MICROBIAL GENOMICS

RESEARCH ARTICLE
Han *et al.*, *Microbial Genomics* 2022;8:000812
DOI 10.1099/mgen.0.000812

Characterization of CRISPR-Cas systems in *Bifidobacterium breve*

Xiao Han[1,2], Xingya Zhou[1,2], Zhangming Pei[1,2], Catherine Stanton[3,4,5], R. Paul Ross[3,4], Jianxin Zhao[1,2,6], Hao Zhang[1,2,6,7], Bo Yang[1,2,3,*] and Wei Chen[1,2,6]

Abstract

The clustered regularly interspaced short palindromic repeat (CRISPR)-CRISPR-associated protein (Cas) system is an important adaptive immune system for bacteria to resist foreign DNA infection, which has been widely used in genotyping and gene editing. To provide a theoretical basis for the application of the CRISPR-Cas system in *Bifidobacterium breve*, the occurrence and diversity of CRISPR-Cas systems were analysed in 150 *B. breve* strains. Specifically, 47% (71/150) of *B. breve* genomes possessed the CRISPR-Cas system, and type I-C CRISPR-Cas system was the most widely distributed among those strains. The spacer sequences present in *B. breve* can be used as a genotyping marker. Additionally, the phage assembly-related proteins were important targets of the type I-C CRISPR-Cas system in *B. breve*, and the protospacer adjacent motif sequences were further characterized in *B. breve* type I-C system as 5′-TTC-3′. All these results might provide a molecular basis for the development of endogenous genome editing tools in *B. breve*.

DATA SUMMARY

All genome sequences used in this study were available on the National Center for Biotechnology Information (NCBI) Genome database or SRA database. In total, 44 new genome sequences were deposited under the project accession no. PRJNA755456. The genome size, CDS number, GC content and accession numbers of the genomes used in this study were shown in Table S1 (available in the online version of this article). The CRISPR-Cas system type, repeat sequence, repeat length and the number of spacers of *B. breve* used in this study were shown in Table S2.

INTRODUCTION

Gut microbiota play an important role in the host metabolism and health maintenance. Early microbial exposure in the intestine is believed to promote the development of the systemic immune system in early life [1]. In addition to competing with other species for survival, a major challenge faced by bacteria is phage infection. In fact, bacteria and their bacteriophages are the most important members of the gut microbiome, and their struggle and co-evolution have been ongoing [2]. In addition to restriction-modification systems, the clustered regularly interspaced short palindromic repeats (CRISPR) is another important defence system for bacteria against phage infection. Initially, the CRISPR-CRISPR-associated proteins (Cas) systems were found to be widely present in bacteria and archaea, which were assigned to two classes, six types and 33 subtypes according to the presence/absence and arrangement of *cas* genes [3]. The immune function of the CRISPR-Cas system is mainly divided into the following three stages. (1) The adaptation stage. When the strain is invaded by phage,

Received 19 August 2021; Accepted 15 March 2022; Published 22 April 2022

Author affiliations: [1]State Key Laboratory of Food Science and Technology, Jiangnan University, Wuxi, PR China; [2]School of Food Science and Technology, Jiangnan University, Wuxi, PR China; [3]International Joint Research Laboratory for Pharmabiotics & Antibiotic Resistance, Jiangnan University, Wuxi, PR China; [4]APC Microbiome Ireland, University College Cork, Cork, Ireland; [5]Teagasc Food Research Centre, Moorepark, Co., Cork, Ireland; [6]National Engineering Research Center for Functional Food, Jiangnan University, Wuxi, PR China; [7]Wuxi Translational Medicine Research Center and Jiangsu Translational Medicine Research Institute Wuxi Branch, Wuxi, PR China.

***Correspondence:** Bo Yang, bo.yang@jiangnan.edu.cn

Keywords: *Bifidobacterium breve*; CRISPR-Cas system; genotyping; prophage.

Abbreviations: Acr, anti-CRISPR; bp, base pair; Cas, CRISPR-associated proteins; CCFM, Culture Collection of Food Microorganisms; CDS, coding sequences; CRISPR, clustered regularly interspaced short palindromic repeats; crRNA, CRISPR RNA; dCas9, dead Cas9; MLST, multi-locus sequence typing; NCBI, National Center for Biotechnology Information; ORFs, open reading frames; PAMs, protospacer adjacent motifs; pre-crRNA, CRISPR RNA precursor; TMP, tape measure protein.

In total, 44 new genome sequences were deposited under the project accession no. PRJNA755456. The accession numbers of the genomes used in this study were shown in Table S1.

Data statement: Two supplementary tables are available with the online version of this article.

Impact Statement

The CRISPR-Cas systems identified in *B. breve* showed high occurrence and diversity, and type I-C CRISPR-Cas systems were relatively common in *B. breve*. Type I-E and I-U CRISPR-Cas systems were identified, which were not reported in *B. breve* previously. In addition, an 'undetermined' type CRISPR-Cas system in which only the repeat-spacer arrays were detected but no *cas* genes was identified, seemed to evolve from the loss of *cas* genes in subtype I-C during genetic recombination or the repeat-spacer arrays obtained through horizontal gene transfer. Variable spacers can be used as a reference for genotyping, and the match between the spacers and the prophages indicated that those CRISPR loci may be active in resisting foreign DNA infection. By tracing the source of protospacers, the genes encoding phage assembly-related proteins were considered to be important targets of subtype I-C CRISPR-Cas system in *B. breve*. Hence, the characterization of CRISPR systems in *B. breve* could provide a perspective for further analysis on the defence strategy of strains against bacteriophages, and provide a reference for the use of endogenous CRISPR systems for genetic manipulation.

Cas1 and Cas2 scan and recognize the invading nucleic acid and integrate small fragments of the invading DNA into the genome. (2) The expression stage. The CRISPR locus is transcribed to form a long CRISPR RNA precursor (pre-crRNA), which is then processed into mature crRNA (CRISPR RNA) by different Cas proteases. (3) The interference stage. Mature crRNA and specific Cas protease form a complex, which will find the target through base complementary pairing and cut the foreign DNA by the nuclease activity of the Cas protein [4]. Due to the precise and efficient genome-editing capabilities of Cas9 nuclease, it has been widely used in gene modification in eukaryotes and prokaryotes [5–7].

As one of the first microbial colonizers in the gastrointestinal tract of infants, bifidobacteria have been confirmed to be useful in the prevention and treatment of various diseases, such as *Helicobacter pylori* infection, allergic disease, inflammatory bowel disease and irritable bowel syndrome [8]. *Bifidobacterium breve* is an important species, which is usually isolated from the gastrointestinal tract of infants, especially breastfed infants [9]. Some *B. breve* strains have been proven to be effective against a series of diseases in animal models and clinical trials, including necrotizing enterocolitis, coeliac disease, paediatric obesity and allergies [10–13]. Targeted genetic modification and transformation of *B. breve* will contribute to enhancing the benefits of probiotics and the development of next-generation food micro-organisms. Currently, the nuclease spCas9 from *Streptococcus pyogenes* is the most widely used in CRISPR-Cas-based gene-editing systems, but some studies have shown that low transformation efficiencies were obtained with Cas9/dead Cas9 (dCas9) bearing plasmids due to its toxicity [14, 15]. There is a report that the expression of exogenous dCas9 has the risk of inducing abnormal cell morphology in *Escherichia coli* [16]. In addition, the existence of the restriction-modification system in bifidobacteria brings difficulties to the action of exogenous nucleases [17, 18]. Given the high occurrence of CRISPR-Cas systems in bifidobacteria, using the endogenous CRISPR-Cas systems seems to be an alternative for gene editing. Therefore, exploring the detailed characterization of the CRISPR-Cas systems in *B. breve* is necessary for developing subsequent gene-editing tools.

In this study, the incidence of the CRISPR-Cas system in *B. breve* was analysed. CRISPR-Cas system subtypes, locus integrity, characteristics of the repeats and spacers as well as the PAM sequences were assessed, which will provide a reference for the development of spacer-based genotyping and gene-editing systems in *B. breve*.

METHODS

CRISPR locus identification

A total of 150 *B. breve* genomes were used in this study, including 104 draft genomes sequenced in our lab previously [19] and 46 publicly available *B. breve* genomes from GenBank database until April 2021 (Table S1). CRISPRCasFinder (https://crisprcas.i2bc.paris-saclay.fr/CrisprCasFinder/Index) was used to analyse the CRISPRs and *cas* genes [20]. Through manual screening, only the results with evidence-level four were retained. The types of CRISPR-Cas system were predicted and that type only containing repeat-spacer arrays but without *cas* gene in the genome was defined as type 'undetermined'. CRISPRDetect (http://crispr.otago.ac.nz/CRISPRDetect/predict_crispr_array.html) was used to verify the CRISPR arrays [21].

Characterization of CRISPRs

Spacer and repeat data were derived from online prediction as described and graphed using GraphPad Prism 8.0. The secondary structure of repeat sequence was predicted using RNAfold WebServer (http://rna.tbi.univie.ac.at/cgi-bin/RNAWebSuite/RNAfold.cgi). MEGA-X was used to construct a phylogenetic analysis tree based on Cas1 amino acid sequences [22]. Alignments were performed using MUSCLE algorithm and the neighbor-joining tree was depicted with 1000 bootstrap replicates [23]. The optimization of the phylogenetic tree was performed using ITOL (https://itol.embl.de) [24].

Prophage identification

Prophages in *B. breve* genomes were predicted using PHASTER (http://phaster.ca) [25]. As described in PHASTER, the integrity of the prophage region in the genome was scored based on the size of the region, the number of total coding sequences (CDS), and the presence of each phage-related protein (such as 'capsid', 'head', 'integrase'). According to the above score, the predicted prophage of this region was marked as follows: incomplete (less than 70), questionable (between 70 to 90) or intact (greater than 90). The heatmap of prophages in *B. breve* genomes was completed with TBtools [26]. The target of the CRISPR spacer in the prophage was identified by BLASTn search. The results with the percentage of identical matches greater than 95% and an E value less than e^{-5} were considered as statistical reliability. The ORFs of the prophage were predicted using Prokka and then the protein targeted was annotated by the spacers through BLASTp search in NCBI nr database [27]. The prediction of PAMs was performed based on 8 bp extracted from upstream and downstream of the identified protospacer. Then the WebLogo server was used to visualize the predicted PAM sequence (https://weblogo.berkeley.edu/logo.cgi) [28].

RESULTS

Occurrence and diversity of CRISPR in *B. breve*

In order to characterize the occurrence and diversity of CRISPR in *B. breve*, 150 *B. breve* genomes were investigated by *in silico* analysis (including 46 publicly available complete genomes from GenBank database). Overall, CRISPR repeat/spacer loci were identified in 81 out of 150 (54%) genomes (Fig. 1a). Furthermore, 47% of *B. breve* strains (71/150) harboured intact CRISPR-Cas systems (CRISPR repeat/spacer loci accompanied by *cas* gene) in their genomes. According to different *cas* genes, CRISPR-Cas systems can be divided into two classes, six types and 33 subtypes [3]. Type I systems were the most common CRISPR-Cas system in *B. breve*. Specifically, 64 type I-C systems, 10 type I-E systems, and two type I-U systems were identified in 150 strains (Fig. 1a). *B. breve* genomes from GenBank database were only identified as harbouring type I-C systems, while the strains isolated in our lab from Chinese subjects presented a variety of CRISPR-Cas systems.

In order to assess the integrity of the CRISPR-Cas systems in *B. breve*, some representative strains for CRISPR loci visualization were selected (Fig. 2). Interestingly, there were two different subtype systems in the same strain, similar to previous reports [23, 29, 30]. Subtype I-C and I-E systems coexisted in some strains such as AHWH11M1, AHWH13M7, GuXi201667, Sunxiaoran13 and ZJHZD20M12 (Fig. 2, Table S2). Complete type I-C systems were identified in FHNXY43M2 and UCC2003. NRBB52 harbouring type I-C system and HeNJZ2M1 harbouring type I-E system showed the absence of Cas2 in the CRISPR loci. Surprisingly, no *cas3* gene was identified in type I-E CRISPR loci in *B. breve* strains in this study. Furthermore, partial repeat-spacer arrays in *B. breve* AHWH13M7, HuNan2016415 and HeNJZ2M1 were inserted into the *cas* gene clusters, and repeat-spacer arrays in *B. breve* UCC2003 and NRBB52 appeared truncated.

Cas1 and Cas2 are two conserved proteins in all CRISPR-Cas systems [31]. Due to the absence of Cas2 protein in some strains, a phylogenetic analysis was further performed based on Cas1 protein to examine the conservation and difference of CRISPR-Cas systems in *B. breve* (Fig. 3). Cas1 from the same subtype tended to cluster together. Subtype I-C branch further diverged into two main subclusters, interestingly, the strains in the right subcluster were all isolated from Chinese subjects except for *B. breve* lw01.

Analysis of CRISPR repeat-spacer arrays

In the expression stage, repeat-spacer arrays are transcribed as a precursor transcript (pre-crRNA), which is then processed by nucleases into mature CRISPR RNAs (crRNAs) [32]. The size of repeat sequences is usually conserved within CRISPR subtypes. In *B. breve*, the length of repeat sequences was 33 nucleotides for subtype I-C, while it was 36 nucleotides for subtype I-U. The shortest length of repeat sequences was observed for subtype I-E, which was 28 nucleotides (Fig. 1b). When the type only detected with the repeat-spacer arrays but no *cas* genes, it was defined as ‘undetermined’. Interestingly, either the length or the secondary structure of repeat sequences in type ‘undetermined’ was more similar to that in subtype I-C (Fig. 1d), indicating that the type ‘undetermined’ in *B. breve* might evolve from subtype I-C with the loss of *cas* genes during genetic recombination or the repeat-spacer arrays obtained through horizontal gene transfer.

Spacer arrays are the memory of foreign DNA infection generated during the immune function of CRISPR-Cas systems and are the hypervariable regions of the CRISPR array [33]. Each spacer represented an immune record, and the number of spacers was highly diverse across the different subtypes. Subtype I-C seemed to be the most active CRISPR-Cas system in *B. breve*, with an average of 52 spacers per locus. In total, 127 spacers in the *B. breve* NMGEL2M1 harbouring subtype I-C system represented the largest number of spacers among all the strains assayed. There was no CRISPR array around the *cas* gene in the subtype I-C system in *B. breve* Sunxiaoran13, which represented the smallest number of spacers in *B. breve*. The number of spacers for subtypes I-E and I-U was similar, with an average of 26 and 30 spacers per locus, respectively (Fig. 1c). Due to the absence of *cas* genes and the inability to obtain new spacers, type ‘undetermined’ had the least number of spacers (with an average of 15 spacers per locus).

Subtype I-C represented the highest occurrence and the most active CRISPR-Cas system in *B. breve*, some strains harbouring type I-C systems were further selected for spacer visualization to analyse the evolutionary relationship among them. In the adaptation

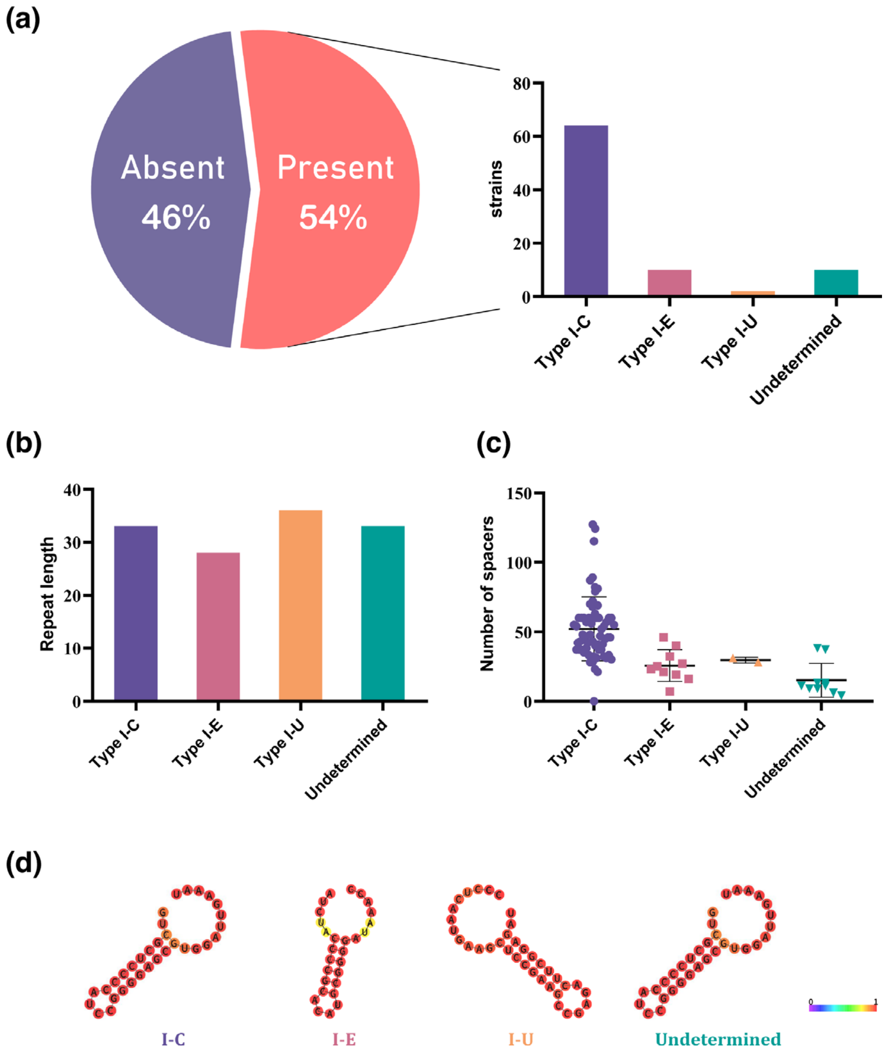

Fig. 1. CRISPR-Cas systems in *B. breve*. (a) Occurrence of CRISPR array. (b) Repeat length of different CRISPR-Cas subtypes. (c) The number of spacers in each CRISPR-Cas subtype. (d) Secondary structure of repeat sequences of different subtypes. Secondary structure prediction was performed by the RNAfold web server. Bases were coloured by base-pair probability.

stage, new spacer integration often occurred at the leader end of CRISPR structure. Therefore, the most recently integrated spacer will be close to the leader, while the earliest integrated spacer will be far from the leader in the genome. The spacers were sorted according to the order of the spacer integration, in which the oldest spacer was the first, and the newest spacer was the last. The first spacers among 11 strains were identical (Fig. 4), which indicated they might share the same origin. But the diversity presented from the second spacer suggested that they were evolving in different directions. *B. breve* FFJND6M1, FGZ3I1M6, FHNFQ23M3 and JSWX39M4 shared the first to 24th spacers, indicating that they had a common ancestor. Among them, FFJND6M1 may be in a different niche within a certain period as lacking the 25th to 27th spacer compared with the other three strains. Furthermore, the difference between new spacers captured by strains FFJND6M1 and FGZ3I1M6 in the latest timepoint evidence the activity of these CRISPR loci. The identical arrangement of spacers in *B. breve* NRBB08, NRBB18 and NRBB19 indicated that their evolutionary relationship was extremely close, and may even be with the same strain, given that their isolation sources were the same [17]. Interestingly, three consecutive identical spacers were observed in NCFB2258, possibly due to multiple infections by the same phage.

Traceability of spacer targeting sequences

Each spacer is an immune event in which the strain resisted infection by foreign DNA, which can explain its origin and the changes in ecological niches. Unfortunately, few bifidobacterial phage genomes in public databases presented difficulties for spacer traceability. Prophages are bacteriophage genomes integrated into the bacterial genome. Therefore, in order to

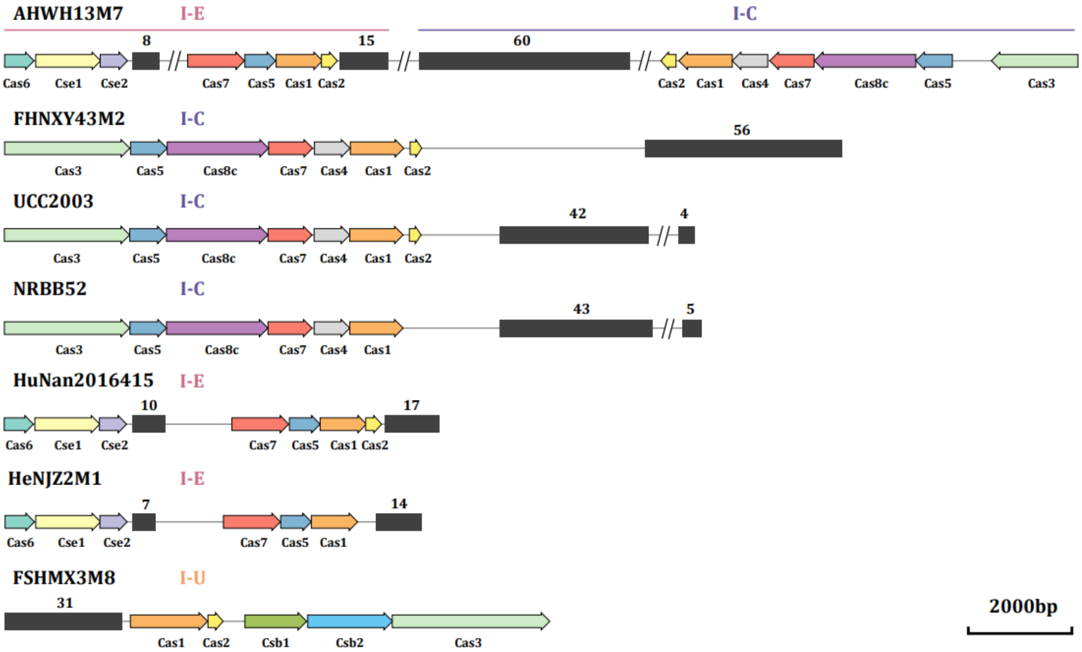
AHWH13M7
I-E
I-C
8
15
60
Cas6 Cse1 Cse2 Cas7 Cas5 Cas1 Cas2
Cas2 Cas1 Cas4 Cas7 Cas8c Cas5 Cas3
FHNXY43M2
I-C
56
Cas3 Cas5 Cas8c Cas7 Cas4 Cas1 Cas2
UCC2003
I-C
42
4
Cas3 Cas5 Cas8c Cas7 Cas4 Cas1 Cas2
NRBB52
I-C
43
5
Cas3 Cas5 Cas8c Cas7 Cas4 Cas1
HuNan2016415
I-E
10
17
Cas6 Cse1 Cse2 Cas7 Cas5 Cas1 Cas2
HeNJZ2M1
I-E
7
14
Cas6 Cse1 Cse2 Cas7 Cas5 Cas1
FSHMX3M8
I-U
31
Cas1 Cas2 Csb1 Csb2 Cas3
2000bp

Fig. 2. Visualization of representative CRISPR-Cas loci. Different *cas* genes were indicated by arrows with different colours. The repeat-spacer sequence was represented by black rectangles. Long irrelevant sequences were shortened by double disconnection. The CRISPR locus was drawn based on its size, and the bar scale indicates 2 Kb.

explore the targets of spacers, the prediction of prophages in *B. breve* was performed and the homology between spacers and prophages was analysed.

Prophages were identified in 148 out of 150 (98.7%) *B. breve* strains, except for FSHMX3M8 and JSWX25M8. Six prophages were detected in FHNFQ22M5 and BR3, although they were incomplete prophages, representing the largest number of prophages in *B. breve*. Subsequently, the nucleotide identity between spacers and prophages was analysed based on BLAST. Overall, 95 out of the 148 (64.2%) prophages were targeted by at least one spacer (Fig. 5), and the prophages detected in CNCM_I_4321 and DRBB30 were most targeted by spacers (both were 317 spacers targeted). Interestingly, spacers targeting their genomes were observed in JCM7019, FBJHD5M2, FGZ3I2M1 and ZJHZ3M2.

In order to further investigate the origin of spacers, the protein function of the predicted prophages in *B. breve* was annotated by BLASTp analyses against the NCBI nr database. The tape measure protein was targeted by spacers the most times, followed by recombinase phage RecT family, N-acetylmuramyl-L-alanine amidase negative regulator of AmpC AmpD, and phage major capsid protein (Fig. 6). The prophages predicted in DRBB28 were displayed in Fig. 7(a), in which the position of targets by the spacers can be visually observed. The tape measure protein in the prophage identified in DRBB28 was highly targeted by spacers, which was consistent with the general trend in *B. breve* (Fig. 7a).

PAMs are necessary for the acquisition and interference of spacers. The PAM is located at the 5′-end or 3′-end of each protospacer, and is usually a 2 to 5 bp sequence, which varied according to CRISPR-based system and species [34]. PAM sequences for type I-C in the 5′ flanking region of the protospacer were inferred to be 5′-TTC-3′ (Fig. 7b). Since there were few prophages detected in *B. breve* genomes that can be matched by type I-E spacers, and no prophages can be matched by type I-U spacers, the PAMs of those two subtypes in *B. breve* remained unknown. Interestingly, the PAMs for type 'undetermined' were also inferred as 5′-TTC-3′, which was the same as that in subtype I-C in *B. breve*.

DISCUSSION

The gut microbiota that coexists with the host plays an important role in maintaining human health. Among them, *B. breve* is one of the main species in the gastrointestinal tract of breastfed infants. It can provide unique protection for the intestinal health of infants and effectively reduce the incidence of intestinal infections in infants [35]. The gastrointestinal tract is a huge library of natural phages, where the struggle between bacteria and bacteriophages is constantly staged [36]. CRISPR-Cas systems are defence systems that *B. breve* evolved in order to face severe survival challenges. The strain harbouring the CRISPR-Cas system will be a candidate for industrial production and application in the future, due to its strong ability to resist bacteriophages during production.

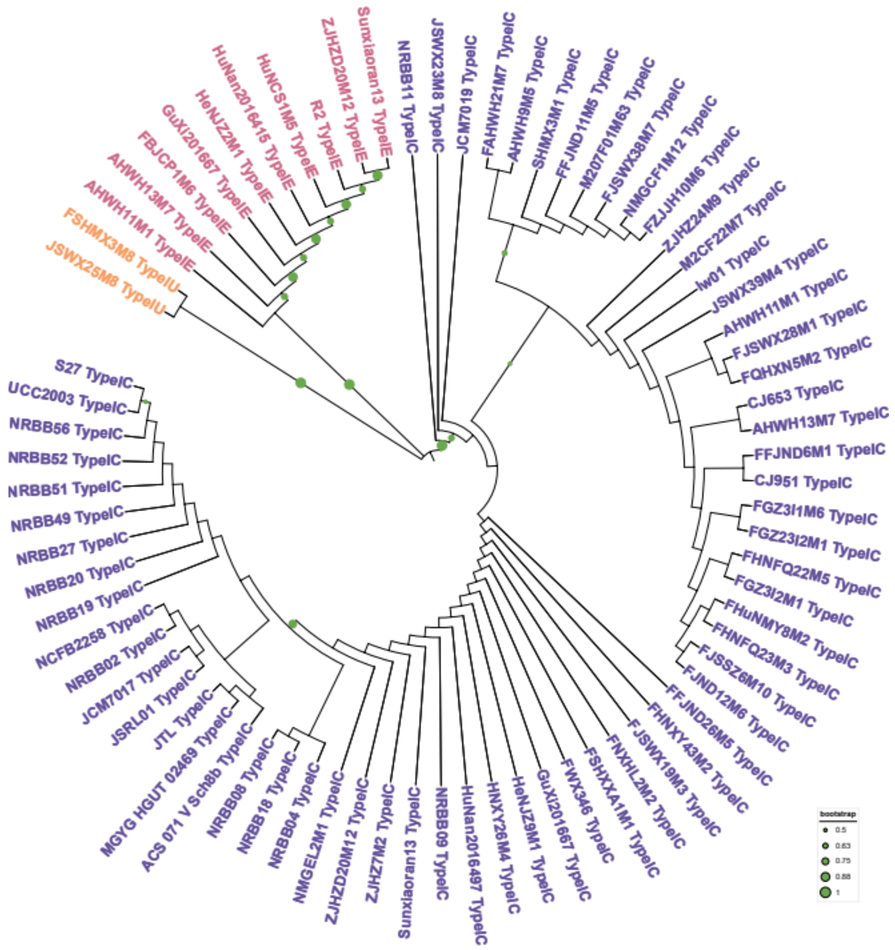

Fig. 3. Phylogenetic tree based on the amino acid sequence of Cas1 protein. Different subtypes were indicated by different colours. The neighbor-joining tree was described with 1000 bootstrap replicates. Bootstrap values were recorded on the branch and represented by the size of the circle.

The diversity of CRISPR systems in 150 *B. breve* strains from different sources was first characterized. The prevalence of CRISPR-Cas systems in *B. breve* was 47% (71/150), which was similar to that in bacteria (45%), but slightly lower than that in other bifidobacterial species (57%) [37, 38]. However, the prevalence of CRISPR-Cas systems in *B. longum* was previously reported as 38% while that in *B. pseudocatenulatum* strains was reported as 62%, suggesting differences among *Bifidobacterium* species [23, 39]. The high occurrence of CRISPR-Cas loci in the genomes of bacteria and archaea indicated that in addition to its survival, there may be other factors that prevented the loss of the CRISPR-Cas systems. A recent study on archaeal type I-B CRISPR-Cas revealed that a 311-base pair (bp) sequence between Cas6 and Cas8 can cause cytotoxicity, but it will be inhibited by the combined action of antitoxin RNA and intact

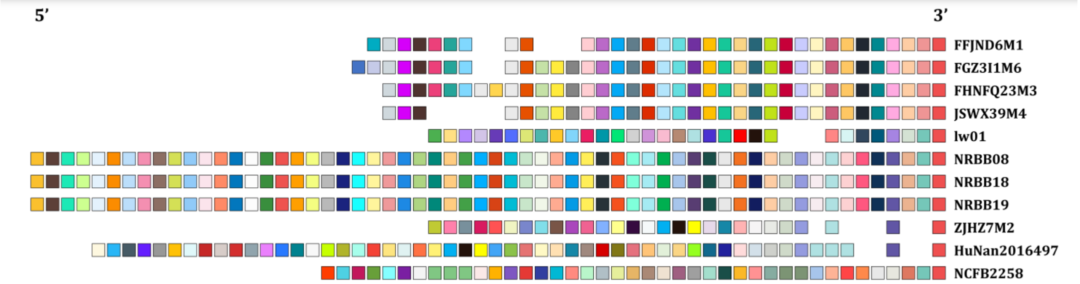

Fig. 4. Visualization of spacers in *B. breve*. Each unique spacer sequence was represented by a square with a unique colour. The first spacer of strain acquisition was shown on the right, and the last spacer was shown on the left.

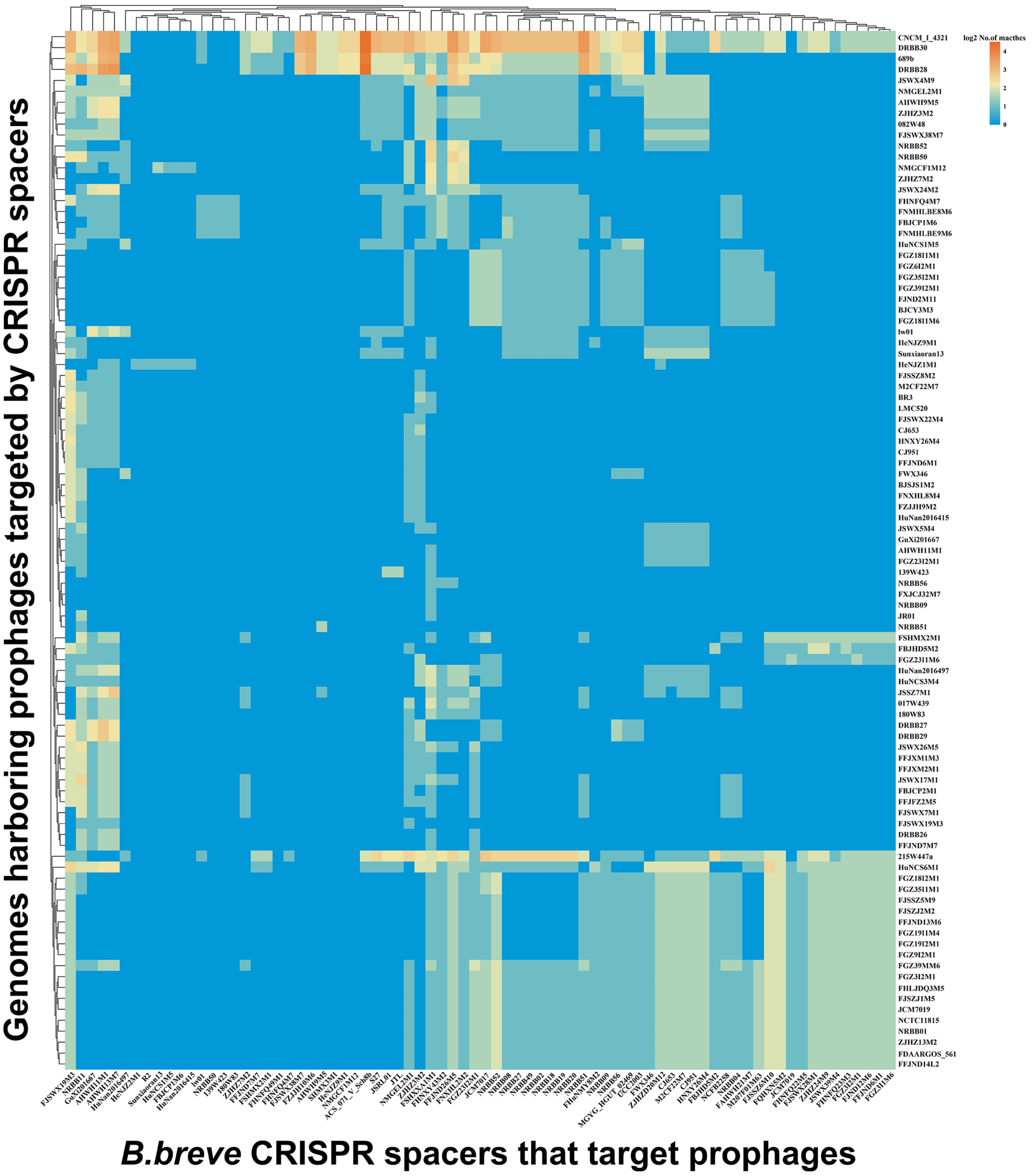

Fig. 5. Prophages in *B. breve* genomes targeted by spacers. The horizontal axis represented those *B. breve* strains carrying the spacers that targeted the prophages, and the vertical axis represented the strains carrying the prophages targeted by the spacers. The number of targeted events was indicated by different colours. The value was calculated by log2.

Cascade (any *cascade* genes deletion will result in the loss of inhibition), which made them addictive to the host [40]. CRISPR-Cas systems in *B. breve* were diverse, including type I-C, type I-E and type I-U systems. Type I-C seemed to be the most popular CRISPR system in *B. breve*, accounting for 84% of the total sum. A previous study on the CRISPR-Cas system diversity in 954 *Bifidobacterium* genomes only identified type I-C and type I* (represents untyped groups) in *B. breve* [38]. It also provided evidence that type I-C was the main CRISPR-Cas system in *B. breve*. Type II systems were absent in *B. breve*, although they have been identified in other

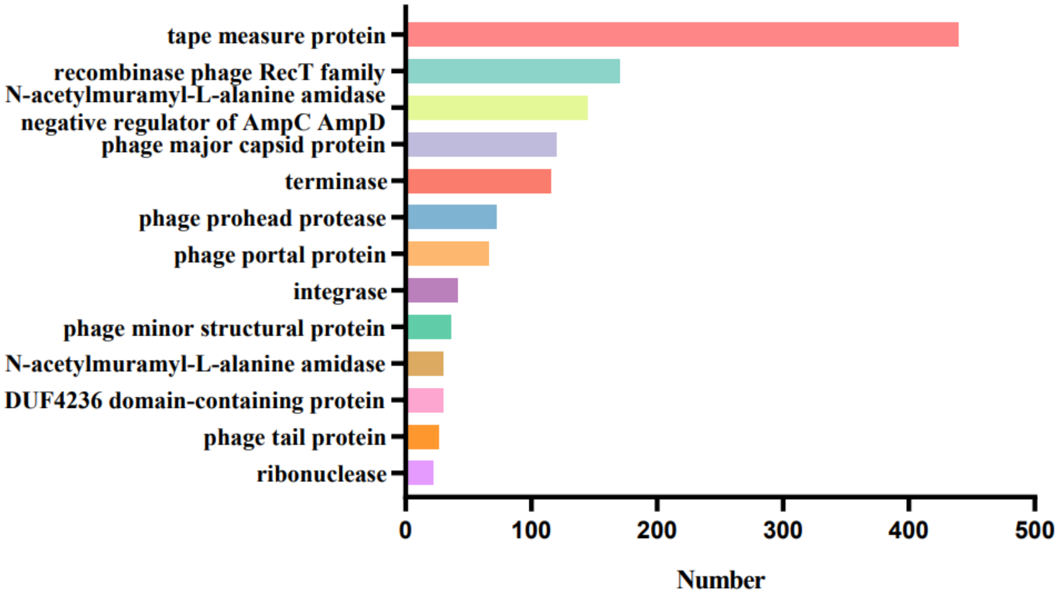

Fig. 6. Prophage protein targeted by spacers. The 13 proteins that were most frequently targeted by spacers were displayed. Protein targeting was analysed based on the NCBI nr database through BLASTp search.

Bifidobacterium, such as *B. longum*, *B. bifidum*, *B. adolescentis* and *B. pseudocatenulatum* [38]. There has been no report on the identification of type III systems in bifidobacteria.

Multiple CRISPR loci sometimes coexisted in the same strain, and they usually belonged to different subtypes. The coexistence of type I-C and type I-E was observed in five strains. The presence of these multi-subtype CRISPR-Cas systems may be due to horizontal gene transfer, which helped the bacteria to resist diverse infections. The lack of *cas2* in the CRISPR locus was found in some strains. In the adaptation stage, the Cas1-Cas2-dual-forked DNA complexes facilitated the integration of the new spacer into the CRISPR array [41]. However, Cas2 may not have a catalytic effect on the process of obtaining the spacer. Instead, it was likely to act as an adaptor protein, bringing the two Cas1 dimers together, stabilizing and measuring the length of the original spacer DNA to bind to the target, or mediating the interaction with other components necessary for catching the spacer [41, 42]. Previous research reported that, with the presence of Cas2, the protospacer integration reaction catalysed by Cas1 was significantly enhanced [43]. In the type I CRISPR-Cas system, the cleavage of the target DNA was performed by its signature protein, Cas3 that had the dual function of helicase and nuclease [44]. The absence of *cas3* gene in type I-E CRISPR loci had been also discovered in other bifidobacteria including *B. longum* subsp. *longum* 1-6B, 2-2B and 44B [23]. The insertion of partial repeat-spacer arrays into the *cas* gene clusters (such as that in *B. breve* AHWH13M7) or the truncation of repeat-spacer arrays (such as that in *B. breve* UCC2003) may result from the rearrangement during cell division or incomplete assembly of the draft genomes [23, 29]. The reverse layout of *cas* genes and repeat-spacer arrays was observed in the type I-U CRISPR-Cas system in *B. breve*, in which the repeat-spacer arrays were located upstream of *cas* genes in the genome. This may be caused by the genomic rearrangement event [29]. Further classification of subtype I-C from the evolutionary relationship found that, except for *B. breve* lw01, all strains gathered in the right subcluster (Fig. 3) were isolated from Chinese subjects in our lab. In fact, *B. breve* lw01 was also isolated from the faeces of a Chinese infant in another work [45], which meant that the strain clustering had a certain correlation with their origin source. The Cas1 protein of the same subtype seemed to be different among species. Our previous phylogenetic analysis based on Cas1 proteins from *L. gasseri* and *L. paragasseri* showed that the same subtypes were clustered in the same branch, suggesting that Cas protein can be used as a potential indicator to differentiate species with similar phylogenetic relationships [46].

The conservation of the repeat sequences was critical to the adaptive immune function of CRISPR-Cas systems [47]. Mature crRNA usually consisted of a spacer surrounded by repeat sequences, in which the 5′ end was a handle formed by a few nucleotides of the repeat sequences, and the 3′ end was a stem-loop formed by some nucleotides of another repeat sequence [48]. The stem-loop at the 3′ end was related to the specific recognition and binding of nucleases [49, 50], while the 5′ handle seemed to distinguish self from non-self through complementary pairing with CRISPR DNA repeats to prevent the occurrence of self-immunity [51]. The characteristics of repeat sequences in *B. breve* CRISPR-Cas systems showed strong subtype dependence. The repeat sequence was 33 nucleotides on average for subtype I-C, and 36 nucleotides for subtype I-U was observed in *B. breve*, which were consistent with other *Bifidobacterium* species [39]. The repeat sequence for subtype I-E was 28 nucleotides on average, while it was 29 nucleotides on

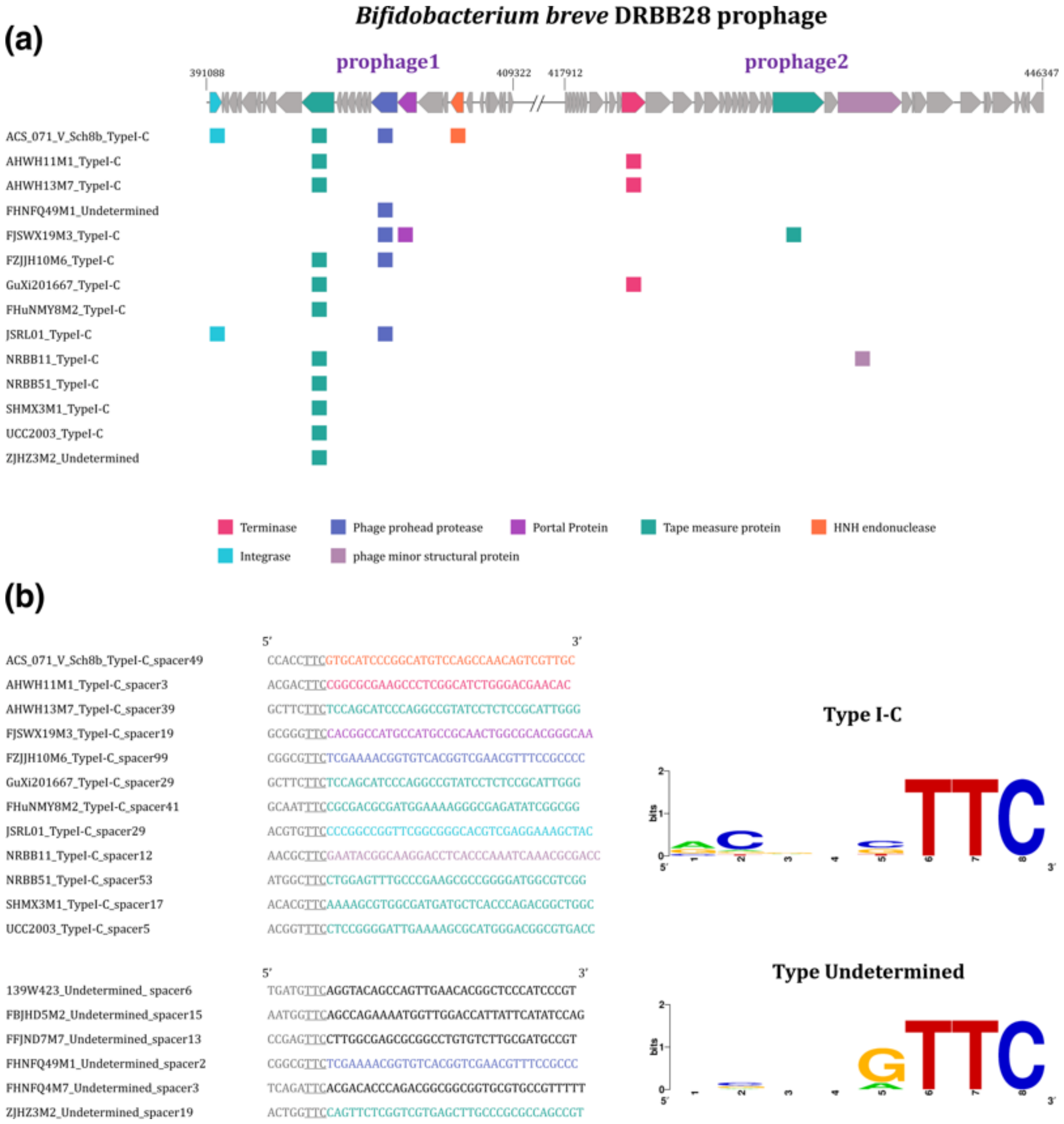

Fig. 7. Visualizing *B. breve* DRBB28 prophage protein targeted by spacers and prediction of PAMs in *B. breve*. (a) *B. breve* DRBB28 prophage protein targeted by spacers. Two of the *B. breve* DRBB28 prophages were displayed and the different prophages were separated by double disconnection. Each square represented a targeted event and was coloured according to different targeted proteins. (b) PAMs prediction. Spacers with different text colours target proteins represented as the same colour in Fig. 7(a). The eight nucleotides (grey text) flanking the 5' of each protospacer were used to predict the PAM sequence (underlined). The corresponding results were displayed on the right after visualization using the WebLogo server.

average in most *Bifidobacterium* species [38]. The order of spacers in the genome represented the temporal and spatial sequence of immune events [52]. Subtype I-C represented the most active CRISPR-Cas system in *B. breve*, ranging from 0 spacers in Sunxiaoran13 to 127 spacers in NMGEL2M1, with an average of 52 spacers. The lack of CRISPR arrays in the subtype I-C system in Sunxiaoran13 may be due to any of three reasons: (1) the locus was not active; (2) genetic recombination caused array loss; (3) the genome assembly contained errors. The number of spacers for subtype I-E was much smaller than that of subtype I-C, with an average of 26 spacers. The average number of spacers for subtype I-U was similar to that of subtype I-E, ranging from 28 spacers in JSWX25M8 to 31 spacers in FSHMX3M8, with an average of 30 spacers.

After visualizing the spacers in some strains harbouring type I-C systems, it was observed that FFJND6M1, FGZ3I1M6, FHNFQ23M3 and JSWX39M4 had a close evolutionary relationship (the first to the 24th spacers were the same). This was consistent with the results of our previous phylogenetic tree, which was constructed based on orthologous genes among *B. breve* genomes, that *B. breve* FFJND6M1, FGZ3I1M6 and JSWX39M4 were in the same evolutionary branch (FHNFQ23M3 was not involved in this study) [19]. In fact, since the arrangement of spacers was a trace of immune events, there had been previous studies on the genotyping of bacteria based on spacers. In total, 26 *L. buchneri* were genotyped based on the CRISPR locus and their distinct evolutionary paths were established [53]. CRISPR arrays were used to genotype *Cronobacter sakazakii*, *C. malonaticus* and *C. dublinensis*, which showed a greater ability to distinguish similar species than multi-locus sequence typing (MLST). More importantly, the spacers-based genotyping approach can be used to identify the origin of strains and provide a phylogenetic tree reflecting the common origin [54]. In addition, the shortcomings of CRISPR locus-based genotyping were evident, and this new classification method was limited to the strains harbouring the CRISPR-Cas system.

Tracing the origin of spacers can help us better understand the origin and the changes in the ecological niche of strains. Prophage integrated into the bacterial genome can be induced to be infectious phage under certain conditions [55], hence, we first predicted the prophage in *B. breve* genomes. The high detection rate was obtained because our aim was to find more targets of the spacers, rather than discarding the 'suspicious' prophage similar to other previous reports [39, 56]. At the same time, the results of matching spacers and prophages can also be used as a reference for the reliability of prophages prediction. Interestingly, *B. breve* JCM7019, FBJHD5M2, FGZ3I2M1 and ZJHZ3M2 contained the spacers that targeted the prophages in their genomes, which may be an accident of the CRISPR insertion mechanism. The spacers in FBJHD5M2 and ZJHZ3M2 were classified as type 'undetermined', and only repeat-spacer arrays but no *cas* genes were identified in the genomes. We speculated that those two strains may have lost *cas* genes in order to avoid self-targeting fatal events [57]. Since most phages entered the cell as linear DNA, the double-strand break repair helicase/nuclease complex RecBCD could bind, untie, and degrade any exposed linear DNA. A large number of 'fragments' generated in this process will serve as a substrate for the Cas1-Cas2 complex to obtain the spacer. Because RecBCD usually stopped degrading DNA after recognising *Chi* sites, the high density of *Chi* sites on the bacterial chromosome further protected it from the acquisition of the spacer [58]. In fact, spacer self-targeting was very common in bacteria, especially in the strains possessing prophage. This phenomenon of both prophage and spacer targeted prophage in the same strain may be caused by two infections of the same phage, one of which was successful and the other was resisted by the CRISPR-Cas system. In this case, anti-CRISPR (Acr) proteins encoded by prophages may help bacteria to develop tolerance to spacer self-targeting [59]. This mean that after the phages integrated into the bacterial genomes, the arms race between phages and bacteria were still not over. However, due to the paucity of currently available bifidobacterial phage genomes, the functions of many prophage proteins cannot be annotated and were described as 'hypothetical proteins'. If the hypothetical proteins were not considered, the spacers will be more targeted at tape measure protein, recombinase RecT, phage portal protein, phage major capsid protein and terminase. The length of tape measure protein (TMP) determined the length of the tail, which was crucial in the phage assembly process [60]. Recombinase RecT bound to ssDNA, worked with a partner protein RecE, and participated in the pairing of the homologous DNA [61]. The major capsid protein was assembled cooperatively with the help of the capsid scaffold protein to form the procapsid, which was an important step in the phage assembly [62]. During the assembly process, proteases degraded the scaffold proteins or delta domain proteases, then the terminase bound to portal protein and sent phage DNA into the capsid, and HNH endonuclease played an important role in this process [63, 64]. Therefore, CRISPR systems may mainly prevent the assembly in the phage life cycle to make the strain survive. The PAMs of subtype I-C and type 'undetermined' in *B. breve* were all inferred as 5′-TTC-3′. Additionally, the frequency of 5′-TTC-3′ in each protein in the prophages was analysed to explore whether the preference of the spacer targeting protein was related to the amount of PAM sequences. Unfortunately, no correlation was found in this work.

CRISPR-based gene-editing tools have been widely used in bacteria, including *L. reuteri*, *L. casei*, *Enterococcus Faecium* and *Lactococcus lactis* [65–68]. However, due to the active restriction-modification system in *B. breve*, it was difficult to introduce exogenous plasmids carrying nucleases [69]. Therefore, the endogenous CRISPR-Cas technology may be an ideal approach for genetic modification in *B. breve*. A previous study had successfully used the endogenous type I-E CRISPR-Cas system of *L. crispatus* for efficient genetic engineering, including insertions, deletions and single-base substitution [70]. Although currently there is no report on the application of endogenous CRISPR-Cas system for gene editing in *Bifidobacterium*, it has broad prospects because of the high incidence of CRISPR-Cas systems in *Bifidobacterium*. In summary, this study characterized the CRISPR systems in *B. breve* and explored the association between CRISPR-Cas systems and prophages, which could provide a theoretical basis for genotyping and the further development of next-generation genome editing tools for *B. breve*.

Funding information

This study was funded by National Natural Science Foundation of China (Nos. 32021005, 31820103010), 111 Project (BP0719028), and Collaborative Innovation Center of Food Safety and Quality Control in Jiangsu Province.

Author contributions

X.H.: methodology, validation, investigation, data curation, writing-original draft, visualization. X.Z.: formal analysis, investigation, data curation. Z.P.: formal analysis, investigation, data curation. C.S.: writing-review and editing. R.P.R.: writing-review and editing. J.Z.: resources, data curation. H.Z.: conceptualization, supervision, funding. B.Y.: conceptualization, validation, investigation, data curation, writing-original draft, writing-review and editing, visualization, supervision. W.C.: conceptualization, supervision, project administration, funding.

Conflicts of interest

The authors declare that there are no conflicts of interest.

Ethical statement

The 104 strains used in this study were all isolated from fecal samples of healthy Chinese people. The collection of fecal samples was approved by the Ethics Committee of Jiangnan University, China (SYXK 2012–0002). Written informed consent for the use of fecal samples was obtained from the participants or their legal guardian before sampling. The health questionnaire was conducted before sampling, and no human experiments were involved. These were the only human materials used in this study. The collection of fecal samples did not pose a predictable risk of harm or discomfort to the participants. All those strains were deposited at Culture Collection of Food Microorganisms (CCFM), Jiangnan University.

References

1. Kelly D, King T, Aminov R. Importance of microbial colonization of the gut in early life to the development of immunity. *Mutat Res* 2007;622:58–69.
2. Lloyd-Price J, Abu-Ali G, Huttenhower C. The healthy human microbiome. *Genome Med* 2016;8:1–11.
3. Makarova KS, Wolf YI, Iranzo J, Shmakov SA, Alkhnbashi OS, *et al*. Evolutionary classification of CRISPR-Cas systems: a burst of class 2 and derived variants. *Nat Rev Microbiol* 2020;18:67–83.
4. Goh YJ, Barrangou R. Harnessing CRISPR-Cas systems for precision engineering of designer probiotic lactobacilli. *Curr Opin Biotechnol* 2019;56:163–171.
5. Jiang W, Zhou H, Bi H, Fromm M, Yang B, *et al*. Demonstration of CRISPR/Cas9/sgRNA-mediated targeted gene modification in Arabidopsis, tobacco, sorghum and rice. *Nucleic Acids Res* 2013;41:20.
6. Harms DW, Quadros RM, Seruggia D, Ohtsuka M, Takahashi G, *et al*. Mouse genome editing using the CRISPR/Cas system. *Curr Protoc Hum Genet* 2014;83:15..
7. Giacalone JC, Sharma TP, Burnight ER, Fingert JF, Mullins RF, *et al*. CRISPR-Cas9-based genome editing of human induced pluripotent stem cells. *Curr Protoc Stem Cell Biol* 2018;44:5B.
8. Hidalgo-Cantabrana C, Delgado S, Ruiz L, Ruas-Madiedo P, Sánchez B, *et al*. Bifidobacteria and their health-promoting effects. *Microbiol Spectr* 2017;5.
9. Turroni F, Peano C, Pass DA, Foroni E, Severgnini M, *et al*. Diversity of bifidobacteria within the infant gut microbiota. *PLoS One* 2012;7:e36957.
10. Braga TD, da Silva GAP, de Lira PIC, de Carvalho Lima M. Efficacy of *Bifidobacterium breve* and *Lactobacillus casei* oral supplementation on necrotizing enterocolitis in very-low-birth-weight preterm infants: a double-blind, randomized, controlled trial. *Am J Clin Nutr* 2011;93:81–86.
11. Klemenak M, Dolinšek J, Langerholc T, Di Gioia D, Mičetić-Turk D. Administration of *Bifidobacterium breve* decreases the production of TNF-α in children with celiac disease. *Dig Dis Sci* 2015;60:3386–3392.
12. Solito A, Bozzi Cionci N, Calgaro M, Caputo M, Vannini L, *et al*. Supplementation with *Bifidobacterium breve* BR03 and B632 strains improved insulin sensitivity in children and adolescents with obesity in a cross-over, randomized double-blind placebo-controlled trial. *Clin Nutr* 2021;40:4585–4594.
13. Enomoto T, Sowa M, Nishimori K, Shimazu S, Yoshida A, *et al*. Effects of bifidobacterial supplementation to pregnant women and infants in the prevention of allergy development in infants and on fecal microbiota. *Allergol Int* 2014;63:575–585.
14. Misra CS, Bindal G, Sodani M, Wadhawan S, Kulkarni S, *et al*. Determination of Cas9/dCas9 associated toxicity in microbes. *Microbiology* 2019:848135.
15. Wurihan W, Huang Y, Weber AM, Wu X, Fan H. Nonspecific toxicities of *Streptococcus pyogenes* and *Staphylococcus aureus* dCas9 in *Chlamydia trachomatis*. *Pathog Dis* 2019;77:ftaa005.
16. Cho S, Choe D, Lee E, Kim SC, Palsson B, *et al*. High-level dCas9 expression induces abnormal cell morphology in *Escherichia coli*. *ACS Synth Biol* 2018;7:1085–1094.
17. Bottacini F, Morrissey R, Roberts RJ, James K, van Breen J, *et al*. Comparative genome and methylome analysis reveals restriction/modification system diversity in the gut commensal *Bifidobacterium breve*. *Nucleic Acids Res* 2018;46:1860–1877.
18. O' Connell Motherway M, Watson D, Bottacini F, Clark TA, Roberts RJ, *et al*. Identification of restriction-modification systems of *Bifidobacterium animalis* subsp. *lactis* CNCM I-2494 by SMRT sequencing and associated methylome analysis. *PLoS One* 2014;9:e94875.
19. Liu R, Yang B, Stanton C, Paul Ross R, Zhao J, *et al*. Comparative genomics and gene-trait matching analysis of *Bifidobacterium breve* from Chinese children. *Food Biosci* 2020;36:100631.
20. Couvin D, Bernheim A, Toffano-Nioche C, Touchon M, Michalik J, *et al*. CRISPRCasFinder, an update of CRISRFinder, includes a portable version, enhanced performance and integrates search for Cas proteins. *Nucleic Acids Res* 2018;46:W246–W251.
21. Biswas A, Staals RHJ, Morales SE, Fineran PC, Brown CM. CRISPRDetect: A flexible algorithm to define CRISPR arrays. *BMC Genomics* 2016;17:356.
22. Kumar S, Stecher G, Li M, Knyaz C, Tamura K. MEGA X: molecular evolutionary genetics analysis across computing platforms. *Mol Biol Evol* 2018;35:1547–1549.
23. Hidalgo-Cantabrana C, Crawley AB, Sanchez B, Barrangou R. Characterization and exploitation of CRISPR Loci in *Bifidobacterium longum Front Microbiol* 2017;8:1851.
24. Letunic I, Bork P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res* 2021;49:W293–W296.
25. Arndt D, Grant JR, Marcu A, Sajed T, Pon A, *et al*. PHASTER: a better, faster version of the PHAST phage search tool. *Nucleic Acids Res* 2016;44:W16-21.
26. Chen C, Chen H, Zhang Y, Thomas HR, Frank MH, *et al*. TBtools: an integrative toolkit developed for interactive analyses of big biological data. *Mol Plant* 2020;13:1194–1202.
27. Seemann T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 2014;30:2068–2069.
28. Crooks GE, Hon G, Chandonia JM, Brenner SE. WebLogo: a sequence logo generator. *Genome Res* 2004;14:1188–1190.
29. Briner AE, Lugli GA, Milani C, Duranti S, Turroni F, *et al*. Occurrence and diversity of CRISPR-cas systems in the genus *Bifidobacterium*. *PLoS One* 2015;10:e0133661.
30. Crawley AB, Henriksen ED, Stout E, Brandt K, Barrangou R. Characterizing the activity of abundant, diverse and active CRISPR-Cas systems in lactobacilli. *Sci Rep* 2018;8:11544.
31. Makarova KS, Haft DH, Barrangou R, Brouns SJJ, Charpentier E, *et al*. Evolution and classification of the CRISPR-Cas systems. *Nat Rev Microbiol* 2011;9:467–477.
32. Leon LM, Mendoza SD, Bondy-Denomy J. How bacteria control the CRISPR-Cas arsenal. *Curr Opin Microbiol* 2018;42:87–95.
33. Garrett SC. Pruning and tending immune memories: spacer dynamics in the CRISPR array. *Front Microbiol* 2021;12:664299.
34. Shah SA, Erdmann S, Mojica FJM, Garrett RA. Protospacer recognition motifs: mixed identities and functional diversity. *RNA Biol* 2013;10:891–899.
35. Bozzi Cionci N, Baffoni L, Gaggìa F, Di Gioia D. Therapeutic microbiology: the role of *Bifidobacterium breve* as food supplement for the prevention/treatment of paediatric diseases. *Nutrients* 2018;10:E1723.
36. Manrique P, Dills M, Young MJ. The human gut phage community and its implications for health and disease. *Viruses* 2017;9:E141.
37. Crawley AB, Henriksen JR, Barrangou R. CRISPRdisco: an automated pipeline for the discovery and analysis of CRISPR-Cas systems. *CRISPR J* 2018;1:171–181.

38. Pan M, Nethery MA, Hidalgo-Cantabrana C, Barrangou R. Comprehensive mining and characterization of CRISPR-cas systems in *Bifidobacterium*. *Microorganisms* 2020;8:720.

39. Wang G, Liu Q, Pei Z, Wang L, Tian P, *et al*. The diversity of the CRISPR-Cas system and prophages present in the genome reveals the co-evolution of *Bifidobacterium pseudocatenulatum* and phages. *Front Microbiol* 2020;11:1088.

40. Li M, Gong L, Cheng F, Yu H, Zhao D, *et al*. Toxin-antitoxin RNA pairs safeguard CRISPR-Cas systems. *Science* 2021;372:eabe5601.

41. Wang J, Li J, Zhao H, Sheng G, Wang M, *et al*. Structural and mechanistic basis of PAM-dependent spacer acquisition in CRISPR-Cas systems. *Cell* 2015;163:840–853.

42. Rollie C, Schneider S, Brinkmann AS, Bolt EL, White MF. Intrinsic sequence specificity of the Cas1 integrase directs new spacer acquisition. *Elife* 2015;4.

43. Nuñez JK, Lee ASY, Engelman A, Doudna JA. Integrase-mediated spacer acquisition during CRISPR-Cas adaptive immunity. *Nature* 2015;519:193–198.

44. Huo Y, Nam KH, Ding F, Lee H, Wu L, *et al*. Structures of CRISPR Cas3 offer mechanistic insights into Cascade-activated DNA unwinding and degradation. *Nat Struct Mol Biol* 2014;21:771–777.

45. Wang L, Wang Y, Li Q, Tian K, Xu L, *et al*. Exopolysaccharide, isolated from a novel strain *Bifidobacterium breve* lw01 possess an anticancer effect on head and neck cancer - genetic and biochemical evidences. *Front Microbiol* 2019;10:1044.

46. Zhou X, Yang B, Stanton C, Ross RP, Zhao J, *et al*. Comparative analysis of *Lactobacillus gasseri* from Chinese subjects reveals a new species-level taxa. *BMC Genomics* 2020;21:119.

47. Hille F, Charpentier E. CRISPR-Cas: biology, mechanisms and relevance. *Philos Trans R Soc Lond B Biol Sci* 2016;371:20150496.

48. Brouns SJJ, Jore MM, Lundgren M, Westra ER, Slijkhuis RJH, *et al*. Small CRISPR RNAs guide antiviral defense in prokaryotes. *Science* 2008;321:960–964.

49. Sorek R, Kunin V, Hugenholtz P. CRISPR--a widespread system that provides acquired resistance against phages in bacteria and archaea. *Nat Rev Microbiol* 2008;6:181–186.

50. Kunin V, Sorek R, Hugenholtz P. Evolutionary conservation of sequence and secondary structures in CRISPR repeats. *Genome Biol* 2007;8:R61.

51. Marraffini LA, Sontheimer EJ. Self versus non-self discrimination during CRISPR RNA-directed immunity. *Nature* 2010;463:568–571.

52. Mohanraju P, Makarova KS, Zetsche B, Zhang F, Koonin EV, *et al*. Diverse evolutionary roots and mechanistic variations of the CRISPR-Cas systems. *Science* 2016;353:aad5147.

53. Briner AE, Barrangou R. *Lactobacillus buchneri* genotyping on the basis of clustered regularly interspaced short palindromic repeat (CRISPR) locus diversity. *Appl Environ Microbiol* 2014;80:994–1001.

54. Zeng H, Li C, He W, Zhang J, Chen M, *et al*. *Cronobacter sakazakii*, *Cronobacter malonaticus*, and *Cronobacter dublinensis* genotyping based on CRISPR locus diversity. *Front Microbiol* 2019;10:1989.

55. Pei Z, Sadiq FA, Han X, Zhao J, Zhang H, *et al*. Identification, characterization, and phylogenetic analysis of eight new inducible prophages in *Lactobacillus*. *Virus Res* 2020;286:198003.

56. Lugli GA, Milani C, Turroni F, Tremblay D, Ferrario C, *et al*. Prophages of the genus *Bifidobacterium* as modulating agents of the infant gut microbiota. *Environ Microbiol* 2016;18:2196–2213.

57. Stern A, Keren L, Wurtzel O, Amitai G, Sorek R. Self-targeting by CRISPR: gene regulation or autoimmunity? *Trends Genet* 2010;26:335–340.

58. Levy A, Goren MG, Yosef I, Auster O, Manor M, *et al*. CRISPR adaptation biases explain preference for acquisition of foreign DNA. *Nature* 2015;520:505–510.

59. Stanley SY, Borges AL, Chen K-H, Swaney DL, Krogan NJ, *et al*. Anti-CRISPR-associated proteins are crucial repressors of anti-CRISPR transcription. *Cell* 2019;178:1452–1464.

60. Katsura I, Hendrix RW. Length determination in bacteriophage lambda tails. *Cell* 1984;39:691–698.

61. Murphy KC. Phage recombinases and their applications. *Adv Virus Res* 2012;83:367–414.

62. Oh B, Moyer CL, Hendrix RW, Duda RL. The delta domain of the HK97 major capsid protein is essential for assembly. *Virology* 2014;456–457:171–178.

63. Kala S, Cumby N, Sadowski PD, Hyder BZ, Kanelis V, *et al*. HNH proteins are a widespread component of phage DNA packaging machines. *Proc Natl Acad Sci U S A* 2014;111:6022–6027.

64. Aksyuk AA, Rossmann MG. Bacteriophage assembly. *Viruses* 2011;3:172–203.

65. Oh J-H, van Pijkeren J-P. CRISPR-Cas9-assisted recombineering in *Lactobacillus reuteri*. *Nucleic Acids Res* 2014;42:e131.

66. Song X, Huang H, Xiong Z, Ai L, Yang S. CRISPR-Cas9^{D10A} nickase-assisted genome editing in *Lactobacillus casei*. *Appl Environ Microbiol* 2017;83:e01259-17.

67. de Maat V, Stege PB, Dedden M, Hamer M, van Pijkeren J-P, *et al*. CRISPR-Cas9-mediated genome editing in vancomycin-resistant *Enterococcus faecium*. *FEMS Microbiol Lett* 2019;366:fnz256.

68. Guo T, Xin Y, Zhang Y, Gu X, Kong J. A rapid and versatile tool for genomic engineering in *Lactococcus lactis*. *Microb Cell Fact* 2019;18:22.

69. O'Connell Motherway M, O'Driscoll J, Fitzgerald GF, Van Sinderen D. Overcoming the restriction barrier to plasmid transformation and targeted mutagenesis in *Bifidobacterium breve* UCC2003. *Microb Biotechnol* 2009;2:321–332.

70. Hidalgo-Cantabrana C, Goh YJ, Pan M, Sanozky-Dawes R, Barrangou R. Genome editing using the endogenous type I CRISPR-Cas system in *Lactobacillus crispatus Proc Natl Acad Sci U S A* 2019;116:15774–15783.