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

Tony J. Lam[1], Kate Mortensen[1], and Yuzhen Ye[1*]

[1]School of Informatics, Computing and Engineering, Indiana University, Bloomington, IN, USA

[*]Corresponding author: Yuzhen Ye, 700 N. Woodlawn Avenue, Bloomington, IN 47408, yye@indiana.edu

Abstract

The CRISPR-Cas systems are important prokaryotic adaptive immune systems and effectively record the arms-race between bacteria and invading mobile elements in their CRISPR arrays. Using data from culture-based population genomics and metagenomics, we investigated the CRISPR-Cas systems associated with *Bacteriodes fragilis*, an important gut bacterium, and studied their diversity and dynamics. We analyzed genomes of 601 *B. fragilis* isolates derived from 12 healthy individuals, among which include time-series isolates from 6 individuals. In addition, we analyzed 222 reference *B. fragilis* genomes. Three different types of CRISPR-Cas systems (Type I-B, II-C and III-B) were found in analyzed B. fragilis genomes, with Type III-B being the most prevalent whereas the Type II-C being the most dynamic with varying arrays among isolates. Our graph-based summary and visualization of CRISPR arrays provided a holistic view of the organization and the composition of the CRISPR spacers. We observed insertions of different new spacers at the leader ends of the CRISPR arrays of Type II-C CRISPR-Cas systems in different *B. fragilis* sub-populations in one individual, an example of *B. fragilis*' adaptation and parallel evolution within individual microbiomes. A network of *B. fragilis* and its predicted invaders also revealed microdynamics of *B. fragilis* CRISPR array, with hairball-like structures representing spacers with multi-target capabilities and network modules revealing the collective targeting of selected mobile elements by many spacers. This work demonstrates the power of using culture-based population genomics to reveal the activities and evolution of the CRISPR-Cas systems of the important gut bacterium *B. fragilis* in human population.

Introduction

Microorganisms play a crucial role in the human health, especially those of the human gut microbiome [1]. A majority of endosymbiotic relationships between the human microbiota and host reside in the gut [2], a complex network comprised of hundreds of species [3].

Gut bacteria benefit their host by providing important secondary metabolites among others, and in turn gut bacteria depend on their host to maintain a stable ecosystem and provide accessible nutrients. In addition, the gut flora is thought to play an important role in the mucosal immune system [1]. The mutually beneficial interactions imply that gut bacteria may adapt to environmental changes imposed by their mammalian host such as the intake of antibiotics or dietary changes [4].

The ecological relationships between host and microorganisms in the human gut are ever changing. While a vast majority of relationships between gut microorganisms and the human host are symbiotic in nature, commensal organisms can become pathogenic depending on the state of the ecosystem. Such organisms are known as pathobionts. Approximately 25% of the human gut flora are gram-negative, obligate anaerobes of the *Bacteroides* phylum[5]. These bacteria are generally commensal but are occasionally known to cause infection. *Bacteroides fragilis* in particular is known for its pathobiont nature and accounts for a disproportionate amount of *Bacteroides* infections than its representative population size in the *Bacteroides* phylum residing in the gut [6]. While such infections are considered to be medically severe, previous studies have shown that *B. fragilis* can facilitate in the alleviation of certain disease conditions, thereby restoring systemic immune defects. These positive attributions fuel the proposed use of *B. fragilis* as a potential probiotic therapy [7]. This interesting bacteria-host behavior has inspired further investigation to understand the dynamics at play.

The extent to which commensal bacteria diversify and adapt within the human gut was examined by Zhao et al [4] using a combination of culture-based population genomics and metagenomics. This study provides insight on within-microbiome evolution of *B. Fragilis* and found that intra-individual *B. fragilis* populations contain substantial *de novo* nucleotide and mobile element diversity, preserving years of within-person history. Mammalian hosts and their gut symbionts were considered to have parallel evolutionary histories, as represented by their congruent phylogeny [8]. Such co-phylogenetic patterns can be explained by co-evolution of the mammalian hosts and their bacteria, or the hosts' allopatric speciation

(which limits the inter-host bacterial dispersal and genomic recombination) without relying on co-evolution [9].

Aside from bacteria-host interactions, gut bacteria are constantly engaged in an evolutionary arms-race with mobile genetic elements (MGEs), both of phage and plasmid origins. CRISPR-Cas systems are important adaptive immune systems that bacteria and archaea use to protect against invaders such as phage and plasmids [10, 11, 12]. Such systems are highly prevalent, existing in approximately half of bacterial and most of the archaeal genera [13, 14, 15, 16, 17]. The extreme diversity of CRISPR-Cas systems is reflected by the ever-changing classification scheme, owing to the constant discovery of new CRISPR-Cas system types and subtypes [11, 18, 19]. System types can be grouped into two main classes: Class I and Class II. Class I includes types I, II and V and use a complex of multiple Cas proteins to degrade foreign nucleic acids. Class II includes 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) [20]. The diversity of CRISPR-Cas systems in prokaryotes provides a fitness edge against invaders and is suggested to be a product of advantageous evolution [21, 22, 23]. Similarly, evolution of invaders has been observed to occur in tandem with host adaptive immunity as to evade host defense mechanisms such as anti-CRISPR genes [10, 11, 24, 25, 26, 27]. CRISPR-Cas systems have utility outside the scope of immunity. These systems are conveniently useful for a wide variety of applications, including nucleic acid editing [28, 29] and recently for detection of SARS-CoV-2 [30].

The mechanism necessary for recognition of foreign genetic material lies in the anatomy of the CRISPR-Cas system itself. CRISPR arrays are comprised of short DNA segments, known as spacers, and these provide a cornerstone to CRISPR-Cas derived adaptive immunity. Spacers, which are originally segments of invaders' genomes, retain the memory of past immunological encounters, and are primarily acquired as a result of Cas protein complex mediated acquisition [11]. Newly acquired spacers are typically integrated towards the leader ends of arrays [31]. Additionally, leader sequences usually found upstream of CRISPR arrays

are attributed to the efficiency of CRISPR-Cas derived immune response [32]. Several studies have also suggested that spacer acquisition remains possible through several alternative means such as homologous recombination [31, 33, 34], and ectopic spacer integration where spacers are inserted into middle of arrays as a result of leader sequence mutations [32, 35].

Bacteroides species is one of the most prevalent anaerobic bacteria in the lower intestinal tract. One member of the Bacteroides genus is *B. fragilis*, an opportunistic pathogen, accounting for a small fraction (about 2%) of the total gut Bacteroides, but it is the agent of $> 70\%$ of Bacteroides infections (causing anaerobic bacteremia and intraabdominal abscesses, and surgical site infections following colorectal surgery [36, 37, 38]). A previous study of 109 reference genomes reported four CRISPR-Cas systems in *B. fragilis*, three CRISPR-Cas system types and a fourth which they described as an orphaned CRISPR array [6]. Taking advantage of the increasing number *B. fragilis* reference genomes, and more importantly, the large number of *B. fragilis* isolates from 12 individuals, we re-investigated the CRISPR-Cas systems in B. fragilis. Similarly, our study revealed the existence of type I-B, II-C and III-B CRISPR-Cas systems in *B. fragilis* genomes; however, based on the analysis of the spacer turnover, we believe the fourth putative one reported in [6] that having no adjacent *cas* genes was a false one that superficially reassemble the repeat-spacer structure of genuine CRISPR arrays. More importantly, the availability of hundreds of time-resolved genomes from *B. fragilis* isolates from 6 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. Insights into how *B. fragilis* interacts with its invaders, as well as its CRISPR-Cas systems confer immunity help provide insights into factors that contribute to *B. fragilis* virulence, horizontal gene transfer, and evolution.

Methods

Genomic data processing and assembly

Reads from 601 *B. fragilis* isolates from the Zhao et al. study [4] were downloaded from the NCBI BioProject Accession PRJNA524913, henceforth referred to as the 'Zhao2019 dataset'. Raw shotgun sequencing reads had low quality reads and Illumina Truseq2 adapters trimmed using Trimmomatic v0.39 [39] (parameters used: LEADING:5 TRAILING:5 SLIDINGWINDOW:4:10 MINLEN:60). Trimmed reads were then assembled using SPAdes v3.12 [40] with default settings. FragGeneScan [41] was 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 CRISPRone website `http://omics.informatics.indiana.edu/CRISPRone/Bfragilis`.

Characterization of CRISPR-Cas systems

To identify CRISPR-Cas systems in *B. fragilis* genomes, we utilized CRISPRone [42] 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 with 90% sequence identity. Consensus reference repeats were then used as input for CRISPRAlign [43], 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 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 [44] 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. Spacers sequences that clustered together were considered identical spacer sequences. Spacer clusters were reserved for downstream computation of spacer content heterogeneity (Figure 1A) and construction of compressed spacer graphs (Figure 1B).

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. To help filter out 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. Here we define spacer content heterogeneity score as:

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

Where n is defined as the number 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 its removal.

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

Compressed spacer graphs [45] were constructed for each CRISPR-Cas type to summarize and illustrate CRISPR array dynamics. To construct compressed spacer graphs, for a given group of CRISPR arrays, all spacer sequences are clustered to 90% sequence identity to identify representative spacer clusters of similar/identical spacers. Then, 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 structure, the spacer graph was then simplified by collapsing neighboring nodes if two neighboring nodes share an "in-degree" and "out-degree" equal to or less than one (Figure 1B). Compressed spacer graphs highlight CRISPR array structure and dynamics (e.g. branching structures representing spacer gain and loss). Arrays that share no spacers will result in disconnected components in the compressed spacer graph.

Mobile Genetic Element Databases

A collection of mobile genetic element (MGE) databases were collected, including phage and plasmid databases. The phage databases included the Gut Phage Database [46] (GPD), MicrobeVersusPhage [47] (MVP) database, and the reference viral database [48] (RVDB). The plasmid databases included the Comprehensive and Complete Plasmid Database [49] (COMPASS), and PLSDB [50]. 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

Spacer sequences were first extracted from CRISPRone outputs. Spacer sequences were then queried against the MGE database using BLASTN [51] to search for putative Host-MGE interactions. 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. Due to the use of multiple MGE databases, subject sequences were dereplicated using dRep [52] with default parameters to obtain a dereplicated reference set of subject targets. Only BLASTN hits matching the dereplicated reference subject set were kept for downstream analysis. All remaining results were considered to represent a positive association between a given CRISPR spacer and their protospacer containing MGE targets.

Using the filtered results, a CRISPR-based MGE-host network was constructed. Within a given network graph, spacer sequence clusters and MGEs are represented as nodes. An edge between a spacer node and MGE node was constructed if a given spacer has a matching protospacer found in the corresponding MGE. All visualizations and manual inspection of the networks were performed using Cytoscape [53].

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 reference genomes and 601 isolates from the Zhao2019 dataset. Our analysis showed that among all *B. fragilis* genomes, three types of CRISPR-Cas systems were identified, Type I-B, Type II-C, and Type III-B. Example

illustrations of these CRISPR-Cas systems are depicted in Figure 2A, and Table 1 shows the signature *cas* genes and the repeat sequence of the CRISPR arrays for these CRISPR-Cas systems. We note that besides the three 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. Among the discarded CRISPR groups include a putative fourth CRISPR-Cas system that was previously reported in *B. fragilis genomes* [6]. The 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.

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 (Figure 2B and Table 2). CRISPR-Cas system Type I-B was among the least prevalent with only isolates from two individuals (Subjects 07 and 08) containing this CRISPR-Cas system type. CRISPR-Cas systems Type II-C and Type III-B were similarly prevalent with isolates from six individuals (see Table 2 for the lists of subjects that contain these systems). No CRISPR-Cas systems were found to be present within Zhao2019 Subjects S04, S05, 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. Similarly, there was an observed lack of shared inter-individual spacer content, with majority of the spacers observed being individual specific (Figure 4), with the most common shared spacer being the anchor spacer on the trailer end of observed CRISPR arrays.

Spacer content heterogeneity score (Figure 2) and compressed spacer graph of Type I-B CRISPR-Cas systems (Figure 3A) found within Zhao2019 isolates shows that the CRISPR

array structure 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 (Figure 3B). Each branch within the spacer graph represents an unique CRISPR array structure; bottle-neck nodes (e.g. 9, 35, 40) 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 hetegenerity score shows similar trends among the *B. fragilis* reference genomes (see Figure 2C).

The Type II-C CRISPR-Cas systems found within the Zhao2019 isolates were among the most diverse between 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) which have been previously hypothesized as ancient spacers or anchor spacers [54, 55, 45]. The spacer sequence diversity can be seen in Figure 4, 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. While previous studies [56, 57, 42] have shown that intra-individual populations of *B. fragilis* are dominated by a single strain, our findings here show that in some cases many "lineages" or strains of *B. fragilis* remain present within the same individual at any given time point. For example, there are multiple "lineages" of *B. fragilis* containing different Type II-C CRISPR arrays in S02 at time point 1 (S02-0001), one of which was also present in S02 at a later time point (S02-0024, 24 days apart from the first time point) (see Figure 5). Similarly, there are also multiple lineages of *B. fragilis* containing diverse Type II-C CRISPR arrays in S06 (S06-0001) (see Figure 5).

Overall CRISPR array structures remain relatively stable across samples from the same individual, with some slight variations between observed CRISPR arrays (e.g Figure 4 S01, S02, S03, S06, S07). A notable example is the arrays of type III-B CRISPR-Cas systems found in S01 isolates (Figure 5A). 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 can be observed from samples from the same individual. For example, Figure 4 shows that various strains of *B. fragilis* with varying CRISPR array structure can be 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 converges into a single neighboring node. This pattern observed in the spacer graph is suggestive of multiple "lineages" that have gained an alternative leader end spacer compared to each other, specifically when the bacteria are being exposed to different MGEs and are evolving accordingly to the ob-

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

Interaction network of *B. fragilis* and its invaders

To better understand the dynamics between host and MGEs, we analyzed the potential interactions between *B. fragilis* and their invader MGEs. By matching spacer sequences to protospacer sequences found in MGEs, we were able to identify putative targets of *B. fragilis* CRISPR-Cas systems and infer putative host-MGE interactions. A computational network was then built using the putative host-MGE interactions, where an edge between spacer and MGE nodes was assigned if and only if a MGE contained a matching protospacer to a given spacer (Figure 6). Host-MGE interaction networks were then built for each CRISPR-Cas system and its putative hosts, as well as individual specific networks; See supplementary website for all computational networks included in this study.

Analyzing the putative CRISPR-Cas targets of both Zhao2019 and *B. fragilis* reference genomes, we found that Type I-B and Type II-C CRISPR-Cas systems favored the targeting of phages compared to plasmids. A total of 276 putative phage and 13 plasmid targets were identified for Type I-B CRISPR-Cas systems. In Type II-C CRISPR-Cas systems, 2093 putative phage and 127 putative plasmid targets were identified. In Type I-B and Type II-C CRISPR-Cas systems, phage targets outnumbered plasmid targets by a magnitude of about 7.58 and 9.77, respectively. In contrast, Type III-B CRISPR-Cas systems favored the targeting of plasmids over phages, where 2 putative phage and 11 putative plasmid targets were identified resulting in a 5.5 plasmid:phage target ratio.

A network analysis of Host-MGE associations showed varying levels of micro-dynamics within *B. fragilis* CRISPR-Cas systems. Several spacers formed hairball-like network structures, indicative of the multi-target functionality of those spacers. It is likely that the spacers have adapted to target conserved regions of several MGEs, enabling broad range targeting of potential invaders. Many of the spacers associated with the hairball-like structures were

situated at the leader end (i.e. newer spacers) of observed CRISPR-Cas systems. Additionally, many leader end spacer sequences were also observed to share targets between other spacers, targeting the same MGE but with different protospacers (Figure 7). In contrast, trailer end spacers (i.e. older spacers) have shown to have fewer MGE associations.

Discussion

Here we expanded upon previous works [6] 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. As a common commensal bacteria of the human gut microbiome, sometimes probiotic candidate and sometimes pathogen, the role of *B. fragilis* as one of the most virulent member of the *Bacteroides* genus should not be overlooked [58]. Part of *B. fragilis* virulence is due to its potent virulence factors, and as such, it becomes important to better understand the mechanisms and factors that contribute to its virulence, horizontal gene transfer, and evolution. 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.

Analyzing genomes from both NCBI reference genomes and the Zhao2019 dataset, we were able to update the evaluation of known CRISPR-Cas systems found within *B. fragilis*, where we found three types of CRISPR-Cas systems (Type I-B, Type II-C, and Type III-B) with various levels of distributions among the genomes. In our analysis, we also showed 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 by [6], 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.

By focusing our analysis on bacterial isolates, we were also able to reduce the chance

of artifacts (e.g., chimeric contigs) from the assembly process, reducing the noise in our analysis of the evolutionary processes in *B. fragilis* CRISPR-Cas systems. Our analysis showed 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 (Figure 4), 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 (Figure 5A). Unsurprisingly, most inter-individual CRISPR-Cas systems did not share many spacers between individuals, suggesting that strains found within individuals were derived from a common ancestor, but strains have since diverged (as evident in their spacer content and exposure to different MGEs) and have not been circulating between individuals. 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.

Interestingly, when exploring the interplay/dynamics of *B. fragilis* and its MGEs, we found that *B. fragilis* CRISR-Cas systems seemed to prefer targeting phage genomes over plasmid genomes. This was in contrast to some studies which found CRISPRs favoring the targeting of plasmids over phages [59, 60]. 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 sev-

eral 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 [61, 62]. Alternatively, it may be possible that these instances of multiple targeting is 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 [63, 64] 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. Of particular interest were funneling patterns which 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' areas of the CRISPR array.

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. While our works here can help better understand how *B. fragilis* adapts to MGE exposure,

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. Additionally, 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. By better understanding how *B. fragilis* and other pathobionts interact with its invading mobile elements will better understand its evolution and elements which lead to its pathogenicity.

Conflict of interest

The authors declare no competing interests.

References

[1] O'Hara AM, Shanahan F. The gut flora as a forgotten organ. EMBO reports. 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. Journal of Allergy and Clinical Immunology. 2015;135(1):3-13.

[3] Quigley EM. Gut bacteria in health and disease. Gastroenterology & hepatology. 2013;9(9):560.

[4] 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.

[5] Troy EB, Kasper DL. Beneficial effects of Bacteroides fragilis polysaccharides on the immune system. Frontiers in bioscience: a journal and virtual library. 2010;15:25.

[6] Tajkarimi M, Wexler HM. CRISPR-Cas systems in Bacteroides fragilis, an important pathobiont in the human gut microbiome. Frontiers in microbiology. 2017;8:2234.

[7] 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. Frontiers in immunology. 2018;9:1040.

[8] Youngblut N, Reischer G, Walters W, Schuster N, Walzer C, Stalder G, et al. Host diet and evolutionary history explain different aspects of gut microbiome diversity among vertebrate clades. Nature. Communications. 2019;10(1).

[9] Groussin M, Mazel F, Alm EJ. Co-evolution and co-speciation of host-gut Bacteria systems. Cell Host & Microbe. 2020;28(1):12-22.

[10] Labrie SJ, Samson JE, Moineau S. Bacteriophage resistance mechanisms. Nature Reviews Microbiology. 2010 mar;8(5):317-27. Available from: https://doi.org/10.1038/nrmicro2315.

[11] Koonin EV, Makarova KS, Wolf YI. Evolutionary Genomics of Defense Systems in Archaea and Bacteria. Annual Review of Microbiology. 2017 sep;71(1):233-61. Available from: https://doi.org/10.1146/annurev-micro-090816-093830.

[12] 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 03;359(6379).

[13] Alexander Bolotin ASSDE Benoit Quinquis. Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. Microbiology. 2005 aug;151(8):2551-61. Available from: https://doi.org/10.1099/mic.0.28048-0.

[14] Mojica FJM, Dez-Villaseor C, Garca-Martnez J, Soria E. Intervening Sequences of Regularly Spaced Prokaryotic Repeats Derive from Foreign Genetic Elements. Journal of Molecular Evolution. 2005 Feb;60(2):174-82. Available from: https://doi.org/10.1007/s00239-004-0046-3.

[15] Barrangou R, Fremaux C, Deveau H, Richards M, Boyaval P, Moineau S, et al. CRISPR Provides Acquired Resistance Against Viruses in Prokaryotes. Science. 2007 mar;315(5819):1709-12. Available from: https://doi.org/10.1126/science.1138140.

[16] 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 nov;468(7320):67-71. Available from: https://doi.org/10.1038/nature09523.

[17] 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 apr;520(7548):505-10. Available from: https://doi.org/10.1038/nature14302.

[18] 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 06;115(23):E5307-16.

[19] 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 03;15(3):169-82.

[20] Makarova KS, Zhang F, Koonin EV. SnapShot: class 2 CRISPR-Cas systems. Cell. 2017;168(1):328-8.

[21] Stern A, Sorek R. The phage-host arms race: Shaping the evolution of microbes. BioEssays. 2010 oct;33(1):43-51. Available from: https://doi.org/10.1002/bies.201000071.

[22] Takeuchi N, Wolf YI, Makarova KS, Koonin EV. Nature and Intensity of Selection Pressure on CRISPR-Associated Genes. Journal of Bacteriology. 2011 dec;194(5):1216-25. Available from: https://doi.org/10.1128/jb.06521-11.

[23] Koonin EV, Wolf YI. Evolution of the CRISPR-Cas adaptive immunity systems in prokaryotes: models and observations on virus–host coevolution. Molecular BioSystems. 2015;11(1):20-7. Available from: https://doi.org/10.1039/c4mb00438h.

[24] Deveau H, Barrangou R, Garneau JE, Labonte J, Fremaux C, Boyaval P, et al. Phage Response to CRISPR-Encoded Resistance in Streptococcus thermophilus. Journal of Bacteriology. 2007 dec;190(4):1390-400. Available from: https://doi.org/10.1128/jb.01412-07.

[25] 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. Nature Communications. 2018 jul;9(1). Available from: https://doi.org/10.1038/s41467-018-05092-w.

[26] 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 Research. 2018 aug;46(19):10395-404. Available from: https://doi.org/10.1093/nar/gky687.

[27] 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 sep;526(7571):136-9. Available from: https://doi.org/10.1038/nature15254.

[28] Cong L, Ran FA, Cox D, Lin S, Barretto R, Habib N, et al. Multiplex genome engineering using CRISPR/Cas systems. Science. 2013;339(6121):819-23.

[29] Cox DB, Gootenberg JS, Abudayyeh OO, Franklin B, Kellner MJ, Joung J, et al. RNA editing with CRISPR-Cas13. Science. 2017;358(6366):1019-27.

[30] Broughton JP, Deng X, Yu G, Fasching CL, Servellita V, Singh J, et al. CRISPR–Cas12-based detection of SARS-CoV-2. Nature biotechnology. 2020;38(7):870-4.

[31] 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.

[32] McGinn J, Marraffini LA. CRISPR-Cas Systems Optimize Their Immune Response by Specifying the Site of Spacer Integration. Molecular Cell. 2016 nov;64(3):616-23. Available from: https://doi.org/10.1016/j.molcel.2016.08.038.

[33] 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. Molecular Microbiology. 2010 nov;79(1):35-49. Available from: https://doi.org/10.1111/j.1365-2958.2010.07452.x.

[34] Garrett RA, Shah SA, Vestergaard G, Deng L, Gudbergsdottir S, Kenchappa CS, et al. CRISPR-based immune systems of the Sulfolobales: complexity and diversity. Biochemical Society Transactions. 2011 feb;39(1):51-7. Available from: https://doi.org/10.1042/bst0390051.

[35] 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. Scientific Reports. 2017 mar;7(1). Available from: https://doi.org/10.1038/srep43438.

[36] Yekani M, Baghi HB, Naghili B, Vahed SZ, Sóki J, Memar MY. To resist and persist: Important factors in the pathogenesis of Bacteroides fragilis. Microbial pathogenesis. 2020:104506.

[37] 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. Surgical infections. 2010;11(1):79-109.

[38] Wexler H, Geffen D. The genus Bacteroides. Rosenberg EY, DeLong EF, Thompson F, Lory S, Stackebrandt E et al. 2014:459-84.

[39] Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30(15):2114-20.

[40] 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. Journal of computational biology. 2012;19(5):455-77.

[41] Rho M, Tang H, Ye Y. FragGeneScan: predicting genes in short and error-prone reads. Nucleic acids research. 2010;38(20):e191-1.

[42] Zhang Q, Ye Y. Not all predicted CRISPR–Cas systems are equal: isolated cas genes and classes of CRISPR like elements. BMC Bioinformatics. 2017 Feb;18(1). Available from: `https://doi.org/10.1186/s12859-017-1512-4`.

[43] Rho M, Wu YW, Tang H, Doak TG, Ye Y. Diverse CRISPRs Evolving in Human Microbiomes. PLoS Genetics. 2012 jun;8(6):e1002441. Available from: `https://doi.org/10.1371/journal.pgen.1002441`.

[44] Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics. 2006 may;22(13):1658-9. Available from: `https://doi.org/10.1093/bioinformatics/btl158`.

[45] Lam TJ, Ye Y. Long reads reveal the diversification and dynamics of CRISPR reservoir in microbiomes. BMC genomics. 2019;20(1):1-12.

[46] 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.

[47] Gao NL, Zhang C, Zhang Z, Hu S, Lercher MJ, Zhao XM, et al. MVP: a microbe–phage interaction database. Nucleic acids research. 2018;46(D1):D700-7.

[48] 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.

[49] 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. Frontiers in microbiology. 2020;11:483.

[50] Galata V, Fehlmann T, Backes C, Keller A. PLSDB: a resource of complete bacterial plasmids. Nucleic acids research. 2019;47(D1):D195-202.

[51] Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST: architecture and applications. BMC Bioinformatics. 2009;10(1):421. Available from: `https://doi.org/10.1186/1471-2105-10-421`.

[52] 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. The ISME journal. 2017;11(12):2864-8.

[53] 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 research. 2003;13(11):2498-504.

[54] Weinberger AD, Sun CL, Pluciński MM, Denef VJ, Thomas BC, Horvath P, et al. Persisting viral sequences shape microbial CRISPR-based immunity. PLoS computational biology. 2012;8(4):e1002475.

[55] 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. Molecular microbiology. 2012;85(6):1057-71.

[56] 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.

[57] 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.

[58] Wexler HM. Bacteroides: the good, the bad, and the nitty-gritty. Clinical microbiology reviews. 2007;20(4):593-621.

[59] 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. Nature Microbiology. 2020 Nov;6(1):123-35. Available from: `https://doi.org/10.1038/s41564-020-00794-8`.

[60] Touchon M, Rocha EP. The small, slow and specialized CRISPR and anti-CRISPR of Escherichia and Salmonella. PloS one. 2010;5(6):e11126.

[61] 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 biology. 2019;16(4):566-76.

[62] Wimmer F, Beisel CL. CRISPR-Cas systems and the paradox of self-targeting spacers. Frontiers in microbiology. 2020;10:3078.

[63] 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.

[64] Shmakov SA, Wolf YI, Savitskaya E, Severinov KV, Koonin EV. Mapping CRISPR spaceromes reveals vast host-specific viromes of prokaryotes. Communications biology. 2020;3(1):1-9.

Tables

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*
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.

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		

Figures

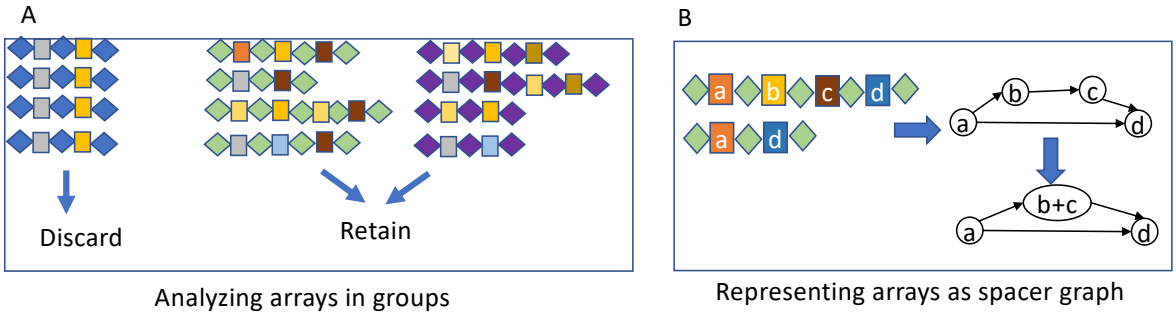

Figure 1: Approaches used for the identification and refinement of the CRISPR arrays and construction of spacer graphs. (A) CRISPR arrays are analyzed in groups such that each group shares identical or very similar repeats. CRISPR arrays containing no spacer variance or no CRISPR spacer hetergeneity are discarded. (B) Example of spacer sharing CRISPR arrays can be represented as a simplified graphical structure, which we call spacer graphs. Nodes represents unique spacer or a group of consecutive spacers. Edges between nodes are represented when the spacers represented by the node are adjacent to one another on any given CRISPR array used to construct the spacer graph.

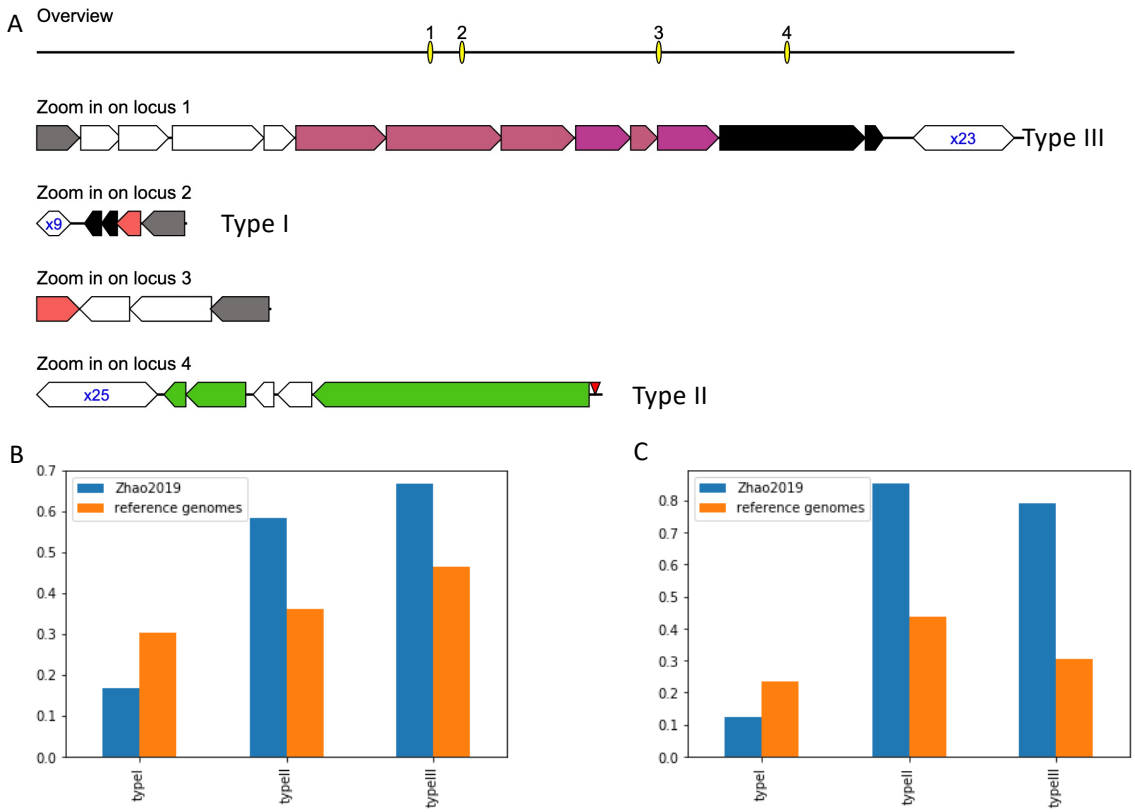

Figure 2: The three types of CRISPR-Cas systems found in *B. fragilis* genomes. (A) Representatives of *B. fragilis* CRISPR-Cas systems; example systems were obtained from the complete genome of *B. fragilis* strain S14 (accession number: GCA_001682215.1_ASM168221v1). (B) Prevalence of CRISPR-Cas systems between NCBI reference genomes and isolates from the Zhao2019 dataset. (C) Spacer content heterogeneity of CRISPR arrays between NCBI reference genomes and isolates from the Zhao2019 dataset.

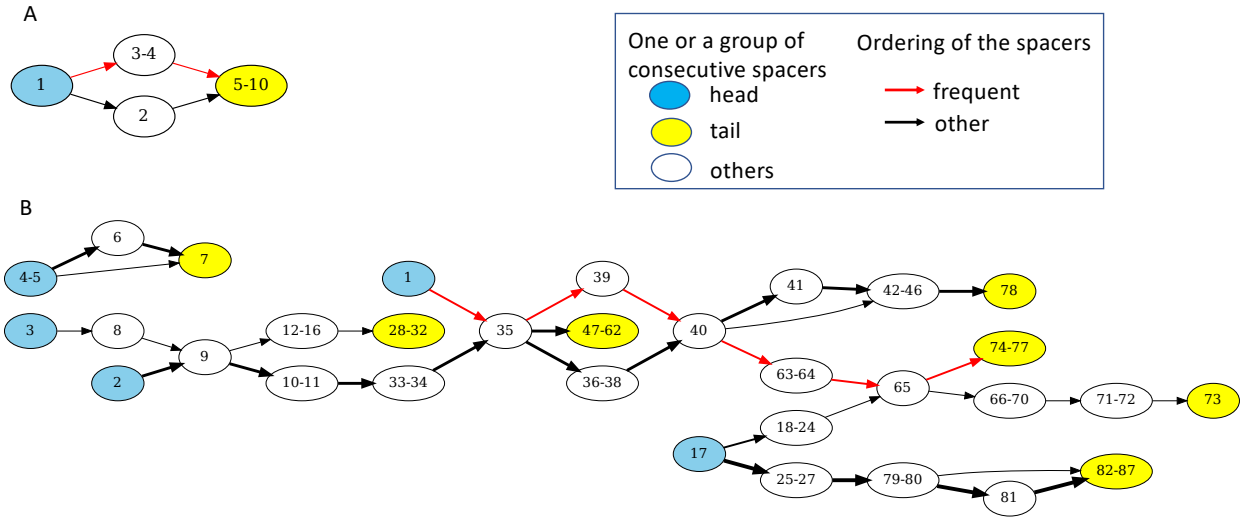

Figure 3: Compressed spacer graphs of Type I-B and Type III-B CRISPR-Cas systems in *B. fragilis*. Compressed spacer graphs are simplified models of multiple CRISPR arrays, where nodes are representative of consecutive arrays, and edges representative of adjacent spacers within a given set of CRISPR arrays. Nodes representing leader end spacers are highlighted in blue, and nodes representing trailer end spacers are shown in yellow. (A) Compressed spacer graph of Type I-B CRISPR-Cas systems. (B) Compressed spacer graph of Type III-B CRISPR-Cas system.

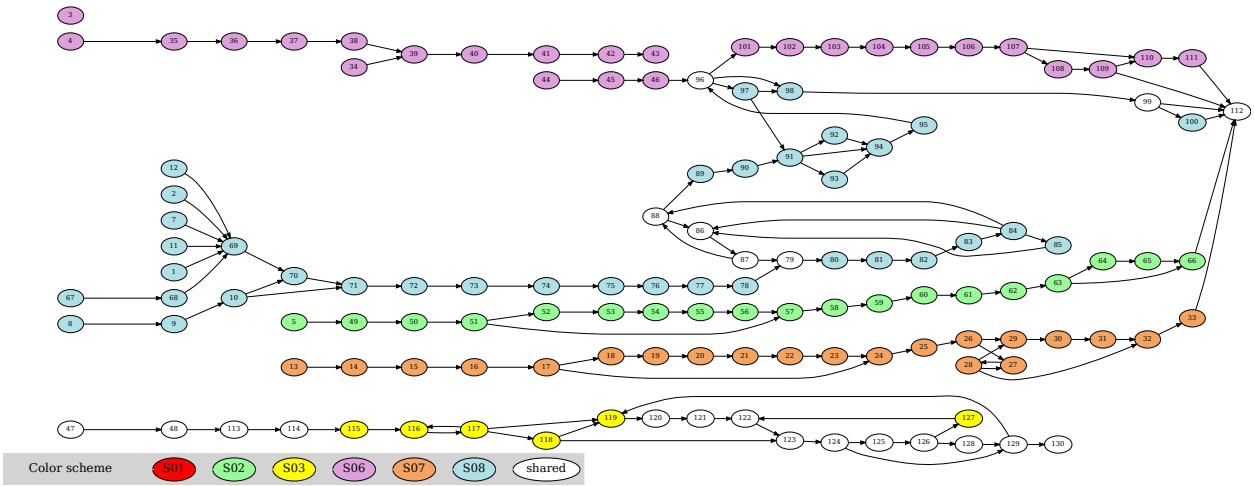

Figure 4: Spacer graph of the Type II-C CRISPR-Cas systems in *B. fragilis*. Spacers that are unique to individuals are highlighted in colors as indicated (see color scheme in the plot). Spacers shared by isolates from at least two individuals are shown colored as white. Each node is representative of a single unique spacer found in *B. fragilis* Type II-C CRISPR arrays.

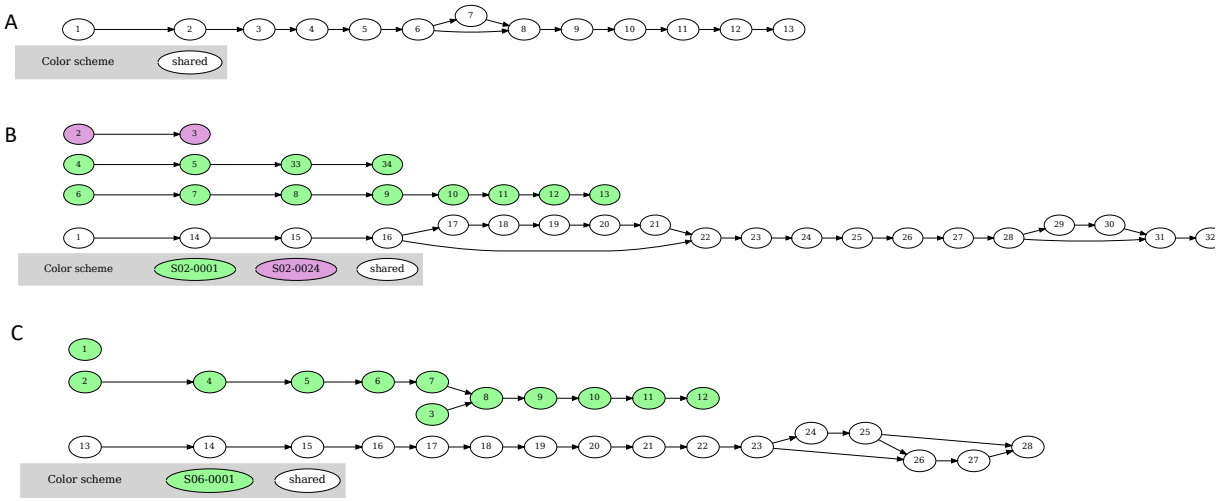

Figure 5: Intra-individual CRISPR array diversity. (A) Spacer graph of CRISPR arrays from individual S01 isolates, consisting of 10 time points, showing low spacer diversity. (B) Spacer graph from individual S02. Green nodes are unique spacers identified from time point 1, and purple nodes are from time point 24, white nodes represent spacers found in multiple time points. (C) Spacer graph from Individual S06. Green nodes are unique spacers identified from time point 1, and white nodes represent spacers found in multiple time points.

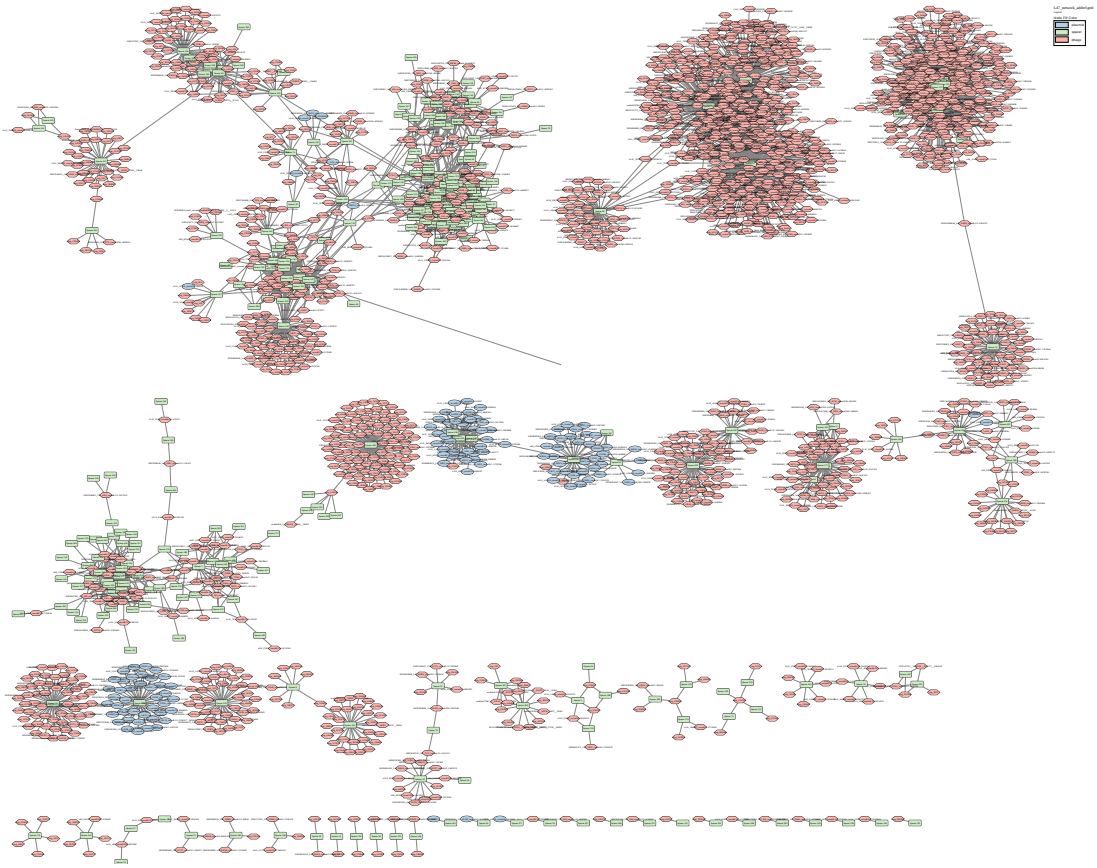

Figure 6: Spacer-MGE network of Type II-C CRISPR-Cas systems. Genomes included in the graph include those from the NCBI *B. fragilis* reference genomes and Zhao2019 isolates. Red nodes represent phage genomes, blue nodes represent plasmid genomes, green nodes represent Type II-C spacers found in *B. fragilis*. A edge is built between a spacer node and MGE node if a protospacer matching a given spacer is present within the MGE.

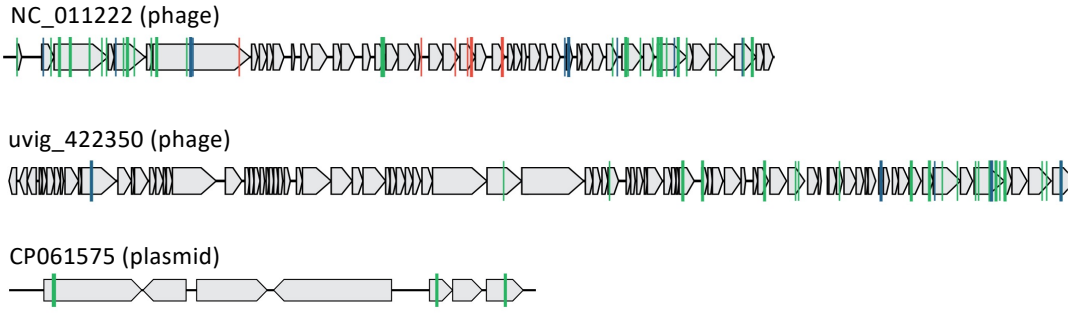

Figure 7: Representative invaders that contain protospacers matching CRISPR spacers identified in *B. fragilis*. Genes encoded by each MGE genome are shown as arrows. Vertical lines indicate protospacer locations within a given MGE genome. Green lines indicate protospacers associated with spacers found in type I-B CRISPR-Cas systems. Red lines indicate protospacers associated with spacers found in Type II-C CRISPR-Cas systems.