburg J. The Bacteroides fragilis pathogenicity island links virulence and strain competition. Gut Microbes. 2017;8(4):374–83.
36. Husain F, Tang K, Veeranagouda Y, Boente R, Patrick S, Blakely G, et al. Novel large-scale chromosomal transfer in Bacteroides fragilis contributes to its pan-genome and rapid environmental adaptation. Microb Genomics. 2017;3(11):e000136.
37. Tajkarimi M, Wexler HM. CRISPR-Cas systems in Bacteroides fragilis, an important pathobiont in the human gut microbiome. Front Microbiol. 8;2017:2234.
38. Zhao S, Lieberman TD, Poyet M, Kauffman KM, Gibbons SM, Groussin M, et al. Adaptive evolution within gut microbiomes of healthy people. Cell Host Microbe. 2019;25(5):656–67.
39. Makarova KS, Wolf YI, Iranzo J, Shmakov SA, Alkhnbashi OS, Brouns SJ, et al. Evolutionary classification of CRISPR–Cas systems: a burst of class 2 and derived variants. Nat Rev Microbiol. 2020;18(2):67–83.
40. Weinberger AD, Sun CL, Pluciński MM, Denef VJ, Thomas BC, Horvath P, et al. Persisting viral sequences shape microbial CRISPR-based immunity. PLoS Comput Biol. 2012;8(4):e1002475.
41. Lam TJ, Ye Y. Long reads reveal the diversification and dynamics of CRISPR reservoir in microbiomes. BMC Genomics. 2019;20(1):1–12.
42. Lee SM, Donaldson GP, Mikulski Z, Boyajian S, Ley K, Mazmanian SK. Bacterial colonization factors control specificity and stability of the gut microbiota. Nature. 2013;501(7467):426–9.
43. Verster AJ, Ross BD, Radey MC, Bao Y, Goodman AL, Mougous JD, et al. The landscape of type VI secretion across human gut microbiomes reveals its role in community composition. Cell Host Microbe. 2017;22(3):411–9.
44. 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(1):. https://doi.org/10.1186/s12859-017-1512-4.
45. Shang J, Jiang J, Sun Y. Bacteriophage classification for assembled contigs using graph convolutional network. Bioinformatics. 2021;37(Supplement 1):i25-33.
46. Bailey TL, Elkan C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proc Int Conf Intell Syst Mol Biol. 1994;2:28–36.
47. Wexler HM. Bacteroides: the good, the bad, and the nitty-gritty. Clin Microbiol Rev. 2007;20(4):593–621.
48. Arbas SM, Narayanasamy S, Herold M, Lebrun LA, Hoopmann MR, Li S, et al. Roles of bacteriophages, plasmids and CRISPR immunity in microbial community dynamics revealed using time-series integrated meta-omics. Nat Microbiol. 2020;6(1):123-35. https://doi.org/10.1038/s41564-020-00794-8.
49. Touchon M, Rocha EP. The small, slow and specialized CRISPR and anti-CRISPR of Escherichia and Salmonella. PloS ONE. 2010;5(6):e11126.
50. Nicholson TJ, Jackson SA, Croft BI, Staals RH, Fineran PC, Brown CM. Bioinformatic evidence of widespread priming in type I and II CRISPR-Cas systems. RNA Biol. 2019;16(4):566–76.
51. Wimmer F, Beisel CL. CRISPR-Cas systems and the paradox of self-targeting spacers. Front Microbiol. 2020;10:3078.
52. Shmakov SA, Sitnik V, Makarova KS, Wolf YI, Severinov KV, Koonin EV. The CRISPR spacer space is dominated by sequences from species-specific mobilomes. MBio. 2017;8(5):e01397–17.

53. Shmakov SA, Wolf YI, Savitskaya E, Severinov KV, Koonin EV. Mapping CRISPR spaceromes reveals vast host-specific viromes of prokaryotes. Commun Biol. 2020;3(1):1–9.
54. Mortensen K, Lam TJ, Ye Y. Comparison of CRISPR-Cas immune systems in healthcare-related pathogens. Front Microbiol. 2021;12:758782.
55. Lam TJ, Ye Y. CRISPRs for strain tracking and their application to microbiota transplantation data analysis. CRISPR J. 2019;2(1):41–50.
56. Barrangou R, Dudley EG. CRISPR-based typing and next-generation tracking technologies. Ann Rev Food Sci Technol. 2016;7:395–411.
57. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30(15):2114–20.
58. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol. 2012;19(5):455–77.
59. Rho M, Tang H, Ye Y. FragGeneScan: predicting genes in short and error-prone reads. Nucleic Acids Res. 2010;38(20):e191–1.
60. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics. 2006;22(13): 1658-9. https://doi.org/10.1093/bioinformatics/btl158.
61. Rho M, Wu YW, Tang H, Doak TG, Ye Y. Diverse CRISPRs Evolving in Human Microbiomes. PLoS Genet. 2012;8(6):e1002441. https://doi.org/10.1371/journal.pgen.1002441.
62. Camarillo-Guerrero LF, Almeida A, Rangel-Pineros G, Finn RD, Lawley TD. Massive expansion of human gut bacteriophage diversity. Cell. 2021;184(4):1098–109.
63. Gao NL, Zhang C, Zhang Z, Hu S, Lercher MJ, Zhao XM, et al. MVP: a microbe–phage interaction database. Nucleic Acids Res. 2018;46(D1): D700–7.
64. Goodacre N, Aljanahi A, Nandakumar S, Mikailov M, Khan AS. A reference viral database (RVDB) to enhance bioinformatics analysis of high-throughput sequencing for novel virus detection. MSphere. 2018;3(2):e00069–18.
65. Douarre PE, Mallet L, Radomski N, Felten A, Mistou MY. Analysis of COMPASS, a new comprehensive plasmid database revealed prevalence of multireplicon and extensive diversity of IncF plasmids. Front Microbiol. 2020;11:483.
66. Galata V, Fehlmann T, Backes C, Keller A. PLSDB: a resource of complete bacterial plasmids. Nucleic Acids Res. 2019;47(D1):D195–202.
67. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST: architecture and applications. BMC Bioinformatics. 2009;10(1):421. https://doi.org/10.1186/1471-2105-10-421.
68. Olm MR, Brown CT, Brooks B, Banfield JF. dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. ISME J. 2017;11(12):2864–8.
69. Lefkowitz EJ, Dempsey DM, Hendrickson RC, Orton RJ, Siddell SG, Smith DB. Virus taxonomy: the database of the International Committee on Taxonomy of Viruses (ICTV). Nucleic Acids Res. 2018;46(D1):D708–17.
70. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003;13(11):2498–504.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.